



TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANABİLİM DALI

**16S rRNA GEN DİZİLEME VERİLERİNİ
KULLANARAK PEDIATRİK CROHN
HASTALIĞININ TANISINDA MAKİNE
ÖĞRENMESİ MODELLERİNİN
UYGULANMASI**

Sinem ÖZDEMİR

DOKTORA TEZİ

Prof. Dr. Kemal TURHAN

TRABZON-2025

BEYAN

Bu tez çalışmasının Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Tez Hazırlama ve Yazım Kılavuzu standartlarına uygun olarak yapıldığını ve yazıldığını, tezin akademik ve etik kurallara bağlı kalınarak gerçekleştirilmiş özgün bir bilimsel araştırma eserim olduğunu, tezde yer alan ve bu tez çalışmasıyla elde edilmeyen tüm bilgi ve yorumlara kaynak gösterdiğimi ve kullanılan kaynakların kaynaklar listesinde yer aldığını, tezin çalışılması ve yazımı aşamalarında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

03/02/2024

Sinem ÖZDEMİR

İthaf

Bu doktora tezimi, benim bu günlere gelmem için hiçbir fedakârlıktan kaçınmayan ve her zaman yanımda olan sevgili anneme ithaf ediyorum.



TEŞEKKÜR

Doktora eğitimim süresince bilgi birikimi, rehberliği ve geniş bakış açısıyla beni yönlendiren, kendimi geliştirmemde büyük katkıları olan değerli hocam Prof. Dr. Kemal TURHAN'a en içten teşekkürlerimi sunarım.

Doktora eğitimime başladığım günden itibaren destek ve yardımlarını esirgemeyen, her zaman yanımda olan değerli hocalarım Doç. Dr. Burçin KURT ve Dr. Serbülent ÜNSAL'a şükranlarımı sunarım. Ayrıca eğitim sürecim boyunca, sadece akademik yönlendirmeleriyle değil, motivasyon ve samimi destekleriyle yanımda olan Dr. Zeliha AYDIN KASAP'a ve tez sürecimde desteğini esirgemeyen Mehmet Erşan KALAYCI'ya teşekkür ederim. Bu zorlu yolda benimle birlikte yürüyen, her daim yanımda olan değerli arkadaşım Işık ÇAKMAK'a da teşekkürlerimi iletmek isterim.

Hayatımın her anında bana güç veren, varlıklarıyla beni her zaman daha ileriye taşıyan aileme, özellikle emeklerini hiçbir zaman ödeyemeyeceğim sevgili anneme sonsuz teşekkürlerimi sunuyorum. Güzel gülümsemeleriyle hayatı daha anlamlı kılan canım yeğenlerim Işıl ve Emir'e de sevgilerimle teşekkür ederim.

Doktora eğitimime YÖK 100/2000 bursu ile katkı sağlayan YÖK'e teşekkür ederim. Bu araştırmada yer alan tüm/kısmi nümerik hesaplamalar TÜBİTAK ULAKBİM, Yüksek Başarım ve Grid Hesaplama Merkezi'nde (TRUBA kaynaklarında) gerçekleştirilmiştir.

Sinem ÖZDEMİR

İÇİNDEKİLER

Sayfa

İÇ KAPAK SAYFASI

ONAY

BEYAN

İthaf

TEŞEKKÜR

İÇİNDEKİLER

vi

TABLolar DİZİNİ

ix

ŞEKİLLER DİZİNİ

xi

KISALTMALAR, SİMGELER ve FORMÜLLER DİZİNİ

xiii

ÖZET

xvi

ABSTRACT

xvii

1. GİRİŞ ve AMAÇ

1

2. GENEL BİLGİLER

3

2.1. Crohn Hastalığı

3

2.1.1. Klinik Bulgular Semptomlar

4

2.1.2. Risk Faktörleri

5

2.1.2.1. Çevresel Faktörler

5

2.1.2.2. Genetik Faktörler

6

2.2. Pediatrik Crohn Hastalığı

6

2.3. Crohn Hastalığı Tanı ve Görüntüleme Yöntemleri

7

2.3.1. Endoskopik Görüntüleme

7

2.3.2. Kesitsel Görüntüleme

8

2.3.3. Histoloji

8

2.3.4. Biyobelirteç

8

2.3.5. Serolojik Belirteçler

8

2.3.6. C-Reaktif Protein (CRP)

9

2.3.7. Dışkıdaki Kalprotektin

9

2.3.8. Yeni Potansiyel İnvazif Olmayan Belirteçler

9

2.4. İnsan Mikrobiyomu

10

2.4.1. Mikrobiyom Profilleme

11

2.4.1.1. Marker Gen Sekanslama Verisi

12

2.4.1.2. Tüm Metagenom Dizileme Verileri (Shotgun Metagenomik Dizileme)	12
2.5. Çeşitlilik Analizi	12
2.5.1. Alfa Çeşitlilik Analizi	12
2.5.1.1. Chao1 İndeksi	13
2.5.1.2. Faith'in Filogenetik Çeşitliliği	13
2.5.1.3. Pielou Düzgünlük İndeksi	13
2.5.1.4. Shannon-Wiener Çeşitlilik İndeksi	13
2.5.2. Beta Çeşitlilik	14
2.5.2.1. UniFrac Mesafesi	14
3. GEREÇ ve YÖNTEM	16
3.1. Veri Kümesinin Elde Edilmesi	17
3.2. Temsil Vektörlerinin Oluşturulması	17
3.2.1. K-mer Temsilleri	17
3.2.2. ASV Temsil Vektörlerinin Oluşturulması	19
3.2.3. DNABERT-S Temsil Vektörlerinin Oluşturulması	21
3.2.3.1. Müfredat Karşıt Öğrenim	21
3.2.3.2. SimCLR Metodu	22
3.3. Yeniden Örnekleme Yöntemleri	24
3.3.3.1. Sentetik Azınlık Örnekleme Arttırma Yöntemi (SMOTE)	24
3.3.3.2. Rastgele Örnek Seçimi ile Veri Arttırımı	24
3.3.4. Sınıflayıcılar	25
3.3.4.1. Destek Vektör Makineleri	25
3.3.4.2. Rastgele Orman Algoritması	25
3.3.4.3. Uyarlanabilir Güçlendirme Algoritması	26
3.3.4.4. Karar Ağacı Algoritması	27
3.3.5. Performans Değerlendirme Ölçütleri	28
3.3.6. Kullanılan Platform ve Kütüphaneler	32
3.3.7. Hiper Parametre Optimizasyonu	32
4. BULGULAR	33
4.1. Veri Setine Ait Bulgular	33
4.2. Sınıflayıcılara Ait Sonuçlar	44
5. TARTIŞMA ve SONUÇ	54
KAYNAKLAR	58

EKLER	72
Ek 1. Model Sonuçları	73
Ek 2. Modelin Parametre Uzayının GridSearchCV ile Taranması Sonucu En İyi Modele Ait Hiper Parametre Uzayı	77
Ek 3. Test Verisi Üzerinde Poligon Alan Metriğine Ait Gösterim	80
Ek 4. Ağırlıklandırılmış Test Verisi Sonuçları ve Ağırlıklandırılmış Metrikler Üzerinden Hesaplanan Poligon Alan Metrik Sonuçları	84
ÖZGEÇMİŞ	89



TABLÖLAR DİZİNİ

Tablo No	Sayfa
Tablo 1. Montreal sınıflamasına göre Crohn hastalığı	3
Tablo 2. Crohn hastalığı çevresel risk faktörleri	5
Tablo 3. İnsan bağırsak mikrobiyotasının altı ana şubesi ve baskın türleri	11
Tablo 4. Hastalıklara bağlı olarak 16s rRNA verisinin alındığı bölgelere göre değişen veri dağılımı	17
Tablo 5. 3-mer,4-mer,5-mer yapısının oluşturulmasına ait sekans örneği	18
Tablo 6. Hata matrisi	29
Tablo 7. Performans ölçütleri	30
Tablo 8. Sınıflandırıcılara ait parametre uzayları	32
Tablo 9. Kategorik verilere ait tanımlayıcı istatistikler	33
Tablo 10. Nümerik verilere ait tanımlayıcı istatistikler	35
Tablo 11. Özellik tablo özeti	36
Tablo 12. Vücut bölgesine bağlı Faith_pd değerlerinin Mann-Whitney-U analizine ait sonuçları	36
Tablo 13. Teşhis değişkenine bağlı Faith_pd değerlerinin Mann-Whitney U analizine ait sonuçları	37
Tablo 14. Cinsiyet değişkenine bağlı Faith_pd değerlerinin Mann Whitney-U analizine ait sonuçları	38
Tablo 15. Vücut bölgesi değişkenine bağlı Pielou düzgünlük değerlerinin Mann Whitney-U analizine ait sonuçları	38
Tablo 16. Teşhis değişkenine bağlı Pielou düzgünlük değerlerinin Mann Whitney-U analizine ait sonuçları	39
Tablo 17. Cinsiyet değişkenine bağlı Pielou düzgünlük değerlerinin Mann Whitney-U analizine ait sonuçları	40
Tablo 18. Vücut bölgesi değişkenine bağlı Shannon Entropi değerlerinin MannWhitney-U analizine ait sonuçları	40
Tablo 19. Teşhis değişkenine bağlı Shannon Entropi değerlerinin Mann Whitney-U analizine ait sonuçları	41
Tablo 20. Cinsiyet değişkenine bağlı Shannon Entropi değerlerinin Mann Whitney-U analizine ait sonuçları	42

Tablo 21. Vücut bölgesi değişkenine ait ağırlıklandırılmamış UniFrac uzaklığı PERMANOVA sonuçları	42
Tablo 22. Teşhis değişkenine ait ağırlıklandırılmamış UniFrac uzaklığı PERMANOVA sonuçları	42
Tablo 23. Cinsiyet değişkenine ait ağırlıklandırılmamış UniFrac uzaklığı PERMANOVA sonuçları	43
Tablo 24. Vücut bölgesi değişkenine ait ağırlıklandırılmış UniFrac uzaklığı PERMANOVA sonuçları	43
Tablo 25. Teşhis değişkenine ait ağırlıklandırılmış UniFrac uzaklığı PERMANOVA sonuçları	44
Tablo 26. Cinsiyet değişkenine ait ağırlıklandırılmış UniFrac uzaklığı PERMANOVA sonuçları	44

ŞEKİLLER DİZİNİ

Şekil No	Sayfa
Şekil 1. İnsan mikrobiyotasındaki disbiyoz ve çeşitli hastalıkların ilişkisi	10
Şekil 2. Tez kapsamına ait akış	16
Şekil 3. Meta verinin Keemei aracı ile düzenlenmesi	19
Şekil 4. Cins ve tür seviye temsil vektörlerinin birleştirilmesine ait örnek gösterim	21
Şekil 5. Karşıtlıklı öğrenme için mimari	23
Şekil 6. Doğrusal SVM sınıflayıcı	25
Şekil 7. RF modeli	26
Şekil 8. AdaBoost modeli	27
Şekil 9. Karar ağacı	27
Şekil 10. Poligon alan metriği gösterimi	28
Şekil 11. Kategorik verilere ait tanımlayıcı istatistikler	34
Şekil 12. Nümerik yaş değişkenine ait kutu grafiği	35
Şekil 13. Vücut bölgesine bağlı Faith_pd değerinin dağılımı	36
Şekil 14. Teşhis değişkenine bağlı Faith_pd değerinin dağılımı	37
Şekil 15. Cinsiyet değişkenine bağlı Faith_pd değerinin dağılımı	37
Şekil 16. Vücut bölgesi değişkenine bağlı Pielou Düzgünlük değerinin dağılımı	38
Şekil 17. Teşhis değişkenine bağlı Pielou Düzgünlük değerinin dağılımı	39
Şekil 18. Cinsiyet değişkenine bağlı Pielou Düzgünlük değerinin dağılımı	39
Şekil 19. Vücut bölgesi değişkenine bağlı Shannon Entropi değerinin dağılımı	40
Şekil 20. Teşhis değişkenine bağlı Shannon Entropi değerinin dağılımı	41
Şekil 21. Cinsiyet değişkenine bağlı Shannon Entropi değerinin dağılımı	41
Şekil 22. Orijinal veri kümesi üzerinde eğitilmiş olan modellerin 8 farklı temsil vektörü ile çalıştırılması ile elde edilen ağırlıklandırılmış F1 skorlarına ait grafikler	45
Şekil 23. Rastgele yeniden örnekleme yaklaşımı kullanılarak elde edilen veri kümesi üzerinde eğitilmiş olan modellerin 8 farklı temsil vektörü ile çalıştırılması ile elde edilen ağırlıklandırılmış F1 skorlarına ait grafikler	46
Şekil 24. SMOTE yeniden örnekleme yaklaşımı kullanılarak elde edilen veri kümesi üzerinde eğitilmiş olan modellerin 8 farklı temsil vektörü ile çalıştırılması ile elde edilen ağırlıklandırılmış F1 skorlarına ait grafikler	46

Şekil 25. Eğitim verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait Hamming kayıp skoru	47
Şekil 26. Eğitim verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait doğruluk skoru	48
Şekil 27. Test verisi üzerinde oluşturulan tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait hata matrisi	48
Şekil 28. Test verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait ROC eğrisi	48
Şekil 29. Tür seviye OTU tabloları AdaBoost algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi	49
Şekil 30. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait Hamming kayıp skoru	49
Şekil 31. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait doğruluk skoru	50
Şekil 32. Test verisi üzerinde oluşturulan birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait hata matrisi	50
Şekil 33. Test verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait ROC eğrisi	50
Şekil 34. Birleştirilmiş OTU tabloları AdaBoost algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi	51
Şekil 35. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait Hamming kayıp skoru	51
Şekil 36. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait doğruluk skoru	52
Şekil 37. Test verisi üzerinde oluşturulan birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ai hata matrisi	52
Şekil 38. Test verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait ROC eğrisi	52
Şekil 39. Tür seviye OTU tabloları Rastgele orman algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi	53

KISALTMALAR, SİMGELER ve FORMÜLLER DİZİNİ

Kısaltmalar

CH	Crohn Hastalığı
ASV	Amplicon Sekans Varyant
QIIME2	Mikrobiyal Ekolojiye Nicel İlgörüler-2
DNA	Deoksiriboz Nükleik Asit
DNABERT	DNA Çift Yönlü Kodlayıcı Gösterimleri
ROC-AUC	ROC Eğrisi Altında Kalan Alan
rRNA	Ribozomal Ribonükleik Asit
NOD2	Nükleotid Bağlayıcı Oligomerizasyon Alanı İçeren Protein 2
LRRK2	Lösin Açısından Zengin Tekrar Kinaz 2
IRGM	Bağıışıklık İlişkili GTPaz M
JAK2	Janus Kinaz 2
ATG16L1	Otofaji ile İlişkili 16-benzeri 1
İBH	İnflamatuvar Bağırsak Hastalığı
ER	Endoplazmik Retikulum
ULK1	Unc-51 Benzeri Kinaz 1
MTMR3	Miyotübülün İlişkili Protein 3
XBP1	X-Box Bağlayıcı Protein 1
ORMDL3	ORM1 Benzeri Protein 3
NADPH	Nikotinamid Adenin Dinükleotid Fosfat
CYBA	Sitokrom B-245 Alfa Zinciri
CYBB	Sitokrom B-245 Beta Zinciri
NCF1	Nötrofil Sitoplazmik Faktör 1
NCF2	Nötrofil Sitoplazmik Faktör 2
NCF4	Nötrofil Sitoplazmik Faktör 4
ASCA	IgA Anti-Saccharomyces Cerevisiae Antikorları (İmmünoglobulin A)
IgG	İmmünoglobulin G
pANCA	Perinükleer Anti-Nötrofil Sitoplazmik Antikorlar
ÜK	Ülseratif Kolit
SBCE	İnce Bağırsak Kapsül Endoskopi
IL-17	İnterlökin 17

ROS	Reaktif Oksijen Türleri
BUS	Bağırsak Ultrasonu
CTE	Bilgisayarlı Tomografi Enterografi
MRE	Manyetik Rezonans Enterografi
CRP	C-Reaktif Protein
TNF	Tümör Nekroz Faktörü
IL-1 β	İnterlökin 6
İBS	İrritabl Bağırsak Sendromunu
WMS	Tüm Metagenome Sekanslama
OTU	Operasyonel Taksonomik Birim
ASV	Amplikon Sekans Varyant
MAG	Metagenom Birleştirilmiş Genom
GABA	Gamma-Aminobutirik Asit
MetaPhlAn	Metagenomik Filogenetik Analiz
SGB	Tür Düzeyi Genom Bileşenleri
İK	İskemik Kolit
TF-IDF	Terim Frekansı-Ters Doküman Frekansı
NCBI	Ulusal Biyoteknoloji Bilgi Merkezi
SRA	Sekans Okuma Arşivi
DADA2	Gürültü Bölücü Amplikon Gürültü Giderme Algoritması 2
MI-Mix	Manifold Örnek Karıştırma
C^2LR	Müfredat Karşıt Öğrenme
i-Mix	Örnek Karıştırma
SMOTE	Sentetik Azınlık Örneklem Arttırma Yöntemi
sPLS-DA	Seyrek Kısmi En Küçük Kareler Ayırıcı Analizi
SVM	Destek Vektör Makineleri
SimCLR	Görsel Temsillerin Karşıt Öğrenimi için Basit Çerçeve
AdaBoost	Uyarlanabilir Güçlendirme
RF	Rastgele Orman
NSAID	Nonsteroid Antienflamatuar İlaçlar

Formüller

$$Chao1 = S_{obs} + \frac{F_1^2}{2F_2}$$

$$e = H' / \ln S$$

$$H' = -\sum_{i=1}^S p_i \cdot \ln(p_i)$$

$$\sum_{i=1}^S p_i = 1$$

$$\text{Ağırlıklandırılmamış}_{AB} = \frac{\sum b_A \Delta b_B}{\sum b_A \cup b_B}$$

$$\text{Ağırlıklandırılmış}_{AB} = \frac{\sum_i^n (b_i \times |\frac{A_i}{A_T} - \frac{B_i}{B_T}|)}{\sum_{i=1}^n b_i}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|}$$

$$tfidf_{i,j} = tf_{i,j} * idf_i$$

$$h_i^m = \lambda g_m(x_i) + (1 - \lambda) g_m(\hat{x}_i)$$

$$v_i^{karıştırma} = \lambda v_i + (1 - \lambda) \hat{v}_i$$

$$\hat{\ell}(f_m(h_i^m), v_i^{karıştırma}) =$$

$$- \sum_{n=1}^B v_{i,n}^{karıştırma} \log \frac{\exp(\frac{s(f_m(h_i^m), f(x_{n+}))}{\tau})}{\sum_{j=1}^B \alpha_{ij+} \exp\left(\frac{s(f_m(h_i^m), f(x_{j+}))}{\tau}\right)}$$

$$\alpha_{ij+} = \frac{\exp(\frac{s(f_m(h_i^m), f(x_{j+}))}{\tau})}{\frac{1}{B-1} \sum_{k=1, k \neq i}^B \exp(s(f_m(h_i^m), f(x_{k+}))/\tau)}$$

$$\hat{L} = \frac{1}{B} \sum_{i=1}^B \hat{\ell}(f_m(h_i^m), v_i^{karıştırma})$$

$$\ell(f(x_i), v_i) = - \sum_{n=1}^B v_{i,n} \log \frac{\exp(s(f(x_i), f(x_{n+}))/\tau)}{\sum_{j \neq i} \alpha_{ij} \exp(s(f(x_i), f(x_j))/\tau)}$$

$$\alpha_{ij} = \frac{\exp(s(f(x_i), f(x_j))/\tau)}{\frac{1}{2B-2} \sum_{k \neq i, i+} \exp(\frac{s(f(x_i), f(x_k))}{\tau})}$$

$$p_{ij} = x_i + rand(0,1) \times (x_{ij} - x_i)$$

$$S = \prod_{k=1}^K |L^{(k)}|$$

ÖZET

16S rRNA Gen Dizileme Verilerini Kullanarak Pediatrik Crohn Hastalığının Tanısında Makine Öğrenmesi Modellerinin Uygulanması

Kronik bir gastrointestinal inflamatuvar hastalık olan Crohn hastalığı, sıklığı ve prevalansı gün geçtikçe artmakta olan inflamasyon, fibrotik darlıklar, enterik fistül ve bağırsak neoplazisi gibi sorunlara sebep olabilen bir hastalıktır. Crohn hastalığı tanısı günlük pratikte hastanın tüm şikayetleri, muayene bulguları, laboratuvar sonuçları, kolonoskopi ve görüntüleme bulguları ve patolojik veriler dikkate alınarak teşhisi konulmaktadır. Literatürde yapılan çalışmalar incelendiğinde pediatrik grupta yer alan Crohn hastalarında büyüme ve gelişim bozukluğu sık görülmektedir. Hastalığın erken teşhisi gittikçe önem kazanmaktadır. Tez çalışmasında NCBI PRJEB13679 numaralı projeden alınan 765 hastaya ait Illumina MiSeq verisi kullanılmıştır. Pediatrik Crohn hastaları ve kontrol grubu arasında sınıflama yapabilmek için ileum ve rektum verisi kullanılmıştır. Çalışmada kullanılmış olan 16S rRNA geni bakteri ve arkelerde yüksek oranda korunmuştur. Bununla birlikte türe özgü farklılıkların belirlenebileceği (V1-V9) dokuz hiper değişken bölge vardır. Bu sebeple 16s rRNA taksonomik parmak izi olarak değerlendirilmektedir. Tez çalışmasında Crohn hastalığı teşhisi için 16S rRNA gen dizilemesine ait sekans verisi dikkate alınarak 8 farklı temsil vektörü ve 4 farklı sınıflama modeli kullanılmıştır. Çalışmada 4-mer, 5-mer, 6-mer temsil vektörlerinin yanında QIIME2 aracı kullanılarak oluşturulan tür seviyesinde taksonomik ataması yapılan ASV tabloları ve cins seviye ataması yapılan ASV tabloları ve bu tabloların birleştirilmesi ile oluşturulan temsil vektörlerin yanında DNABERT-S ön eğitilmiş modeli kullanılarak oluşturulan temsil vektörleri kullanılmıştır. En iyi sonuçlar uyarlanabilir güçlendirme algoritmasının birleştirilmiş (cins seviye ASV tablosu + tür seviye ASV tablosu) temsil vektörleri ile kullanılmasıyla elde edilmiştir (Doğruluk 0.86, Ağırlıklandırılmış-Duyarlılık 0.86, Ağırlıklandırılmış-Kesinlik 0.85 ve Ağırlıklandırılmış-AUC skor 0.90, Ağırlıklandırılmış-F1 skor 0.85). Geliştirilen modelin Crohn hastalığının erken teşhisine yardımcı olarak tedavi süreçlerinin başlatılmasında oldukça önemli olması beklenmektedir.

Anahtar Kelimeler: Crohn hastalığı, Derin Öğrenme, Makine Öğrenmesi, OTU Temsilleri, 16S rRNA

ABSTRACT

Implementation of Machine Learning Models for the Diagnosis of Pediatric Crohn's Disease Using 16S rRNA Gene Sequencing Data

Crohn's disease, a chronic gastrointestinal inflammatory disease, is a disease that can cause problems such as inflammation, fibrotic strictures, enteric fistula and intestinal neoplasia, the frequency and prevalence of which are continue to increase. Crohn's disease is diagnosed by considering all the patient's complaints, examination findings, laboratory results, colonoscopy and imaging findings, and pathological data. When the studies in the literature are examined, growth and developmental disorders are frequently seen in Crohn's patients in the pediatric group. Early diagnosis of the disease is becoming increasingly important. Illumina MiSeq data from 765 patients obtained from the NCBI project number PRJEB13679 were used in the thesis study. Ileum and rectum data were used to classify pediatric Crohn's patients and the control group. The 16S rRNA gene used in the study is highly conserved in bacteria and archaea. However, there are nine hypervariable regions (V1-V9) where species-specific differences can be determined. For this reason, 16s rRNA is considered as a taxonomic fingerprint. In the thesis study, 8 different representation vectors and 4 different classification models were used to evaluate the sequence data of 16S rRNA gene sequencing for the diagnosis of Crohn's disease. In the study, in addition to 4-mer, 5-mer, 6-mer representation vectors, ASV tables with taxonomic assignment at the species level created using the QIIME2 tool, ASV tables with genus assignment, and representation vectors created by merging these tables, as well as representation vectors created using DNABERT-S pretrained model were used. The best results were obtained by using the Adaptive Boosting algorithm with combined (genus level ASV table + species level ASV table) representation vectors (Accuracy 0.86, Weighted-Sensitivity 0.86, Weighted-Precision 0.85 and Weighted-AUC score 0.90, Weighted-F1 score 0.85). The developed model is expected to be very important in initiating treatment processes by helping in the early diagnosis of Crohn's disease.

Keywords: Crohn's disease, Deep Learning, Machine Learning, OTU Representations, 16s rRNA

1. GİRİŞ ve AMAÇ

Crohn hastalığı (CH), inflamatuvar özellikler taşıyan, remisyon ve relapslar şeklinde ilerleyen, gastrointestinal bağışıklık sistemi hastalığıdır. Etiyopatogenezi kesin olarak bilinmemekle birlikte, 17. kromozomda 71 duyarlılık lokusuna sahip olduğuna dair kanıtlar bulunmaktadır (1). Genom çapında ilişki çalışmaları, NOD2, LRRK2, IRGM, JAK2 ve ATG16L1 gibi bakteriyel algılama ve bağışıklık sistemiyle ilişkili çok sayıda aleli tanımlamıştır (2). Bu hastalık, bağırsak daralması, perforasyon ve fistül oluşumuna yol açabilen karakteristik yamalı bağırsak iltihabına sebep olur ve genellikle terminal ileum ve proksimal kolonu etkiler. İleokolonoskopi ve histolojik bulgular tanı için altın standart olup, mikro besin eksiklikleri, transmural hastalığın şiddeti ve psikososyal değerlendirmelerle hastalık teşhisi tamamlanır (3). Crohn hastalarının %25'i çocukluk , veya ergenlik döneminde ortaya çıkmaktadır (4). Crohn hastalığının sık görülen semptomlar arasında ishal, karın ağrısı, mide bulantısı ve kusma yer alırken, kilo kaybı, ateş ve yorgunluk gibi sistemik belirtiler de mevcuttur (5).

CH, aynı zamanda deri ve mukoza semptomları (eritema nodozum, oral mukozada aftlar), göz semptomları (konjonktivit, iris iltihabı), eklem semptomları (artrit, ankilozan spondilit), kardiyovasküler semptomlar (perikardit, vaskülit), anemi, obstrüktif akciğer hastalığı, hepatit, pankreas stenozu gibi bağırsak dışı birçok semptoma yol açabilmektedir. Bu semptomlar hastalığın ilerlemesi ve kötü prognozu ile ilişkilidir (6).

CH'nin tedavisi genellikle metabolik aktiviteyi inhibe ederek hastalığın ilerlemesini durdurmayı amaçlar, ancak komplikasyonlar sindirim, osteoartiküler, dolaşım, sinir, boşaltım ve üreme sistemlerini etkileyebilmektedir. En yaygın komplikasyonlardan biri olan ileit, şiddetli karın ağrısı, kanama ve ishal ile karakterizedir. Diğer komplikasyonlar arasında ön üveit, kolelitiazis, sklerit, nefrolitiazis, osteoporoz ve venöz tromboembolizm bulunmaktadır (6).

Ergen başlangıçlı CH, yetişkin başlangıçlı hastalarla daha çok benzerlik gösterse de, yetersiz beslenme, büyüme bozuklukları, ergenlik gecikmesi ve psikososyal zorluklar gibi özel durumlar mevcuttur (7).

Ergenlik çağında başlayan Crohn hastalığı, yetişkinlerdeki fistülize veya striktürel hastalık seyrine benzemektedir. Erken başlangıç nedeniyle, 30 yaşına kadar cerrahi rezeksiyon geçirme riski pediatrik hastalarda 48 ± 5 , yetişkin başlangıçlı hastalarda ise 14 ± 2 'dir (8).

Mikrobiyata verisinin hastalık teşhisinde, sürecin takip edilmesinde kullanılması, maliyet açısından etkin bir yöntemdir. Mikrobiyata verisinin alındığı bölgeye bağlı olarak non-invaziv ya da düşük miktarda biyopsi yapılmasını gerektirebilmektedir (9). Crohn hastalığının erken teşhisi özellikle pediatrik hastaların büyüme ve gelişme bozukluklarının önlenmesi için oldukça önemlidir.

Tez çalışması kapsamında, Crohn hastalığı teşhisi için 16S rRNA gen sekanslama verisi dikkate alınarak, 8 farklı temsil vektörü ve 4 farklı sınıflama modeli kullanılmıştır. Bu temsil vektörleri arasında 4-mer, 5-mer, 6-mer dizileri ile QIIME2 aracı kullanılarak oluşturulan tür ve cins seviyesinde taksonomik atama yapılan ASV (Amplicon Sekans Varyant) tabloları ve bu tabloların birleştirilmesiyle elde edilen vektörler yer almıştır. Ayrıca, DNABERT-S kullanılarak oluşturulan temsil vektörleri kullanılmıştır. En iyi sonuçlar uyarlanabilir güçlendirme algoritmasının birleştirilmiş (cins seviye ASV tablosu + tür seviye ASV tablosu) temsil vektörleri ile kullanılmasıyla elde edilmiştir. Bu yaklaşım, farklı taksonomik seviyelerdeki bilgilerin entegre edilmesiyle daha zengin ve ayrıntılı bir veri temsili sağlamış, modelin performansını artırmıştır. Sonuç olarak, Uyarlanabilir Güçlendirme Algoritmasının birleştirilmiş ASV temsilleriyle sınıflandırılmasında 0.86 doğruluk, 0.86 ağırlıklandırılmış duyarlılık, 0.85 ağırlıklandırılmış kesinlik, 0.90 ağırlıklandırılmış AUC skoru ve 0.85 ağırlıklandırılmış F1 skoru elde edilmiştir.

2. GENEL BİLGİLER

2.1. Crohn Hastalığı

Crohn hastalığı, remisyon ve relapslar şeklinde ilerleyen, ekstraintestinal bulgular taşıyan, inflamatuvar özellik gösteren gastrointestinal bağışıklık sistemi hastalığıdır (10). Kesin etiyopatogenezi tam olarak bilinmese de, 17 kromozomda 71 duyarlılık lokusu olduğuna dair kanıtlar mevcuttur (1). Genom çapında ilişki çalışmalarında Crohn hastalığına özgü çok sayıda alel tanımlanmıştır. Tanımlanan genlerin çoğu, NOD2, LRRK2, IRGM, JAK2 ve ATG16L1 gibi bakteriyel algılama ve doğuştan gelen bağışıklık sistemi ile ilişkilidir (2). Bu durum ağızdan anüse kadar herhangi bir yerde bağırsak daralmasına, perforasyona ve/veya fistül oluşumuna ilerleyebilen yamalı bağırsak lümenal iltihabına neden olmaktadır (3). Genellikle terminal ileum ve proksimal kolonu etkileyen Crohn hastalığı görünüm olarak kesikli ve düzensiz, segmental ve transmuraldır. İleokolonoskopi ve histolojide karakteristik bulguların belirlenmesi tanısal altın standart olmakla birlikte, tam bir değerlendirme için kesitsel görüntüleme ve psikososyal değerlendirme gereklidir (3). Crohn hastalığına sahip hastaların %25'ine yakını çocukluk veya ergenlik döneminde ortaya çıkmaktadır (4,11). Montreal Sınıflamasına göre Crohn hastalığı sınıflaması Tablo 1’de verilmiştir.

Tablo 1. Montreal sınıflamasına göre crohn hastalığı (Silverberg ve ark.'dan, 12)

Tanı Yaşı	A1, ≤ 16
	A2, $\geq 17-40$
	A3, >40
Hastalığın Yeri	L1, Terminal İleum
	L2, Kolonik
	L3, İleokolonik
	L4, Üst gastrointestinal Sistem
Hastalık Davranışı	B1, Non striktüran, non penetran
	B2, Striktüran
	B3, Penetran

Türkiye’de CH için İBH insidansı 1.4/100000 olarak bildirilirken, prevalansı 130/100000 olarak bildirilmiştir (13).

2.1.1. Klinik Bulgular Semptomlar

Sık görülen semptomlar ishal, karın ağrısı, mide bulantısı veya kusma iken kilo kaybı, ateş ve yorgunluk sistemik belirtilerdir (5, 14). Hastaların %3'ünde yoğun gastrointestinal kanama görülebilmekle birlikte kabızlık da görülebilmektedir. Megakolon toksikum, %1.6-6 oranında hastada görülen, bağırsakta genişleme, duvar incilmesi ve septik durum ile karakterize bir komplikasyondur. Şiddetli Crohn hastalığı perforasyonlar, peritonit ve yoğun gastrointestinal kanama gibi ciddi komplikasyonlar oluşturabilmektedir (6, 15). Ayrıca bağırsak lümeninin daralması, kısalması ve displazi gibi komplikasyonlar kolorektal kanser riskini arttırmakla birlikte genellikle hastalığın yedinci yılından sonra görülmektedir, özellikle pankolitli hastalarda kolorektal kanser riski 20 yıl içinde %25 oranında görülmektedir (6).

Deri ve mukoza semptomları (eritema nodozum, oral mukozada aftlar), göz semptomları (konjonktivit, iris iltihabı), eklem semptomları (artrit, ankilozan spondilit), kardiyovasküler semptomlar (perikardit, vaskülit), anemi, obstrüktif akciğer hastalığı, hepatit, pankreas stenozu gibi bağırsak dışı semptomlar, hastalığın ilerlemesi ve kötü prognoz ile ilişkilidir (6).

CH'nin tedavisi genellikle metabolik aktiviteyi inhibe ederek hastalığın ilerlemesini durdurulması temeline dayanmaktadır. Ancak, komplikasyonlar sindirim, osteoartiküler, dolaşım, sinir, boşaltım ve üreme sistemlerini etkileyebilmektedir. En yaygın komplikasyonlardan biri olan ileit şiddetli karın ağrısı, kanama ve ishal ile karakterizedir. Diğer komplikasyonlar arasında ön üveit, kolelitiazis, sklerit, nefrolitiazis, osteoporoz ve venöz tromboembolizm bulunmaktadır (6).

CH'li hastalarda ince bağırsak tutulumu yaklaşık %15, kolon tutulumu %20 ve anorektal bölge tutulumu %15 iken vakaların yarısından çoğu ileum ve çekum tutulumunu içermektedir. Hastalığın karakteristik özelliği, aralıklı granümatöz lezyonlar içerir. Tutulum submukozada olmasına rağmen, bağırsağın tüm katmanları etkilenmektedir. Bağırsak yüzeyi, submukozal ödem alanlarıyla çevrili yarıklar ve çatlaklardan kaynaklanan "arnavut kaldırımı" görünümüne sahiptir. Submukozal tabakada belirgin inflamatuvar ve fibrotik değişiklikler mevcuttur. Bağırsak düz kasları nispeten daha az etkilenmekle birlikte bir süre sonra bağırsak duvarı kalınlaşır ve genellikle esnekliğini kaybetmektedir (13).

2.1.2. Risk Faktörleri

Hastalık etimolojisine etki eden çevresel ve genetik faktörler birleşerek hastalık risk faktörlerini oluşturmaktadır.

2.1.2.1. Çevresel Faktörler

Sigara içme, erken dönemde dumana maruz kalma ve pasif içiciliğin CH riskini arttırdığı bilinmektedir (16). Diyet, bağırsak disbiyozu ile karakterize bir hastalık olan CH hastalığı için bağırsak mikrobiyotasını etkileyen oldukça önemli bir çevresel faktördür (17). Yeni teşhis edilen pediatrik Crohn hastalarının, inflamatuvar bağırsak hastalığı olmayan kontrollere göre daha düşük miktarda meyve ve sebze tükettiği görülmektedir (18). Uzun süre lif ağırlıklı beslenmenin Crohn hastalığı riskini %40 azaltmada etkili olduğu bilinmektedir (19). Ayrıca, Akdeniz diyeti geç dönem başlangıçlı hastalarda düşük risk ile ilişkilendirilmiştir (17, 20).

Antibiyotiklere maruz kalma, çocuklarda Crohn hastalığı riskini belirgin şekilde artırmaktadır (21). Çevresel faktörler popülasyonlara bağlı olarak değişebilmektedir bu durum hastalık oluşumunu önlemek için popülasyon bazlı müdahalelerin yapılmasını gerektirmektedir (17). Crohn hastalığı çevresel risk faktörleri Tablo 2’de gösterilmiştir.

Tablo 2. Crohn hastalığı çevresel risk faktörleri (Roda ve ark.'dan,17)

Çevresel Faktörler	İlişki
Sigara kullanımı	Hastalığın başlangıcı ve daha kötü hastalık seyri ile güçlü pozitif ilişki (22, 23)
Düşük D vitamini beslenme	Hastalık başlangıcı ve seyri ile pozitif ilişki (24, 25)
Menopoz sonrası hormon kullanımı	Hastalık başlangıç ve seyri ile ilişkisi yok (22, 23)
Apandisit ameliyatı	Hastalık başlangıcı ile pozitif etki, hastalık seyri ile ilişki yok (22, 23)
Oral kontraseptif kullanımı	Hastalık başlangıcı ile güçlü ilişki, hastalık seyri ile ilişki yok (22)
NSAID kullanımı	Hastalık başlangıcı ile pozitif ilişki, hastalık seyri ile güçlü pozitif ilişki (22, 23)
Depresyon ve psikososyal stres	Hastalık başlangıcı ile pozitif ilişki, hastalık seyri ile ilişki yok (22, 23, 26)
Düşük diyet lifi	Hastalık başlangıcı ile negatif ilişki, hastalık seyri ile ilişki yok (24, 25)
Yüksek protein diyeti	Hastalık başlangıcı ya da seyri ile ilişki yok (24, 25)
Yüksek yağ diyeti	Hastalık başlangıcı ya da seyri ile ilişki yok (24, 25)

2.1.2.2. Genetik Faktörler

Çevresel faktörlere kıyasla hastalık riskini belirleyen genomik varyasyonlar daha ayrıntılı olarak incelenmiştir. Crohn hastalığı için monozygotik ikizlerde %58.3 ve dizigotik ikizler arasında %0 proband uyum oranı tespit edilmiştir (27). Lökosit antijen bölgesi dışında CH'ye özgü 37 gen bölgesi tanımlanmıştır (28-30). Crohn hastalığında NOD2 varyantları hastalık yeri ile ilişkili iken striktür ile ilişkili olmadığı bilinmektedir (31).

Genom çapında ilişkilendirme çalışmalarında, Crohn hastalığı ile ilişkili çok sayıda gen tanımlanmıştır. Otofajide ATG16L1, IRGM, ULK1, LRRK2 ve MTMR3 genleri CH ile ilişkilendirilirken, hücre içi bakteriyel algılamada NOD2 geni ve endoplazmik retikulum (ER) stresinde XBP1 ve ORMDL3 genleri CH ile ilişkili olarak tanımlanmıştır (32). ATG16L1 Crohn hastalığı ile ilişkili bir genetik risk faktörüdür. Bu risk aleli, Crohn hastalarının yaklaşık %30'unda homozigottur. Bu alel varyantı otofajiyi azaltmaktadır. Ayrıca bağırsak epitelini spontan apoptozdan korumaktadır (33, 34).

Doğuştan gelen bağışıklık sistemi dikkate alındığında fagositik fonksiyondaki kusurlar ve immün yetmezlik unsurları Crohn hastalığı patogenezinde oldukça önemlidir (35). Çok erken başlangıçlı Crohn hastalığında bozulmuş nötrofil NADPH oksidaz aktivitesi tanımlanmıştır (36). CH'de CYBA, CYBB, NCF1, NCF2 ve NCF4'te 26 anlamsız mutasyon tespit edilmiştir. Bu genlerde mutasyon taşıyan pediatrik hastalarda perianal hastalık ($p=0.0008$) ve striktür komplikasyonları ($p=0.002$) olma olasılığının 3 kat daha fazla olduğu tespit edilmiştir (37).

2.2. Pediatrik Crohn Hastalığı

Pediatrik CH insidansı dünya genelinde 100.000'de 2.5-11.4 arasında değişmektedir (38). Hastalık prevalansı ise 20 yaş altı bireylerde 100000'de 31-71 arasında değişmektedir (7, 39). CH'li hastaları yaklaşık %15'i 20 yaşından önce teşhis edilmektedir (38, 40, 41).

Yetişkin başlangıçlı CH ile kıyaslandığında hastalık başlangıcı 2 yaşından küçük olan infantil ve 6 yaşından küçük olan çok erken başlangıçlı hastalıklarda daha agresif bir fenotip sergilemektedir. Bu gruplar için alternatif hastalık yönetim stratejilerinin belirlenmesi gerekmektedir (42). Ergen başlangıçlı CH yetişkin başlangıçlı hastalara daha çok benzerlik göstermekle birlikte yetersiz beslenme, büyüme bozuklukları, ergenlik gecikmesi ve psikososyal ve eğitim zorlukları gibi özel durumlar mevcuttur (7, 42, 43).

Ergenlik çağında başlayan Crohn hastalığı, yetişkin fistülize veya striktürel hastalık seyrine benzer. Erken başlangıç nedeniyle, 30 yaşına kadar cerrahi rezeksiyon geçirme riski pediatrik hastalarda 48 ± 5 yetişkin başlangıçlı hastalarda 14 ± 2 olarak görülmektedir (8, 42).

ASCA IgA veya IgG pozitif pediatrik CH'de terminal ileal veya ileoçekal hastalık prevalansı ve ileride cerrahi girişim ihtiyacı yüksektir. pANCA pozitif hastalarda pankolit veya sol taraflı hastalığa sahip olma olasılığı daha yüksektir (5). Daha önce tedavi uygulanmamış pediatrik hastalardan alınan ileal doku örneklerinde, IL-17, NOD ve Oncostatin-M sinyalini yönlendiren genlerin önemli ölçüde yukarı regüle edilmiş olduğu görülmektedir (44). Pediatrik CH hastalarında daha agresif bir hastalık seyri ve nötrofiller tarafından azaltılmış ROS üretimi ile ilişkilendirilen NADPH oksidazları kodlayan genlerde anlamsız mutasyonlar belirlenmiştir (37). İnflamatuar fenotipli pediatrik CH hastalarında striktür ve delici fenotiplere göre kolon yerleşiminde elafin, katelisin ve ileal yerleşimde α -defensinler artmış olarak bulunmaktadır (45).

2.3. Crohn Hastalığı Tanı ve Görüntüleme Yöntemleri

Crohn hastalığı tanı ve teşhisinde kullanılan yöntem ve görüntüleme teknikleri tedavi süreçlerinin belirlenebilmesi için oldukça önemlidir.

2.3.1. Endoskopik Görüntüleme

İleokolonoskopi, Crohn hastalığının tanısında altın standart olarak kabul edilmektedir ve histolojik inceleme için biyopsi alınmasına olanak tanımaktadır (46).

Güncel kılavuzlar, ileokolonoskopi ve radyolojik görüntüleme sonuçları negatif olan ancak CH şüphesi güçlü bir şekilde devam eden hastalarda ince bağırsak kapsül endoskopisini (SBCE) önermektedir (47). Obstrüktif semptomlar veya stenoz varlığında kapsül retansiyonu riski bulunur; bu risk yerleşik CH hastalarında %13, şüpheli CH hastalarında ise %1.6 olarak bildirilmiştir (48). Bu tür durumlarda, öncelikle ince bağırsak kesitsel görüntüleme yöntemleri tercih edilmelidir. Cihaz destekli enteroskopi ise yalnızca histolojik tanı veya endoskopik tedavinin gerekli olduğu durumlarda kullanılır. Bununla birlikte, invaziv bir prosedür olması, zaman alıcı olması ve yüksek maliyeti nedeniyle, ince bağırsak CH tanısında birinci basamak yöntem olarak genellikle önerilmez (48).

2.3.2. Kesitsel Görüntüleme

Kesitsel görüntüleme yöntemleri (bağırsak ultrasonu (BUS), BT enterografi (CTE) ve manyetik rezonans enterografi (MRE)), Crohn hastalığının transmural yapısından kaynaklanan hastalık yayılımını ve inflamatuvar komplikasyonları (örneğin, stenoz, fistül ve apse) değerlendirmek için kritik bir role sahiptir (47). BUS, CTE ve MRE, CH tanısı ve komplikasyonların tespitinde doğruluğu yüksektir (49).

CTE'nin, radyasyon yayması oldukça önemli bir dezavantajdır (17, 50). CTE'nin Crohn hastalığını tespit etme duyarlılık ve özgüllük değerleri %80'in üzerindedir. Fistüller ve stenozların belirlenmesinde yüksek doğruluk sunmaktadır (17, 51, 52). MRE ise stenozların tespitinde %89 duyarlılık ve %94 özgüllük değerlerine sahiptir (17, 53, 54). Pelvik MRE, perianal fistül ve apseleri değerlendirmek için yaygın olarak kullanılır (17, 55).

BUS, radyasyona sebep olmayan non-invaziv bir yöntemdir. Bazı sınırlamaların yanında bağırsak duvarını ve ekstraluminal bulguları değerlendirmede güvenilir sonuç vermektedir (54).

2.3.3. Histoloji

Endoskopik biyopsi örnekleri veya rezeksiyon materyallerinin histolojik analizi, Crohn hastalığı teşhisinde altın standart olmasına rağmen CH'nin kesin bir histolojik özelliği yoktur. Fakat tanıda yardımcı olabilecek mikroskopik bulgular arasında fokal kronik inflamasyon, odaksal kript düzensizliği, kript hasarıyla ilişkili olmayan granülomlar ve terminal ileumda anormal villöz mimari yer almaktadır (17, 56). Histolojik bulgular spesifik olmasa bile, klinik, endoskopik ve görüntüleme bulgularının değerlendirilmesi, CH tanısını destekleyebilmektedir.

2.3.4. Biyobelirteç

Biyobelirteçler, CH'nin şiddetini ve ilerleyişini değerlendirmek, erken müdahale süreçlerini yönlendirmek ve tedavi yanıtını izlemek açısından önemlidir. Ayrıca cerrahi sonrası nüks riskini tahmin etmekte de kullanılabilmektedir (17).

2.3.5. Serolojik Belirteçler

Otoantikörler, örneğin pANCA ve antimikrobiyal antikörler (örneğin, ASCA), anti-Pseudomonas fluorescens ilişkili sekans I2 antikörleri, anti-dış membran porin C

antikorları, anti-CBir1 antikorları), CH tanısında potansiyel biyobelirteçler arasında yer almaktadır (17, 57). Ayrıca, bakteriyel hücre duvarındaki glikan epitoplarına karşı oluşan dolaşımdaki antikorlar (örneğin, anti-mannobioside, anti-laminaribioside, anti-chitobioside karbonhidrat antikorları ve anti-laminarin antikorları), CH tanısında yardımcı olabilmektedir (17, 58). Ticari olarak en yaygın kullanılan serolojik marker olan ASCA CH hastalarının %60-70'inde pozitiflik göstermektedir. pANCA ise CH vakalarının %10-15'inde saptanmaktadır (17, 57). Ayrıca, pANCA pozitifliği görülen CH hastalarının klinik fenotipleri genellikle Ülseratif Koliti andırmaktadır (17, 59).

2.3.6. C-Reaktif Protein (CRP)

Hepatositler tarafından sentezlenen CRP, TNF, IL-1 β ve IL-6 gibi proinflamatuvar sitokinlere yanıt olarak üretilmesi sebebiyle CH'de hastalık aktivitesini izlemek için kullanılabilir (17, 60, 61). CH'nin aktif döneminde hastaların üçte birinde aktif hastalık bulunmasına rağmen CRP seviyeleri normal düzeydedir, başka bir üçte birlik grupta ise klinik olarak remisyon döneminde olmalarına karşın CRP değerleri yüksektir (61).

2.3.7. Dışkıdaki Kalprotektin

Dışkı biyobelirteçleri, CH aktivitesinin belirlenmesi ve inflamatuvar kolitlerin tanısında kullanılan invaziv olmayan yöntemlerdir (17, 61, 62). İnflamasyon sırasında gastrointestinal sisteme salınan kalprotektin, nötrofillerin sitozolik proteinlerinin %60'ını oluşturur. Kalsiyum bağlayıcı antimikrobiyal protein kompleksidir (17). Dışkı kalprotektin düzeyleri, CH tanısında yüksek duyarlılık ve özgüllüğe sahiptir (17, 62, 63). Kalprotektin genellikle CRP ile birlikte hastalık aktivitesi, nüksler, tedavi yanıtı ve hasta bildirim sonuçlarını değerlendirmek için kullanılmaktadır (17, 62).

2.3.8. Yeni Potansiyel İnvazif Olmayan Belirteçler

Bağırsak mikrobiyotasındaki dengesizlik (disbiyoz), Crohn hastalığının gelişiminde önemli bir faktör olduğu için, dışkı ve serum mikrobiyal belirteçlerinin tanısal değerine yönelik birçok çalışma yapılmıştır. Asya popülasyonunda yapılan çok merkezli bir prospektif çalışmada (95 CH hastası, 81 ÜK hastası, 65 İBS hastası ve 105 sağlıklı birey), dışkıda *Fusobacterium nucleatum* seviyelerindeki artış ve *F. prausnitzii* düzeylerindeki düşüşün, CH hastalarını sağlıklı bireylerden ayırt etmek için değerli bir belirteç olduğu bulunmuştur (sağlıklı bireylerle karşılaştırıldığında eğri altındaki alan AUC 0.841, İBS ile

Bağırsağın farklı kısımlarında farklı fizyolojik koşullar mevcuttur. Bu farklılıklarla, bakteri ve mikroorganizma popülasyonunda değişim gözlemlenmektedir. Bağırsak mikrobiyomundaki düşük oksijen seviyeleri sebebiyle bakterilerin %99'u anaerobdur. Fakat, kör bağırsakta yüksek yoğunlukta aerobik bakteri bulunmaktadır. Bağırsakta altı ana şube Firmicutes, Bacteroidetes, Proteobacteria, Actinobacteria, Euryarchaeota ve Verrucomicrobia baskındır (68). İnsan bağırsak mikrobiyotasının altı ana şubesi ve baskın türler Tablo 3’de gösterilmiştir.

Tablo 3. İnsan bağırsak mikrobiyotasının altı ana şubesi ve baskın türleri (Madhogaria ve ark.'dan, 68)

Ana Şube	Tür	Yüzdelik
Firmicutes	Lactobacillus, Ruminococcus, Clostridium, Eubacterium, Faecalibacterium, Roseburia, Streptococcus	%90
Bacteroidetes	Bacteroides, Prevotella, Xylanibacte	
Proteobacteria	Escherichia, Desulfovibrio	%10
Actinobacteria	Collinsella, Bifidobacterium	
Euryarchaeota	Methanobrevibacter	
Verrucomicrobia	Akkermansia	

Bağırsak mikrobiyotası, hormon benzeri sinyal molekülleri üreten bir mikroorganizma topluluğundan oluşmaktadır ve bu moleküller kan dolaşımı aracılığıyla vücudun diğer bölgelerine taşınmaktadır. Artan kanıtlar, mikrobiyal kompozisyonun çevresel faktörlere göre değişkenlik gösterebileceğini ortaya koymaktadır. Mikrobiyal çeşitliliğin azalması, çeşitli sağlık sorunları ile bağlantılı iken, çeşitliliğin artması metabolik ve bağışıklık sisteminin işlevlerini iyileştirmektedir. Ancak, egzersiz ve bağırsak mikrobiyotası arasındaki etkileşimler üzerine yapılan kontrollü çalışmalar oldukça sınırlıdır (69, 70).

2.4.1. Mikrobiyom Profilleme

Mikrobiyom genetik sekansları 2 şekilde sekanslanabilir.

1. Marker genlerin hedeflenmiş sekanslanması (en genel olarak 16s rRNA genleri)
2. Örnekteki tüm genetik materyalin hedeflenmemiş sekanslanması (Tüm metagenome sekanslama (WMS))

2.4.1.1. Marker Gen Sekanslama Verisi

En çok kullanılan mikrobiyom profillemeye yaklaşımı 16S rRNA gen sekanslamasıdır. 16S rRNA sekanslama, verimli ve uygun maliyetlidir. Sekanslama sürecinde hataları en aza indirmek için biyoenformatik araçlar kullanılmaktadır. Marker gen sekanslama verileri belli bir veri setine özgü olduğu için çalışmalar arasında karşılaştırma yapılmasına uygun değildir. Sekanslama işlemi sonrasında benzer sekanslar operasyonel taksonomik birimler (OTU'lar) olarak kümelenebilirler. Sekanslar gürültüden arındırılarak ampikon sekans varyantlar (ASV'ler) elde edilir. ASV'ler, OTU'lardan daha iyi çözünürlüğe sahiptir. Bu nedenle sıfır yarıçaplı OTU'lar olarak da adlandırılırlar (71).

Gözlemlenen özellikler (OTU'lar veya ASV'ler) taksonomik ağaçlar ve filogenetik ağaçlar olarak düzenlenebilmektedir. Taksonomik ağaçlar, alem, şube, sınıf, takım, aile, cins ve tür sıralamasına göre daha alt organizasyonlara ayrılabilir (71).

2.4.1.2. Tüm Metagenom Dizileme Verileri (Shotgun Metagenomik Dizileme)

Genom bir organizmanın genetik materyalini metagenom ise örnekteki tüm mikroorganizmaların genomlarının koleksiyonunu ifade etmektedir. Gözlemlenen genom parçaları metagenom-birleştirilmiş genom (MAG) oluşturacak şekilde birleştirilirler. Taksonomik bolluğun ölçülmesi metagenomik sekanslama verilerinin Kraken 2 (72) veya MetaPhlAn 4 (73) gibi araçlar kullanılarak sağlanabilmektedir. Bu bağlamda oluşan tür düzeyi genom bileşenleri (SGB) hem bilinen bakteri türlerini hem de tür düzeyinde ilişkili bilinmeyen genom kümelerini ifade etmektedir (71).

2.5. Çeşitlilik Analizi

Bireydeki mikrobiyal çeşitlilik ve dağılımlarının incelenmesi hastalık teşhis, tanı ve tedavi süreçlerine olan yatının takip edilebilmesi açısından oldukça önemlidir.

2.5.1. Alfa Çeşitlilik Analizi

Literatürde, nitel tür tabanlı ölçümler Chao 1 indeksi (74), nicel tür tabanlı ölçümler Faith'in Filogenetik Çeşitliliği (75), Pielou Düzgünlük İndeksi (76), Shannon (veya Shannon Wiener) indeksi (77, 78) oldukça yaygın olarak kullanılmaktadır.

2.5.1.1. Chao1 İndeksi

Chao1 İndeksi, nadir türlerin bolluğunu dikkate alarak tür zenginliğini tahmin etmeye çalışmaktadır. Genellikle ekoloji ve mikrobiyom çalışmalarında kullanılmaktadır (79). Bu indeks “Eşitlik 1” ile hesaplanmaktadır.

$$Chao1 = S_{obs} + \frac{F_1^2}{2F_2} \quad (\text{Eşitlik 1})$$

Burada: S_{obs} : Gözlemlenen toplam tür sayısı, F_1 : Yalnızca bir kez gözlemlenen türlerin sayısı (tekil türler), F_2 : Yalnızca iki kez gözlemlenen türlerin sayısı (çiftli türler).

Mikrobiyom çalışmalarında düşük bolluk sınıflarını kapsayan veri kümelerinde oldukça faydalıdır (79).

2.5.1.2. Faith’in Filogenetik Çeşitliliği

Bir topluluktaki filogenetik ağaçtaki belirli taksonu kapsamak için minimum dal uzunlukları toplamıdır (75).

2.5.1.3. Pielou Düzgünlük İndeksi

Pielou Düzgünlük İndeksi [0,1] arasında değer alırken yüksek değerler yüksek çeşitlilik ve türler arasındaki bireylerin sayısındaki eşitliği ifade etmektedir (76). İndeksin hesaplanması “Eşitlik 2” ile gösterilmiştir.

$$e = H' / \ln S \quad (\text{Eşitlik 2})$$

H' : Shannon İndeks, S : Toplam tür sayısı

2.5.1.4. Shannon-Wiener Çeşitlilik İndeksi

Shannon-Wiener İndeksi, tür çeşitliliğinin hem zenginlik (tür sayısı) hem de bolluk (her bir türün oranı) bileşenlerini dikkate alır. İndeksin hesaplanması “Eşitlik 3” ile verilmiştir (80).

$$H' = - \sum_{i=1}^S p_i \cdot \ln(p_i) \quad (\text{Eşitlik 3})$$

Burada: S , topluluktaki toplam tür sayısını, p_i , i . türün toplam birey sayısına oranı (göreceli bolluk) “Eşitlik 4” ile gösterilmiştir.

$$\sum_{i=1}^S p_i = 1 \quad (\text{Eşitlik 4})$$

Shannon-Wiener İndeksi'nin H' değeri büyüdükçe, topluluğun tür çeşitliliği artmaktadır. Teoride H' değeri çok büyük değerlere ulaşabilirken uygulamada 5 değerini geçemediği görülmektedir (80).

2.5.2. Beta Çeşitlilik

Beta çeşitlilik, mikrobiyom topluluklarının kıyaslanmasında kullanılan genellikle uzaklık ölçümlerini dikkate alan bir çeşitlilik ölçümüdür (81). Çok sayıda beta çeşitlilik ölçümü vardır bunlara örnek olarak Jaccard Benzerlik İndeksi (82), Ağırlıksız UniFrac, Ağırlıklı UniFrac ölçümleri verilebilmektedir.

2.5.2.1. UniFrac Mesafesi

UniFrac mesafeleri, topluluklar arasındaki evrimsel farklılıkları değerlendirmek için kullanılmaktadır. Literatürde çok sayıda çalışma mevcuttur (83,84,85). UniFrac mesafesi, filogenetik ağaçtaki mesafeleri dikkate alarak takson grupları arasındaki farklılıkları değerlendirir. Bu mesafeler, kıyaslanan çevrelere özgü olan, ancak her iki çevre tarafından ortaklaşa paylaşılmayan dalların uzunluklarını temel alır. Ağırlıksız UniFrac mesafesi (83), yalnızca tür varlığı ve yokluğu bilgisini dikkate alır ve her iki topluluğa özgü dalların uzunluğunun bir kesrini hesaplar. Ağırlıklandırılmamış UniFrac Mesafesine ait formül “Eşitlik 5” ile verilmiştir (86).

A ve B iki örneği ifade etmek üzere b filogenetik ağacın dal uzunluklarının kümesini göstermektedir. Δ iki küme arasındaki simetrik farkı göstermektedir.

$$Ağırlıklandırılmamış_{AB} = \frac{\sum b_A \Delta b_B}{\sum b_A \cup b_B} \quad (\text{Eşitlik 5})$$

Bir topluluğa ait olup diğer topluluğa ait olmayan dalların uzunluklarının toplamı, her iki topluluğun da sahip olduğu dalların toplam uzunluğuna oranlanarak hesaplanır.

Buna karşılık, Ağırlıklı UniFrac Mesafesi (84), türlerin bolluk bilgilerini kullanmaktadır ve dalların uzunluğunu topluluklar arasındaki bolluk farklarına göre ağırlıklandırmaktadır. Filogenetik ağacın her dal uzunluğu 2 örnekteki bolluk oranına göre ağırlıklandırılır. A ve B iki örneği ifade etmek üzere b dal uzunluğu kümesini ifade etmektedir. $\frac{A_i}{A_T}$ ve $\frac{B_i}{B_T}$, b_i dal uzunluğu ile orantılı oransal bollukları ifade etmek üzere Ağırlıklı UniFrac Mesafesi “Eşitlik 6” ile hesaplanmıştır (86).

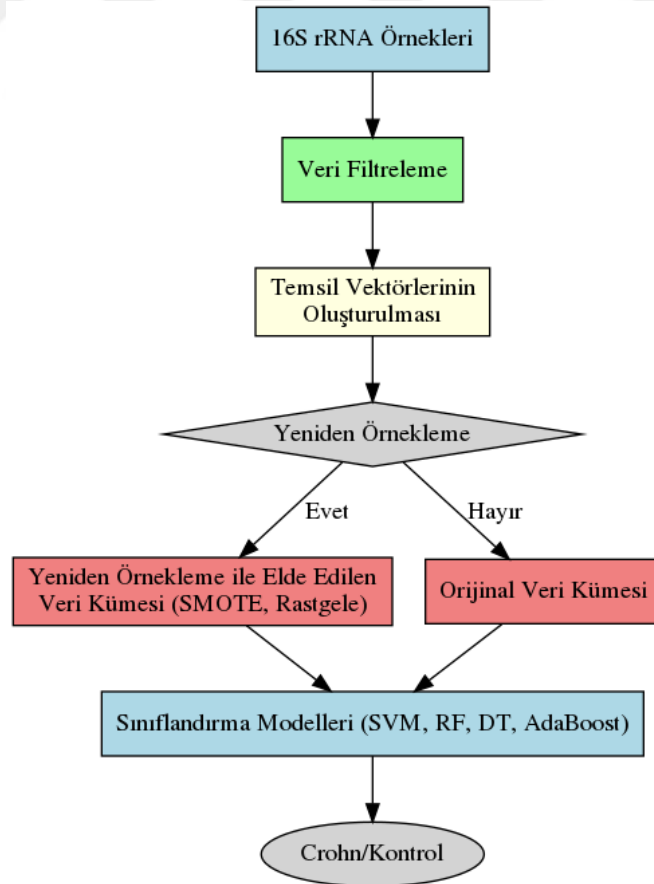
$$Ağırlıklandırılmış_{AB} = \frac{\sum_i^n (b_i \times |\frac{A_i}{A_T} - \frac{B_i}{B_T}|)}{\sum_{i=1}^n b_i} \quad (\text{Eşitlik 6})$$

Ağırlıksız UniFrac mesafesi, nadir türlerin bolluk değişimlerini algılamakta en etkili yöntemdir. Ağırlıklı UniFrac mesafesi, tür bolluğu değişimlerini değerlendirmede daha hassastır, özellikle de bol miktarda bulunan türlerdeki farklılıkları tespit etme konusunda oldukça etkilidir. Ancak, orta derecede bol olan türlerdeki değişiklikleri algılama açısından, hem ağırlıklı hem de ağırlıksız UniFrac mesafeleri sınırlı bir duyarlılığa sahiptir (87).



3. GEREÇ ve YÖNTEM

Çalışmanın yapıldığı veri seti NCBI'nın PRJEB13679 numaralı projesinden alınan 16S rRNA gen (V4) sekans verisidir (88). Veri seti üzerinde filtreleme işleminin gerçekleştirilmesinin ardından makine öğrenmesi tabanlı sınıflama işlemlerinin gerçekleştirilebilmesi için 16S rRNA sekans verisinden 8 farklı temsil vektör (4-mer, 5-mer, 6-mer, DNABERT-S ön eğitilmiş modeli kullanılarak 512 batch boyutu ile temsil vektörlerinin ortalamasının alınmasıyla oluşturulan temsil vektörü, DNABERT-S ön eğitilmiş modeli kullanılarak 512 batch boyutu ile temsil vektörlerinin maksimum değerlerinin alınması ile oluşturulan temsil vektörü, QIIME2 kullanılarak tür seviyesinde, cins seviyesinde, birleştirilmiş (tür + cins seviyesinde) temsil vektörleri) oluşturulmuştur. Sınıflama aşaması (SVM, RF, DT, AdaBoost) için model sonuçlarının iyileştirilebilmesi için orijinal veri kümesine ait sonuçların yanında çeşitli yeniden örnekleme yöntemleri (SMOTE, rastgele örnekleme) kullanılmıştır. Model sonuçları çeşitli performans metrikleri kullanılarak değerlendirilmiştir. Tez çalışmasına ait akış Şekil 2'de gösterilmiştir.



Şekil 2. Tez kapsamına ait akış

3.1. Veri Kümesinin Elde Edilmesi

NCBI PRJEB13679 numaralı projeden (88) alınan 1359 hastaya ait Illumina MiSeq (sürüm 2) platformunda 175 bp tek uç okumalarıyla oluşturulmuş olan 16S rRNA gen (V4) sekans verisi veri boyutunun büyüklüğü sebebiyle Truba platformuna sratoolkit.3.0.0-centos_linux64 kullanılarak kurulmuştur. SRA aracı kullanılarak verilerin indirilmesinin ardından indirilen .sra uzantılı veriler SRA aracı kullanılarak .fastq.gz formatına dönüştürülmüştür. Elimizde verilere ait barkod bilgisi mevcut olmaması sebebiyle dosyalarımıza ait isimlendirme yapısı Casava 1.8 single-end demultiplexed formatına uygun olarak düzenlenmiştir.

1359 örnekten oluşan veri kümesi Yaş değişkenine ait değerler 17 yaşından küçük ve eşit olacak şekilde filtrelenerek 986 hastaya ait meta veri ve sekans verisi elde edilmiştir. Filtreleme işlemi sonucunda, mevcut 906 pediatrik veri arasından veri yetersizliği nedeniyle İskemik Kolit (İK) ve Ülseratif Kolit (UK) hastalarına ait veriler çıkarılmış ve sınıflandırma işlemi Crohn hastalığı (CH) ve kontrol grubu arasında gerçekleştirilmiştir. Bu filtreleme işlemi sonrasında, 765 örnekten oluşan bir veri seti elde edilmiştir.

Tablo 4. Hastalıklara bağlı olarak 16s rRNA verisinin alındığı bölgelere göre değişen veri dağılımı

	İleum (611)	Rektum (295)	Dışkı (15)	Çekum (41)	Kolon (24)
CH (514)	314	154	11	23	12
Kontrol (315)	191	106	3	9	9
UK (105)	72	25	1	5	3
İK (52)	34	10	0	4	0

3.2. Temsil Vektörlerinin Oluşturulması

3.2.1. K-mer Temsilleri

K-mer'ler k uzunluğundaki oligomerlerdir. Genomun çeşitli işlevsel bölgelerini tanımak için kullanılabilmektedirler (89). Sınıflama aşamasına geçmeden önce ham metagenomik verinin incelenmesinde sekans verilerinin referans veriye hizalanmasına ait sayımlar yerine tüm verinin kullanıldığı k-mer sayımları kullanılmıştır. Tüm örneklerde

mevcut olduğu bilinen k-mer temsilleri için örnek bir ilişki matrisi Tablo 5’da gösterilmiştir (90).

Tablo 5. 4-mer,5-mer,6-mer yapısının oluşturulmasına ait sekans örneği

sekans örneği	4-mer sonucu	5-mer sonucu	6-mer sonucu
Atgcgcc	["atgc", "tgcg", "gcgc", "cgcc"]	["atgcg", "tgcgc", "tgcgc"]	["atgcgc", "tgcgcc"]

Tezde kullanılan k-mer temsil modelinde (91) öncelikle k-mer için k değerine bağlı olarak “atcg” sekansını kullanan kombinasyonlar üretilerek sözlük yapısı oluşturulmuştur. Ardından Terim Frekansı-Ters Doküman Frekans (Term frequency-Inverse Document Frequency, TF-IDF) (92) yaklaşımı kullanılarak veri kümesi vektörize edilmiştir. TF-IDF değeri ifadenin dökümanda görülme sıklığı olan terim frekansı ile ifadenin kaç dökümanda görüldüğünü dikkate alarak sık görülen ifadelerle düşük ağırlık vermeyi amaçlayan yaklaşımdır. TF hesaplanmasında kullanılan ifade “Eşitlik 7” ile verilmiş iken IDF değerinin hesaplanması “Eşitlik 8” ile gösterilmiştir.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad \text{Eşitlik 7)}$$

“Eşitlik 7” ’de $n_{i,j}$ kelimenin dokümanda görülmesi sayısını ifade ederken, $\sum_k n_{k,j}$ dokümandaki toplam kelime sayısını ifade etmektedir. “Eşitlik 8” ’de D toplam doküman sayısını, $|\{j: t_i \in d_j\}|$, t_i kelimesini içeren doküman sayısını ifade etmektedir (93). TF-IDF değerinin hesaplanması ise “Eşitlik 9” ile gösterilmiştir.

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad \text{(Eşitlik 8)}$$

$$tfidf_{i,j} = tf_{i,j} * idf_i \quad \text{(Eşitlik 9)}$$

K-mer sayım vektörleri birçok biyoenformatik alanında kullanılabilir. Örneğin 16S rRNA ve metagenom veri kümelerinin taksonomik sınıflandırılmasında kullanılmaktadır (94-97). Çalışmada 4-mer, 5-mer, 6-mer temsil vektörleri (91) kullanılarak oluşturulmuştur.

3.2.2. ASV Temsil Vektörlerinin Oluşturulması

QIIME2 (98) mikrobiyal toplulukların analizinde veri standartlaştırılmasını sağlayan ve veri analizi için çeşitli araçlar sunan bir araçtır. Bu protokolde, 16S rRNA gen sekansına ait Illumina platformunda oluşturulmuş olan çift uç okumaları içeren gen veri kümesine ait veriler kullanılmaktadır (99). Temsil vektörlerinin oluşturulması aşaması için hazırlanmış durumda bulunan FASTQ dosyaları python programlama dili kullanılarak işlenmiştir. Analiz aşamasına geçmeden önce verileri işlemede kullanılacak olan meta veriler QIIME2 tarafından kullanılabilir hale getirilmiştir. Metadata dosyasını uygun formatta kullanabilmek için Keemei (100) aracı kullanılmıştır. Meta verinin Keemei aracı ile doğrulanmasına ait tsv dosyası Şekil 3’de gösterilmiştir. Keemei, Konumsal Olarak Referanslandırılmış Genetik Veri (Spatially Referenced Genetic Data, SRGD) ve QIIME meta verilerini doğrulamak için kullanılan Google e-tablolar eklentisidir (100). Meta veri üzerinde bir hata olması halinde araç hatanın sebebini göstererek düzeltilmesine fırsat tanımaktadır.

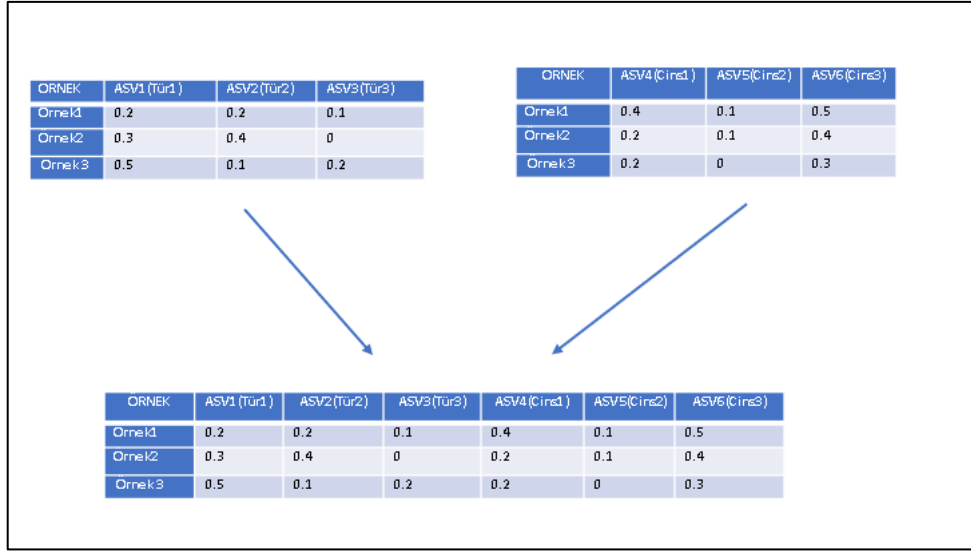
[illegible]

Şekil 3. Meta verinin Keemei aracı ile düzenlenmesi

QIIME2, qza uzantılı giriş ve çıkış dosyaları ile .qzv uzantılı görselleştirme dosya türlerini kullanmaktadır. Tez çalışmasında QIIME2 kullanılarak verilerde meydana gelen gürültü Bölücü amplicon gürültü giderme algoritması 2 (Divisive Amplicon Denoising Algorithm 2, DADA2) (101) aracılığıyla filtrelenmiştir. DADA2, Illumina amplicon dizileme verilerindeki hataları düzeltmeyi amaçlayan yöntem poisson modelini kullanarak hataların gözlemlenme olasılığını dikkate alırken karşılaşılan durumun gerçek bir varyant mı yoksa

hatadan mı kaynaklı olduğuna poisson modeli ile hesaplanan p değeri ile karar verir. Eğer p değeri belirlenen eşikten küçük ise sekans gerçek bir varyanttır. Yani sekans hata olarak değerlendirilmesini sağlayacak tekrarlanma sayısından daha çok gözlemlenmiştir. Ayrıca DADA2 ile kimerik diziler filtrelenmektedir. Çözünürlüğü yüksektir (102). DADA2 aracı kullanılırken Primerlerdeki dejenerasyonun neden olabileceği yanlış pozitif kimerik dizilerin tespitini engellemek için, primerlerin veriden çıkarılmasını gerektirmektedir. DADA2 tarafından üretilen çıktılar ampikon dizisi varyantlar (Amplicon Sequence Variant, ASV) olarak isimlendirilmektedir (103). DADA2 ile gürültü azaltma işleminin ardından elde edilen ASVler dizilenen gen bölgesindeki tek nükleotid farklarını dikkate alır ve duyarlılığı yüksektir (104). Bu yöntemlerin sonucunda her örnekteki benzersiz dizi sayısını içeren özellik tabloları ve özellik tablolarındaki özellik tanımlayıcılarını eşleyen sekans bilgisini içeren özellik verileri oluşturulmuştur (105). Tüm örneklerin %10 undan azında bulunan özellikler veri analizi aşamasında elenmiştir.

Çalışmada taksonomik atamanın gerçekleştirilebilmesi için literatürde yaygın kullanımı olan önceden eğitilmiş Naive Bayes sınıflandırıcı olan gg-13-8-99-515-806-nb-classifier.qza (98,106,107,108) kullanılmıştır. Taksonomik atama için kullanılan Naive Bayes sınıflandırıcı literatürdeki yaygın kullanımı, önceden eğitilmiş bir model olması, QIIME2 üzerinde uygulanma kolaylığı sebebiyle tercih edilmiştir. Naive Bayes sınıflandırıcı 16S rRNA bölgesinden (515F/806R primer çiftiyle bağlı V4 bölgesi) 250 bazı içerecek şekilde bölünen Greengenes 13_8 %99 (109,106) OTU'lar üzerinde eğitilmiştir. Taksonomik atamanın ardından verinin normalize edilmiş bir şekilde sunulduğu göreceli bolluk değerlerinin ifade edildiği temsil vektörleri oluşturulmuştur. Cins seviye ve tür seviyesinde taksonomik atamaların yapıldığı OTU tablolarının sınıflandırıcılar için ayrı ayrı kullanılmasının yanında tabloların uç uca eklenmesiyle elde edilen temsil vektörleri de sınıflandırma aşamasında kullanılmıştır. Temsil vektörlerinin birleştirilmesine ait gösterim Şekil 4'te gösterilmiştir.



Şekil 4. Cins ve tür seviye temsil vektörlerinin birleştirilmesine ait örnek gösterim

3.2.3. DNABERT-S Temsil Vektörlerinin Oluşturulması

DNABERT-2 (110) modeline dayanan yöntemde hata eğilimi yüksek uzun DNA sekanslarından başarılı gömme sonuçlarının elde edilebilmesi için Manifold Örnek Karıştırma (MI-Mix) kaybı ve Müfredat Karşıt Öğrenme (C2LR) yaklaşımlarını kullanan modelin DNA sekanslarını kümeleme ve ayırma yeteneği yüksektir (111).

Metodun birinci aşamasında Ağırlıklı SimCLR yöntemi (112), ikinci aşamada (113) den esinlenen Manifold Örnek Karıştırma yöntemi (111) kullanılmıştır.

3.2.3.1. Müfredat Karşıt Öğrenim

Eğitim aşamasında önce basit örneklerle ait yığınların işlenmesini ardından daha zor örneklerle geçilerek eğitimin sağlanması üzerine kurulu bir yaklaşımdır. Yöntem 2 aşama içermektedir. Birinci aşama SimCLR yöntemi (112), 2. aşama Manifold Örnek Karıştırma yöntemidir. Manifold Örnek Karıştırma yöntemi rastgele gizli katmandaki bağlantı örneklerini karıştırmak için Örnek karıştırma (Instance Mixup, i-Mix) yöntemini (114) kullanmaktadır. i-Mix yaklaşımında daha zorlu pozitif ve negatif örnekler oluşturmak için giriş katmanında çapalar karıştırılmaktadır. Karıştırma ağırlığı λ_i , Beta(α, α) dağılımından elde edilir. α hiperparametredir. Modelde dna dizilerini daha derin bir katmanda karıştırma yaklaşımı olan ve (113) yaklaşımından esinlenen Manifold Örnek Karıştırma yaklaşımı kullanılmıştır (111).

Modeldeki katmanlar arasından rastgele bir katman seçilir. x_i örnek ve v_i etiket olmak üzere, çapayı içeren batch $\{(x_i, v_i)\}_{i=1}^B$ m. inci katmana kadar modelden geçirilerek batch için m. katmandaki gizli durum $\{(g_m(x_i), v_i)\}_{i=1}^B$ olarak elde edilir. Karıştırma işlemi gerçekleştirilerek $\{(g_m(\hat{x}_i), \hat{v}_i)\}_{i=1}^B$ elde edilir. Karıştırılmış gizli durum h_i^m “Eşitlik 10” ve karıştırılmış etiketler “Eşitlik 11” ile hesaplanmıştır. Son gizli durum eldesinde her bir batch için $\{f_m(h_i^m)\}_{i=1}^B$ elde edilir. Ağırlıklar $\alpha_{ii+} = 1$ olmak üzere i. çapa (x_i, v_i) için kayıp fonksiyonu $\hat{L}$ ’nın hesaplanması “Eşitlik 12”, “Eşitlik 13”, “Eşitlik 14” ile gösterilmiştir.

$$h_i^m = \lambda g_m(x_i) + (1 - \lambda)g_m(\hat{x}_i) \quad (\text{Eşitlik 10})$$

$$v_i^{karıştırma} = \lambda v_i + (1 - \lambda)\hat{v}_i \quad (\text{Eşitlik 11})$$

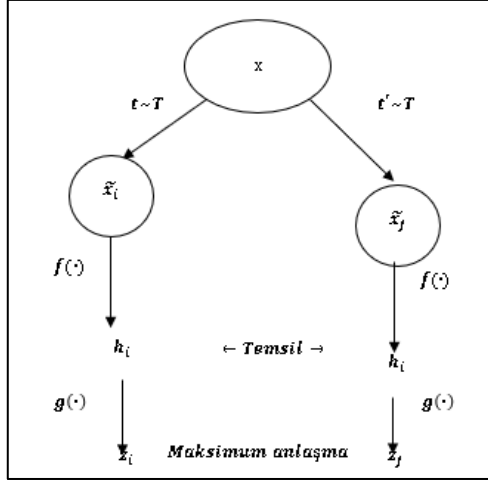
$$\begin{aligned} \hat{\ell}(f_m(h_i^m), v_i^{karıştırma}) = \\ - \sum_{n=1}^B v_{i,n}^{karıştırma} \log \frac{\exp\left(\frac{s(f_m(h_i^m), f(x_{n+}))}{\tau}\right)}{\sum_{j=1}^B \alpha_{ij+} \exp\left(\frac{s(f_m(h_i^m), f(x_{j+}))}{\tau}\right)} \end{aligned} \quad (\text{Eşitlik 12})$$

$$\alpha_{ij+} = \frac{\exp\left(\frac{s(f_m(h_i^m), f(x_{j+}))}{\tau}\right)}{\frac{1}{B-1} \sum_{k=1, k \neq i}^B \exp(s(f_m(h_i^m), f(x_{k+}))/\tau)} \quad (\text{Eşitlik 13})$$

$$\hat{L} = \frac{1}{B} \sum_{i=1}^B \hat{\ell}(f_m(h_i^m), v_i^{karıştırma}) \quad (\text{Eşitlik 14})$$

3.2.3.2. SimCLR Metodu

Aynı verinin farklı veri arttırımı sonuçlarını karşılık kaybını örtük uzayda minimize edecek şekilde temsil vektörlerinin oluşmasında maksimum oy birliğini amaçlayan yöntemdir (112).



Şekil 5. Karşıtlıklı öğrenme için mimari (Chen ve ark.’dan, 112)

Stokastik veri arttırımı aynı örneğin 2 farklı ilişkili görüntüsünü $\tilde{x}_i$, $\tilde{x}_j$ oluşturur. Oluşan bu ilişkili görüntüler pozitif çift olarak adlandırılır. Şekil 5’te $f(\cdot)$ kodlayıcı sinir ağı iken, $g(\cdot)$ projeksiyon başlığıdır (112).

SimCLR yaklaşımında x_i örneği için aynı batch içerisindeki $\{(x_i, x_i^+)\}_{i=1}^B$ pozitif, diğer örnekleri $(2B-2)$ negatif örnek olarak değerlendirir. Seçilen örneğin pozitif örnek ile benzerliğini arttırmaya negatif örneklerle benzerliği azaltmaya çalışır. Seçilen örneğe en yakın olan negatif örnekler daha yüksek ağırlıklı olarak belirlenir. Negatif örneklere ait ağırlıkların seçilen örneğe bağlı olarak değişmesine ağırlıklı SimCLR denir (115). (x_i, x_i^+) $v_i \in \{0,1\}^B$ için $v_{i,i} = 1$ pozitif örneklere ait etiketler, $v_{i,j \neq i} = 0$ negatif örneklere ait etiketleri temsil etmektedir.

x_i için Ağırlıklandırılmış SimCLR kaybı “Eşitlik 15” ile gösterilmiştir.

$$\ell(f(x_i), v_i) = - \sum_{n=1}^B v_{i,n} \log \frac{\exp(s(f(x_i), f(x_n^+))/\tau)}{\sum_{j \neq i} \alpha_{ij} \exp(s(f(x_i), f(x_j))/\tau)} \quad (\text{Eşitlik 15})$$

τ softmax fonksiyonundaki normalizasyonu sağlayan sıcaklık parametresi, $s(\cdot, \cdot)$ cosine benzerliği, ağırlıkları ifade eden α_{ij} x_i örneği için x_j ’nin göreceli önemini ifade etmektedir. $\alpha_{ii^+} = 1$ olmak üzere α_{ij} ’nin hesaplanması “Eşitlik 16” ile verilmiştir.

$$\alpha_{ij} = \frac{\exp(s(f(x_i), f(x_j))/\tau)}{\frac{1}{2B-2} \sum_{k \neq i, i^+} \exp(\frac{s(f(x_i), f(x_k))}{\tau})} \quad (\text{Eşitlik 16})$$

Tez çalışmasında verilere ait sekans uzunluğunun çok büyük olması sebebiyle yığın boyutu 512 olarak seçilmiştir. Boyut seçiminin ardından her bir yığın için elimizde oluşan temsil vektörlerinin örneği temsil edebilmesi için 2 yaklaşım kullanılmıştır. Bu yaklaşımlardan birincisi her bir örneğe ait oluşturulan yığın çıktılarının ortalaması alınarak temsil vektörünün boyutu değişmeden bir temsil vektörü elde edilebilmiştir. 2. yöntem ise her bir yığın için maksimum olan değere sahip olan temsil vektörü seçilmiştir.

3.3.3. Yeniden Örnekleme Yöntemleri

Tez çalışması kapsamında eğitim verisindeki sınıf dengesizliğinin giderilmesi model performansının artırılması için kullanılan önemli yöntemlerdendir. Model geliştirme aşamasında yeniden örnekleme yöntemlerinden Sentetik Azınlık Örneklem Arttırma Yöntemi (Synthetic Minority Over Sampling Technique, SMOTE) (116) ve yeniden yerine koyma ile rastgele örnekleme yaklaşımı (117) kullanılmıştır. Yeniden örnekleme yaklaşımları kullanılarak eğitim veri kümesi 1000 Crohn hastası, 1000 kontrol verisini içerecek şekilde arttırılmıştır.

3.3.3.1. Sentetik Azınlık Örneklem Arttırma Yöntemi (SMOTE)

Veri setinde bazı örnekler üzerinde ve komşu örnekler arasında enterpolasyon yapılarak yeni örnekler üretilmektedir. K - en yakın komşuların bulunabilmesi için azınlık sınıfına ait her bir örneğin diğer örneklerle uzaklığı hesaplanır. Ardından aşırı örneklemenin büyüklüğüne göre her örnek x_i için m en yakın komşu x_{ij} ($j=1,2,...,m$) rastgele seçilir (116). Yeni örneklerin oluşturulmasına ait formül “Eşitlik 17” ile verilmiştir.

$$p_{ij} = x_i + rand(0,1) \times (x_{ij} - x_i) \quad (\text{Eşitlik 17})$$

$rand(0,1)$, $[0,1]$ arasında uniform dağılan değerlerden rastgele değer seçimini ifade etmektedir. p_{ij} değerinin hesaplanması belirli bir dengesizlik oranına erişinceye kadar devam eder (116). Dengesiz veri kümeleri ile çalışmak için kullanılan imbalanced-learn kütüphanesinin SMOTE fonksiyonu kullanılmıştır.

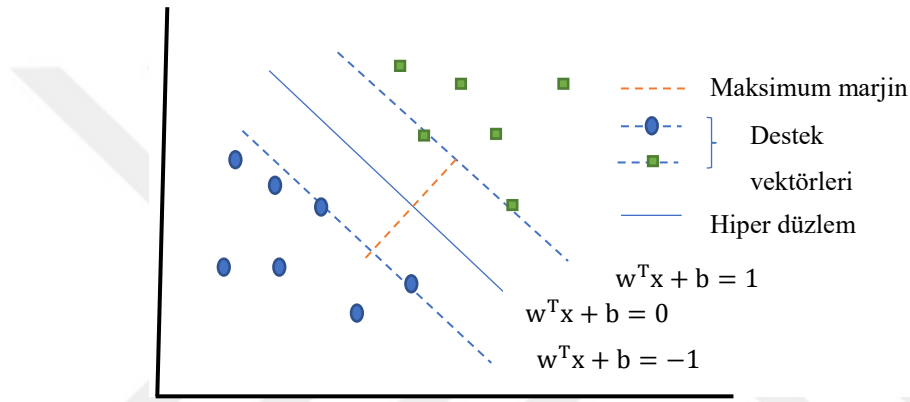
3.3.3.2. Rastgele Örnek Seçimi ile Veri Arttırımı

Eğitim verisi üzerinde rastgele örneğin seçilmesinin ardından tekrar seçilebilmek üzere yerine konulmasına dayanan bir yaklaşım kullanılmıştır. imbalanced-learn kütüphanesinin RandomOverSampler fonksiyonu kullanılmıştır (117).

3.3.4. Sınıflayıcılar

3.3.4.1. Destek Vektör Makineleri

Yapısal riski azaltmayı amaçlayan model olasılık dağılımı bilinmeyen verinin yanlış sınıflama olasılığını minimize etmeye çalışmaktadır. Amaç sınıflar arasındaki minimum uzaklığı maksimize etmektir. En iyi sonuç veriyi maksimum marjin ile ayıracak kanonik hiper düzlemi bulmaktır (118). Doğrusal SVM sınıflayıcıya ait gösterim Şekil 6'da gösterilmiştir.

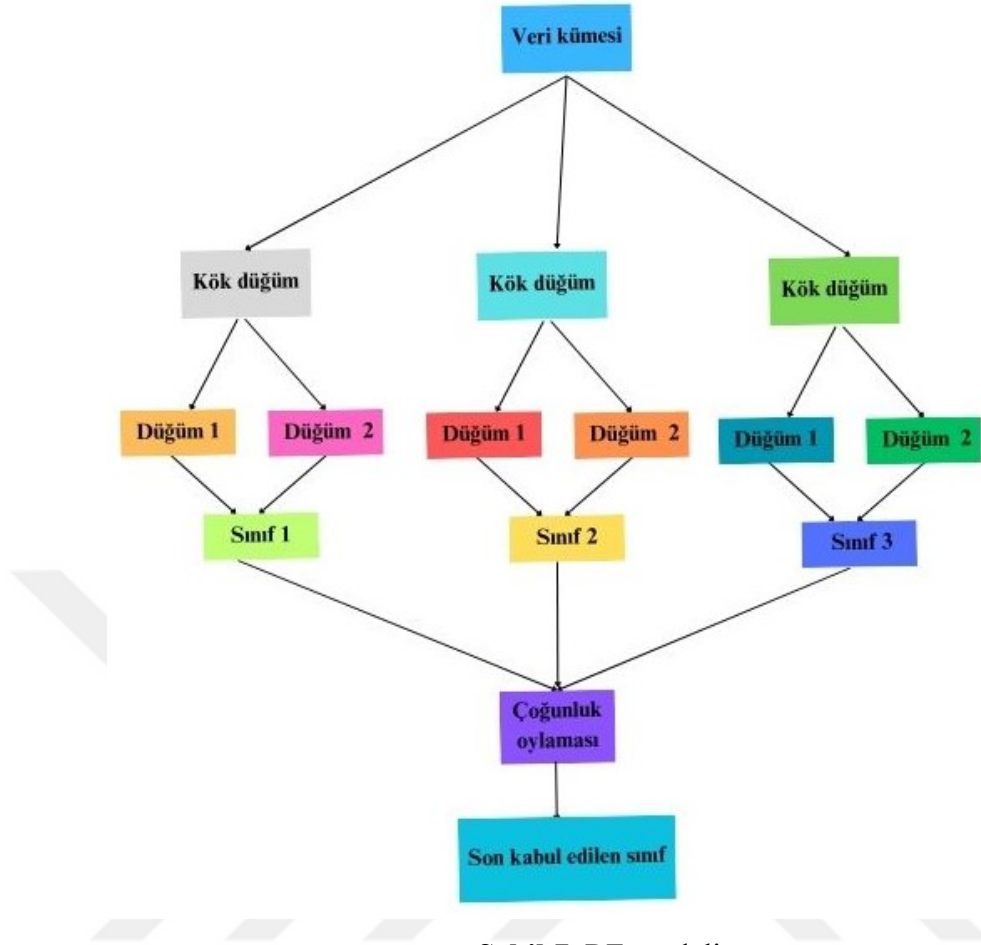


Şekil 6. Doğrusal SVM sınıflayıcı (Rath ve ark.'dan, 119)

Lineer bir düzlem ile ayrılamayan sınıflama problemlerinde kernel yapıları kullanılmaktadır. Eğitim sırasında karar sınırı oluşturulurken eğitim kümesinin genellikle bir alt kümesi kullanılmaktadır. Karar sınırında bulunan örneklere destek vektörleri denir (120). Modelin önemli parametreleri C ve kernel parametresidir. Modelde C parametresi düzenlileştirme parametresidir. C değeri arttıkça örneklerin model üzerindeki etkisi artmaktadır. Çalışmada scikit-learn kütüphanesinin SVC fonksiyonu C, gamma, Kernel, max_iter hiper parametreleri ile birlikte kullanılmıştır.

3.3.4.2. Rastgele Orman Algoritması

Çok sayıda karar ağacının oylama mantığı ile topluluk öğrenmesi yaklaşımı ile değerlendirilmesi ile ortaya çıkar (121). Eğitim örneklerinin bir alt kümesi yeniden yerine koyma ile eğitilir (122). Çalışmada scikit-learn kütüphanesi kullanılmıştır. RF modeline ait gösterim Şekil 7 ile gösterilmiştir.



Şekil 7. RF modeli

3.3.4.3. Uyarlanabilir Güçlendirme Algoritması

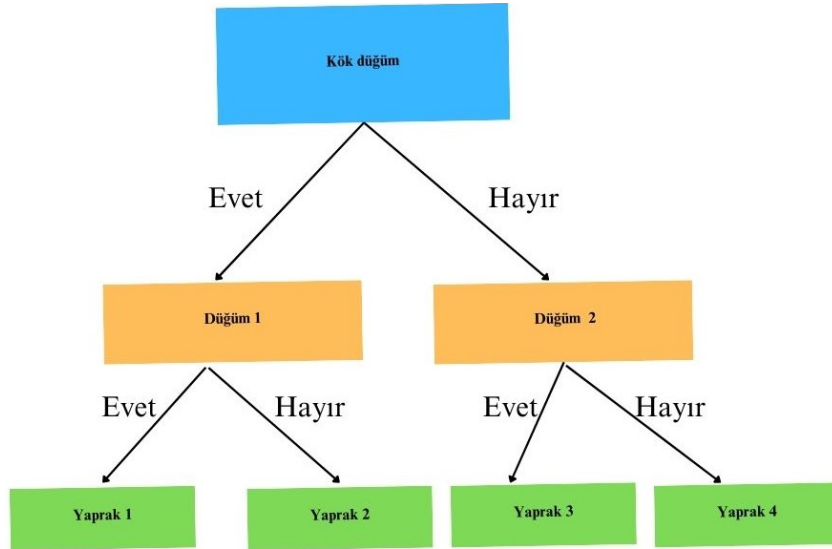
Zayıf sınıflandırıcıların topluluk öğrenmesi modeli olan Uyarlanabilir Güçlendirme (Adaptive Boosting, AdaBoost) her yinelemede hatayı minimize ederek güçlü bir sınıflandırıcı oluşturmayı amaçlar. Zayıf sınıflandırıcının yanlış sınıflandırdığı eğitim verisinin ağırlığı artırılırken, doğru sınıflandırdığı eğitim verisinin ağırlığı azaltılır. Topluluk öğrenmesi yaklaşımı ile model performansını iyileştirmektedir (124). Çalışmada Scikit-learn kütüphanesi kullanılmıştır. AdaBoost algoritmasına ait örnek gösterim Şekil 9’da gösterilmiştir. Şekil 8’de eğitimde kullanılan zayıf öğrenciler tek düğümde oluşan karar ağaçlarıdır. Zayıf öğrencilerin ağırlıkları eşit olmayıp hata miktarına bağlı olarak ayarlanmaktadır.



Şekil 8. AdaBoost modeli

3.3.4.4. Karar Ağacı Algoritması

Ağaç yapısını kullanan denetimli öğrenme modelidir. Belirli koşullara göre veri kayıtlarını kategorilere ayırır. Veri sınıf etiketine göre her nodun Gini indeks, entropi, bilgi kazancı gibi saflık değerini dikkate alarak bölünür. Her bölünme sonrasında saflık değeri hesaplanır. Toplam saflık bu saflık değerlerinin ağırlık ortalamasını içerir. Algoritma, her bir testin saflık ölçütünü değerlendirir ve en düşük saflık değerine sahip bölünmeyi seçer. Bu işlem, her düğümde tekrarlanarak alt düğümlerin sonuç değişkeni açısından daha tutarlı ve homojen hale gelmesi sağlanır (125). Karar ağacına ait gösterim Şekil 9’da gösterilmiştir.



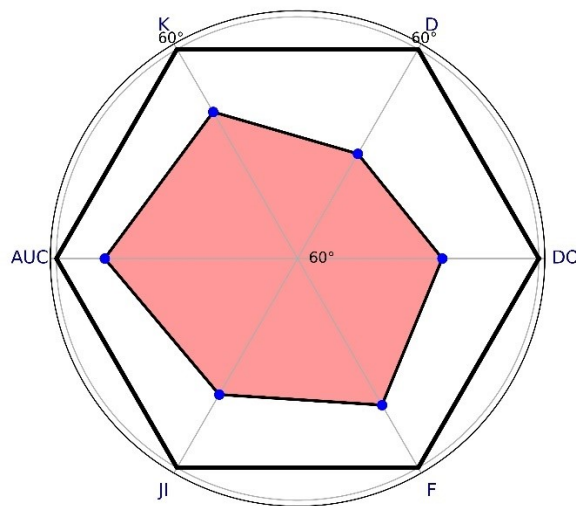
Şekil 9. Karar ağacı

3.3.5. Performans Değerlendirme Ölçütleri

Model performansının değerlendirilmesi doğruluk, hamming kaybı, mikro duyarlılık, mikro kesinlik, mikro AUC, mikro F1 skor, makro duyarlılık, makro kesinlik, makro AUC, makro F1 skor, ağırlıklandırılmış duyarlılık, ağırlıklandırılmış kesinlik, ağırlıklandırılmış AUC, ağırlıklandırılmış F1 skor, hata matrisi metrikleri ve Poligon Alan metriği (PAM) (126) kullanılarak gerçekleştirilmiştir. Performans ölçütleri Tablo 7’de verilmiştir.

Poligon Alan metriği tek bir metrik kullanılarak model performansı değerlendirilmesini mümkün kılmaktadır. Bu performans değerlendirme metriği doğruluk , duyarlılık, kesinlik, Jaccard indeksi, F1 skor ve AUC skorların kullanılmasına dayanmaktadır. Şekil 10’da bu metriklerin isimlendirmesi doğruluk (DO), duyarlılık (D), kesinlik (K), Jaccard İndeksi (JI), F skor (F) ve AUC şeklinde gösterilmiştir.

Poligon alan metriği hesaplanırken kenarları 1 birim olacak şekilde oluşturulan altıgen optimum metrik sonuçlarını ifade ederken oluşturulan altıgenin alanı 2.59807 olması sebebiyle hesaplanan metriklerden oluşturulan çokgen alanının ideal altıgene ait alana bölünmesiyle sağlanmaktadır. Poligon Alan metriği hesaplanırken ağırlıklandırılmış skorlar dikkate alınmıştır (126). Hata matrisine ait gösterim Tablo 6’de gösterilmiştir.



Şekil 10. Poligon alan metriği gösterimi

Tablo 6. Hata matrisi

Tahmin	Gerçek Sınıf	
	Pozitif	Negatif
Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Doğru Pozitif (DP): Gerçekte pozitif olup model tarafından pozitif olarak tahminlenen değer

Doğru Negatif (DN): Gerçekte negatif olup model tarafından negatif olarak tahminlenen değer

Yanlış Pozitif (YP): Gerçekte negatif olup model tarafından pozitif olarak tahminlenen değer

Yanlış Negatif (YN): Gerçekte pozitif olup model tarafından negatif olarak tahminlenen değer

Tablo 7. Performans ölçütler

Performans Ölçütü	Formül	Tanım
Doğruluk	$\frac{DP + DN}{DP + DN + YP + YN}$	Modelin yaptığı tüm tahminlerden ne kadarını doğru olduğunu göstermektedir.
Hamming kaybı	$\frac{DP + DN + YP + YN}{DP + DN + YP + YN}$	Modelin hata oranını gösterir.
Mikro duyarlılık	$\frac{DP}{DP + YN}$	Pozitif sınıfın doğru tahminlenme oranını gösterir.
Makro duyarlılık	$duyarlılık_{pozitif} = \frac{DP}{DP + YN}$	İkili sınıflandırıcı olduğu için negatif ve pozitif sınıflar için hesaplanan duyarlılığın ortalaması alınmaktadır. Bu metrikte sınıflar eşit öneme sahiptir.
	$duyarlılık_{negatif} = \frac{DN}{DN + YP}$	
	$\frac{duyarlılık_{pozitif} + duyarlılık_{negatif}}{2}$	
Ağırlıklandırılmış duyarlılık	$w_{pozitif} = \frac{pozitif\ örnek\ sayısı}{toplam\ örnek\ sayısı}$	Dengesiz veri kümelerinde sınıf dağılımını dikkate almaktadır.
	$w_{negatif} = \frac{negatif\ örnek\ sayısı}{toplam\ örnek\ sayısı}$	
	$w_{pozitif} \times duyarlılık_{pozitif} + w_{negatif} \times duyarlılık_{negatif}$	
Mikro kesinlik	$\frac{DP}{DP + YP}$	Yapılan pozitif tahminlerin ne kadarının gerçekten pozitif olduğunu gösterir.
Makro kesinlik	$kesinlik_{pozitif} = \frac{DP}{DP + YP}$	İkili sınıflandırıcı olduğu için negatif ve pozitif sınıflar için hesaplanan kesinliğin ortalaması alınmaktadır. Bu metrikte sınıflar eşit öneme sahiptir.
	$kesinlik_{negatif} = \frac{DN}{DN + YN}$	
	$\frac{kesinlik_{pozitif} + kesinlik_{negatif}}{2}$	
Ağırlıklandırılmış kesinlik	$w_{pozitif} = \frac{pozitif\ örnek\ sayısı}{toplam\ örnek\ sayısı}$	

Tablo 6. (Devam)

Ağırlıklandırılmış kesinlik	$w_{negatif} = \frac{negatif \ örnek \ sayısı}{toplam \ örnek \ sayısı}$ $w_{pozitif} \times kesinlik_{pozitif} + w_{negatif} \times kesinlik_{negatif}$	Dengesiz veri kümelerinde sınıf dağılımını dikkate almaktadır.
Mikro AUC	$DPO = \frac{DP}{DP + YN}$ $YPO = \frac{YP}{YP + DN}$	ROC eğrisinin altında kalan alanı n ölçüsüdür. Bu eğri doğru pozitif oranı (DPO)ve yanlış pozitif oranı (YPO) ile çizilir.
Makro AUC	$\frac{AUC_{pozitif} + AUC_{negatif}}{2}$	ROC eğrisinin altında kalan alan pozitif sınıf ve negatif sınıf için ayrı ayrı hesaplanarak ikili sınıflama olduğu için ortalaması alınmıştır. Her sınıfa eşit önem vermektedir.
Ağırlıklandırılmış AUC	$w_{pozitif} = \frac{pozitif \ örnek \ sayısı}{toplam \ örnek \ sayısı}$ $w_{negatif} = \frac{negatif \ örnek \ sayısı}{toplam \ örnek \ sayısı}$ $w_{pozitif} \times AUC_{pozitif} + w_{negatif} \times AUC_{negatif}$	Dengesiz veri kümelerinde sınıf dağılımını dikkate almaktadır.
Mikro F1	$2 \times \frac{kesinlik \times duyarlılık}{kesinlik + duyarlılık}$	Sınıf dengesizliklerini göz ardı eder.
Makro F1	$\frac{F1_{pozitif} + F1_{negatif}}{2}$	Pozitif sınıf ve negatif sınıf için ayrı ayrı hesaplanarak ikili sınıflama olduğu için ortalaması alınmıştır. Her sınıfa eşit önem vermektedir.
Ağırlıklandırılmış F1	$w_{pozitif} = \frac{pozitif \ örnek \ sayısı}{toplam \ örnek \ sayısı}$ $w_{negatif} = \frac{negatif \ örnek \ sayısı}{toplam \ örnek \ sayısı}$ $w_{pozitif} \times F1_{pozitif} + w_{negatif} \times F1_{negatif}$	Dengesiz veri kümelerinde sınıf dağılımını dikkate almaktadır.
Poligon Alan Metriği	$PAM = \frac{PA}{2.59807}$	Doğruluk, duyarlılık, kesinlik, Jaccard İndeksi, F skor ve AUC skor metrikleriyle oluşturulan çokgen alanı (PA) ile hesaplanır. Modeller sonuçlarının değerlendirilmesine birçok metriğin birlikte kullanılmasına olanak tanımaktadır.
Jaccard İndeks	$JI = \frac{TP}{TP + FP + FN}$	Tahmin ve gerçek değerler arasında benzerliği hesaplamada kullanılmaktadır.

3.3.6. Kullanılan Platform ve Kütüphaneler

Tezde Truba sistemi kullanılmıştır. Çalışmada farklı sürümlerin kullanılması sebebiyle farklı conda ortamları kullanılmıştır. QIIME2 2022.2, python 3.9.6 ile birlikte kullanılmıştır. K-mer temsil vektörleri için python 3.7.12 kullanılmıştır. DNABERT-S için oluşturulan conda ortamında python 3.9.6 sürümü torch 1.9.1+cu111 versiyonu, torchvision 0.10.1+cu111 versiyonu kullanılmıştır.

3.3.7. Hiper Parametre Optimizasyonu

Izgara aramasında değişkenler için taranacak parametre uzayı kullanılmaktadır. K değişken için parametre uzayı $(L^{(1)} \dots L^{(K)})$, ızgara aramasındaki deneme sayısı “Eşitlik 18” ile gösterilmiştir. Bu hiper parametre optimizasyonu yaklaşımı çok boyutluluk probleminde sahiptir (127, 128). Hiper parametre optimizasyonu için python scikit-learn kütüphanesinin GridSearchCV fonksiyonu kullanılmıştır. 10 katlı çapraz doğrulama doğrulama kullanılmıştır.

Tez çalışmasında kullanılan sınıflayıcılara ait GridSearchCV fonksiyonuyla taranan parametre uzayı Tablo 8 ile gösterilmiştir.

$$S = \prod_{k=1}^K |L^{(k)}| \quad (\text{Eşitlik 18})$$

Tablo 8. Sınıflandırıcılara ait parametre uzayları

Sınıflayıcı	Parametre uzayı
AdaBoostClassifier	"n_estimators": [25, 50, 75], "learning_rate": [0.1, 0.5, 1], "algorithm": ['SAMME', 'SAMME.R']
DecisionTreeClassifier	"criterion": ["gini", "entropy"], "max_depth": [3, 5, 10], "min_samples_split": [5, 10], "min_samples_leaf": [2, 4, 6],
SVM	"C": $[2^{-2}, 2^{-1}, 1, 2]$ "gamma": $[2^{-2}, 2^{-1}, 1, 2]$ "kernel": ["linear", "poly", "rbf"] "max_iter": [25, 50, 100, 200, 300, 400, 500]
RandomForestClassifier	"n_estimators": [5, 10, 15, 20, 25] "max_depth": [2, 3, 4, 5] "min_samples_leaf": [1, 5, 10]

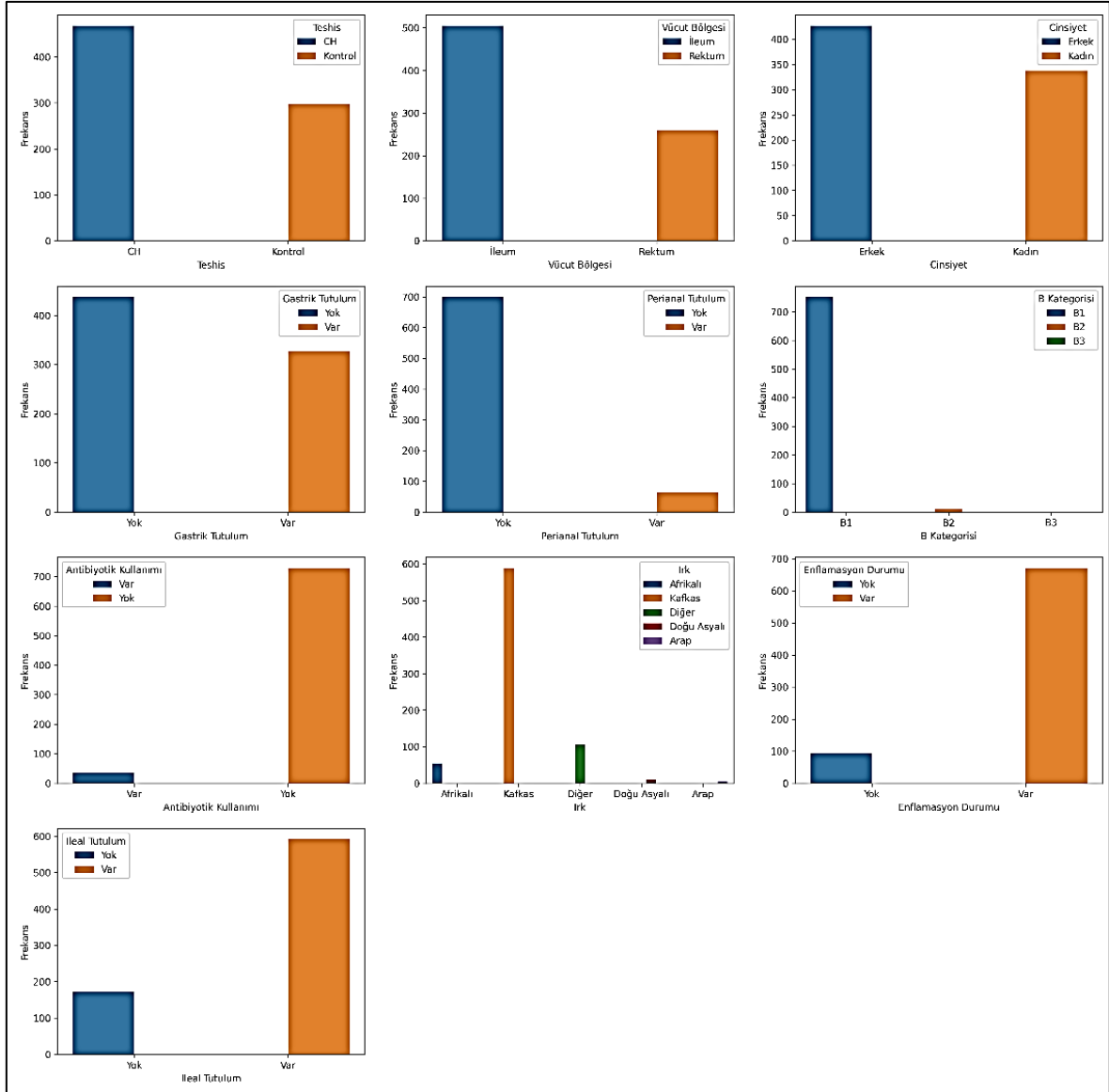
4. BULGULAR

4.1. Veri Setine Ait Bulgular

QIIME2 platformu için kullanılan meta verisine ait sütunlar 'sample-id', 'Vucut_bolgesi', 'Teshis', 'Cinsiyet', 'Yas', 'Gastrik_tutulum', 'Perianal', 'b_kategorisi', 'Antibiyotik', 'Irk', 'İleal tutulum', 'Enflamasyon_durumu' olarak verilmiştir. Meta veriye ait tanımlayıcı istatistikler Tablo 9 ve 10’da gösterilmiştir. Kategorik değişkenlere ait dağılımlar Şekil 11’de histogram grafikleri ile gösterilmiştir. Nümerik bir değişken olan “Yaş” değişkenine ait kutu grafiği Şekil 12’de gösterilmiştir.

Tablo 9. Kategorik verilere ait tanımlayıcı istatistikler

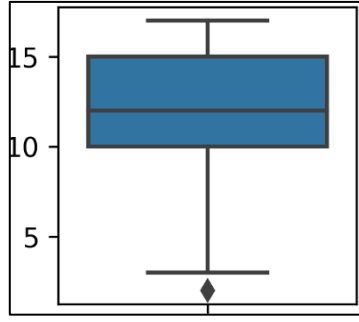
	Sayı	Tekil değer sayısı	En çok tekrar eden	Frekans
Teşhis	765.0	2.0	1.0	468.0
Vücut Bölgesi	765.0	2.0	0.0	505.0
Cinsiyet	765.0	2.0	0.0	427.0
Gastrik Tutulum	765.0	2.0	0.0	438.0
Perianal Tutulum	716.0	2.0	0.0	651.0
B Kategorisi	729.0	3.0	0.0	717.0
Antibiyotik Kullanımı	765.0	2.0	0.0	729.0
Irk	728.0	5.0	3.0	552.0
İleal Tutulum	484.0	2.0	1.0	348.0
Enflamasyon Durumu	379.0	2.0	1.0	285.0



Şekil 11. Kategorik verilere ait tanımlayıcı istatistikler

Tablo 10. Nümerik verilere ait tanımlayıcı istatistikler

	Sayım	Ortalama	Std	Min.	25%	50%	75%	Maks.
Yaş	765.0	12,08	3,29	2.0	10.0	12.0	15.0	17.0



Şekil 12. Nümerik yaş değişkenine ait kutu grafiği

Mevcut veride Yaş değişkenine ait değerler 17 yaşından küçük ve eşit olacak şekilde filtrelenerek 986 hastaya ait meta veri ve sekans verisi elde edilmiştir.

QIIME2 'de DADA2 aracı kullanılarak gerçekleştirilen gürültü filtreleme işleminin ardından veri kümesine ait özellik tablosu Tablo 11'de gösterilmiştir.

Tablo 11. Özellik tablo özeti

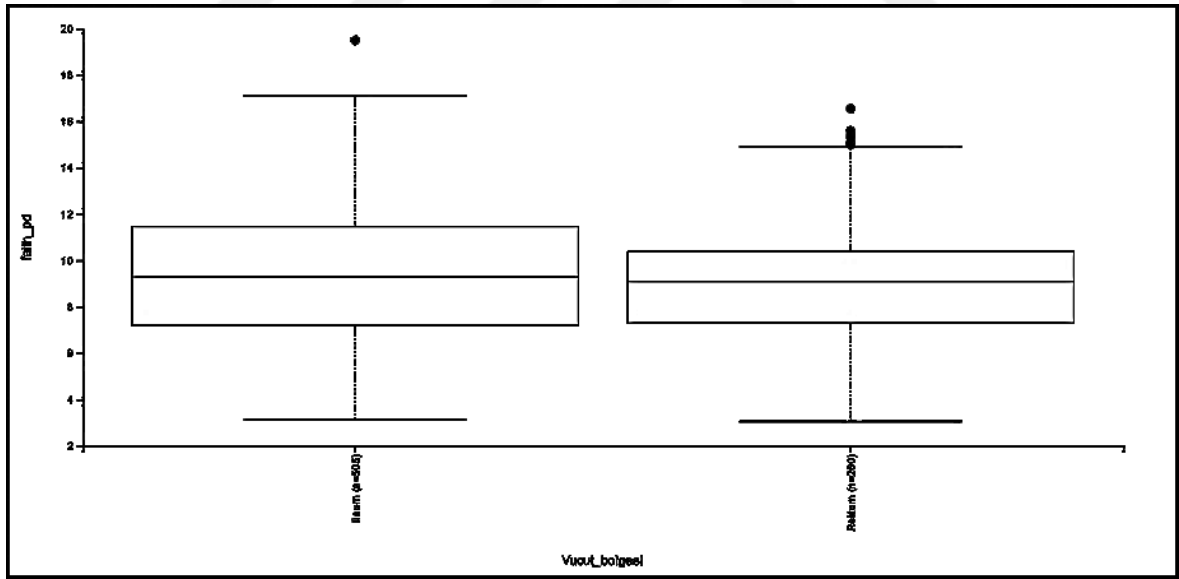
Metrik	Örnek
Örnek sayısı	765
Özellik sayısı	6.768
Toplam frekans	24.338.911

Filogenetik analiz için ağaç oluşturulması aşamasında QIIME2'nin MAFF çoklu sekans hizalama algoritması ve QIIME2'nin FastTree algoritması kullanılarak köklendirme işlemi gerçekleştirilmiştir (99). Alfa ve beta çeşitlilik analizleri QIIME2 üzerinde yapılmıştır.

Verimize ait alfa çeşitlilik analizi sonuçları Şekil 13-21 arasında her bir grup içindeki alfa çeşitlilik ölçümlerine ait dağılımın kutu grafiklerini içermektedir. Tablo 12-20 arasında gruplar arasında anlamlı bir fark olup olmadığına dair Mann Whitney-U testi gerçekleştirilmiştir. Şekil 13-16-19'da vücut bölgesine (ileum, rektum) bağlı olarak sırasıyla Faith Filogenetik Çeşitlilik İndeksi, Pileou Düzgünlük İndeksi, Shannon Çeşitlilik İndeklerine ait dağılımların kutu grafikleri verilmiştir. Bu dağılımlar arasında anlamlı bir fark olmadığı ($p>0.05$) (sırasıyla Faith Filogenetik Çeşitlilik İndeksi: $U=69967$, $p=0.14$; Pileou Düzgünlük İndeksi: $U=65699$, $p=0.99$; Shannon Çeşitlilik İndeksi: $U=64676$, $p=0.74$) Tablo 12, 15, 18'de gösterilmiştir.

Şekil 15, 17’de teşhise (CH, kontrol) bağlı olarak sırasıyla Faith Filogenetik Çeşitlilik İndeksi, Pileou Düzgünlük indekslerine ait dağılımların kutu grafikleri verilmiştir. Bu dağılımlar arasında anlamlı bir fark olmadığı ($p>0.05$) (sırasıyla Faith Filogenetik Çeşitlilik İndeksi: $U=69594$, $p=0.97$; Pileou Düzgünlük İndeksi: $U=65325$, $p=0.16$) Tablo 13, Tablo 16’de gösterilmiştir. Şekil 19’de teşhise (CH, kontrol) bağlı olarak Shannon Çeşitlilik İndeksine ait kutu grafiği gösterilmiştir. Tablo 19’de görüldüğü gibi Shannon Çeşitlilik İndeksi açısından teşhis grupları arasında anlamlı bir fark olduğu $p < 0.05$ gösterilmiştir ($U=81416$, $p=6.31e-5$).

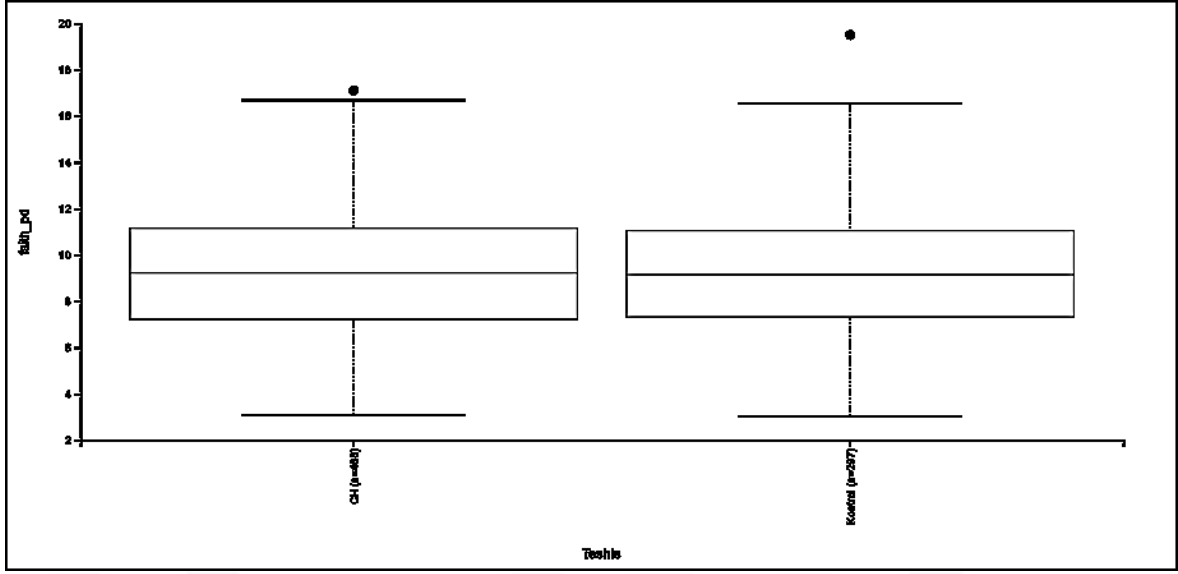
Şekil 15, 18, 21’de cinsiyete (kadın, erkek) bağlı olarak sırasıyla Faith Filogenetik Çeşitlilik İndeksi, Pileou Düzgünlük İndeksi, Shannon Çeşitlilik İndekslerine ait dağılımların kutu grafikleri verilmiştir. Bu dağılımlar arasında anlamlı bir fark olmadığı ($p>0.05$) (sırasıyla Faith Filogenetik Çeşitlilik İndeksi: $U=72833$, $p=0.83$; Pileou Düzgünlük İndeksi: $U=71397$, $p=0.80$; Shannon Çeşitlilik İndeksi: $U=70050$, $p=0.49$) Tablo 14, 17, 20’de gösterilmiştir.



Şekil 13. Vücut bölgesine bağlı Faith_pd değerinin dağılımı

Tablo 12. Vücut bölgesine bağlı Faith_pd değerlerinin Mann-Whitney-U analizine ait sonuçları

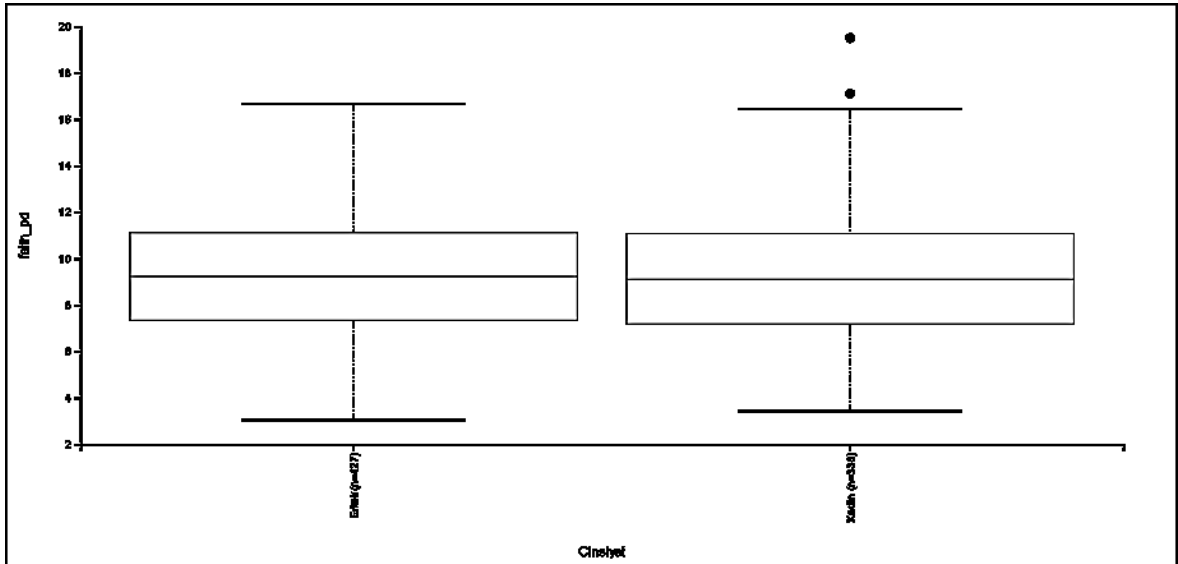
Grup 1	Grup 2	U	p değeri
İleum (n=505)	Rektum (n=260)	69967	0.14



Şekil 14. Teşhis değişkenine bağlı Faith_pd değerinin dağılımı

Tablo 13. Teşhis değişkenine bağlı Faith_pd değerlerinin Mann-Whitney U analizine ait sonuçları

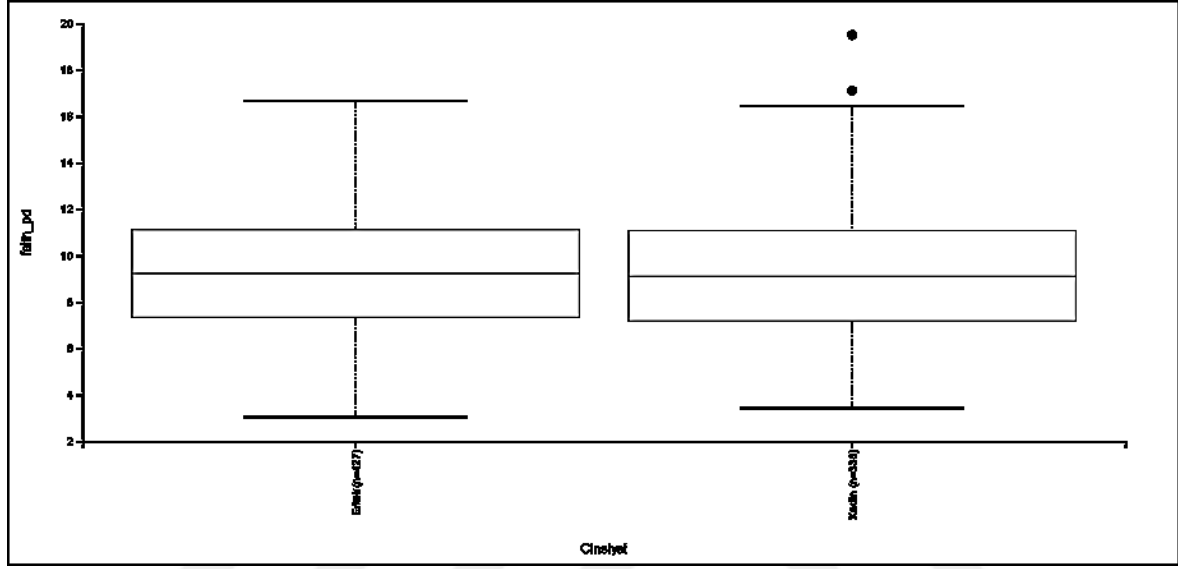
Grup 1	Grup 2	U	p değeri
CH (n=468)	Kontrol (n=297)	69594	0.97



Şekil 15. Cinsiyet değişkenine bağlı Faith_pd değerinin dağılımı

Tablo 14. Cinsiyet değişkenine bağlı Faith_pd değerlerinin Mann Whitney-U analizine ait sonuçları

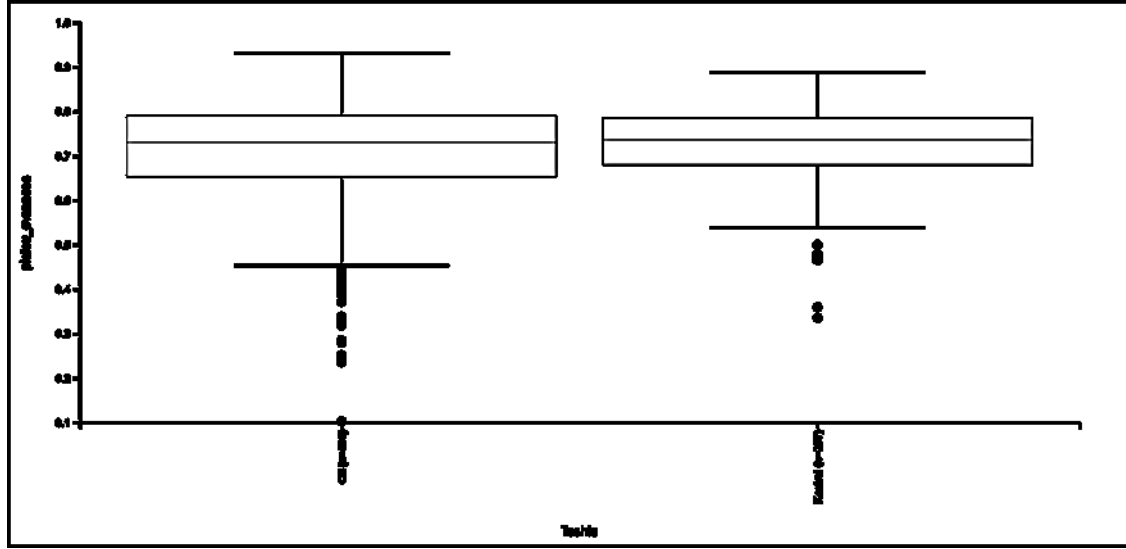
Grup 1	Grup 2	U	p değeri
Erkek (n=427)	Kadın (n=338)	72833	0.83



Şekil 16. Vücut bölgesi değişkenine bağlı Pielou düzgünlük değerinin dağılımı

Tablo 15. Vücut bölgesi değişkenine bağlı Pielou düzgünlük değerlerinin Mann Whitney-U analizine ait sonuçları

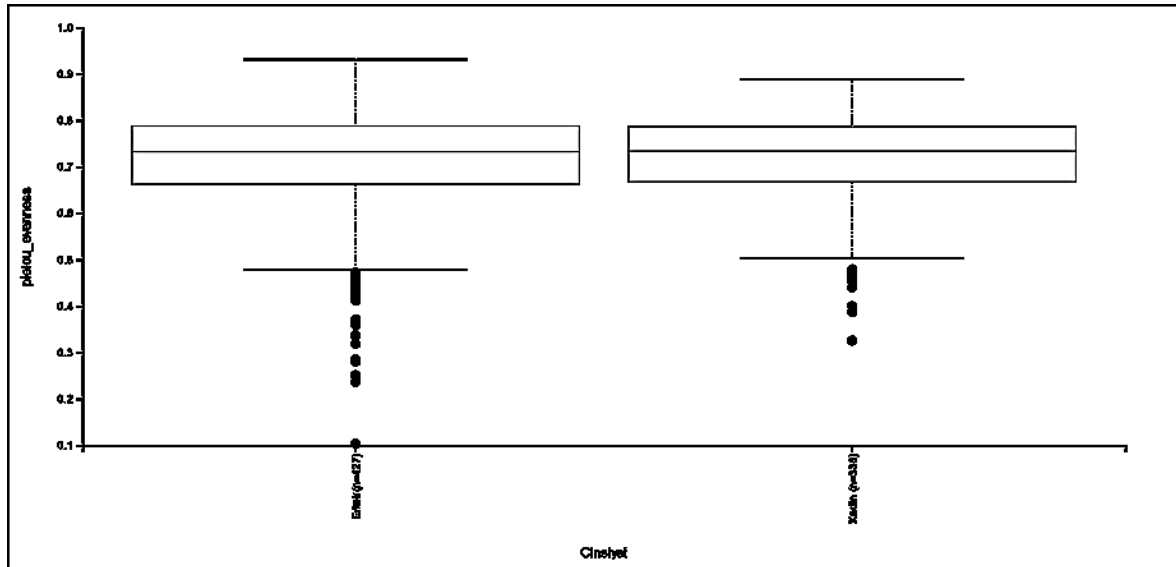
Grup 1	Grup 2	U	p değeri
İleum (n=505)	Rektum (n=260)	65699	0.99



Şekil 17. Teşhis değişkenine bağlı Pielou düzgünlük değerinin dağılımı

Tablo 16. Teşhis değişkenine bağlı Pielou düzgünlük değerlerinin Mann Whitney-U analizine ait sonuçları

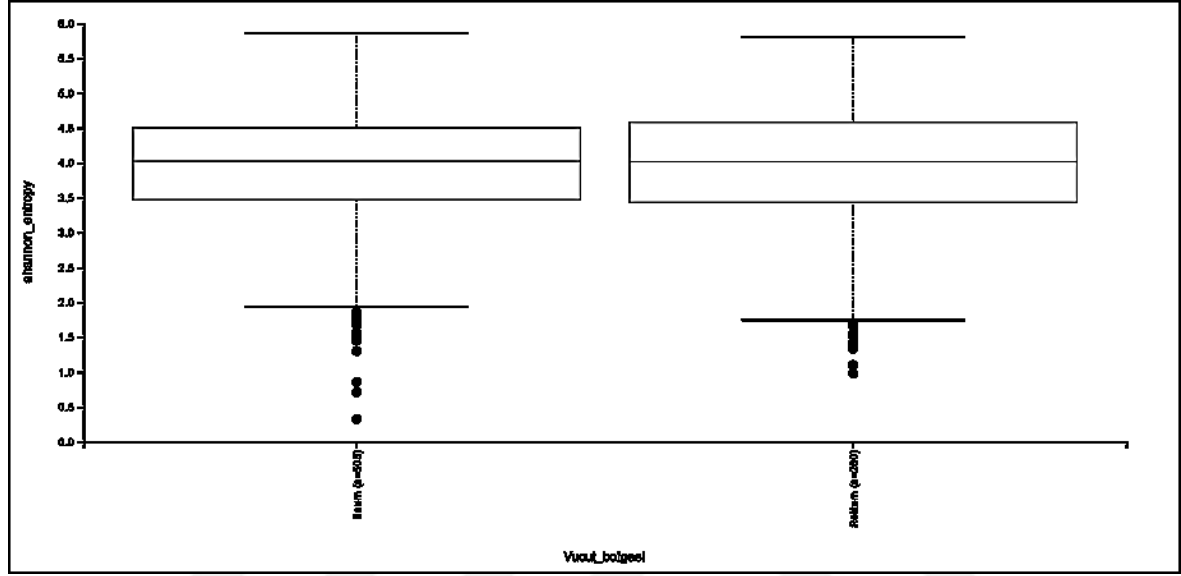
Grup 1	Grup 2	U	p değeri
CH (n=468)	Kontrol (n=297)	65325	0.16



Şekil 18. Cinsiyet değişkenine bağlı Pielou düzgünlük değerinin dağılımı

Tablo 17. Cinsiyet değişkenine bağlı Pielou düzgünlük değerlerinin Mann Whitney-U analizine ait sonuçları

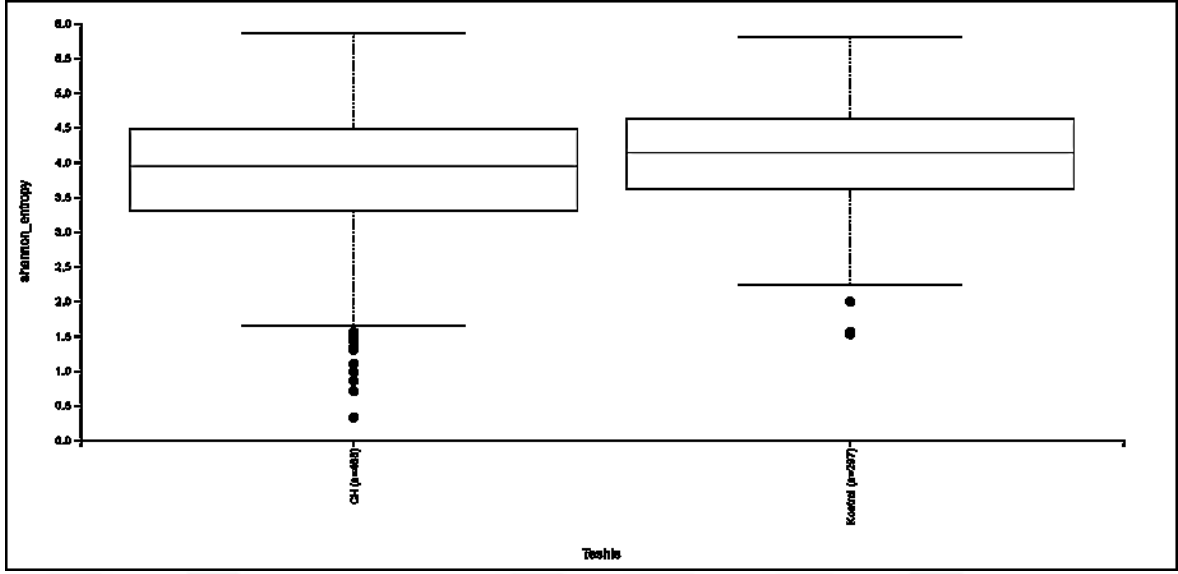
Grup 1	Grup 2	U	p değeri
Erkek (n=427)	Kadın (n=338)	71397	0.80



Şekil 19. Vücut bölgesi değişkenine bağlı Shannon Entropi değerinin dağılımı

Tablo 18. Vücut bölgesi değişkenine bağlı Shannon Entropi değerlerinin MannWhitney-U analizine ait sonuçları

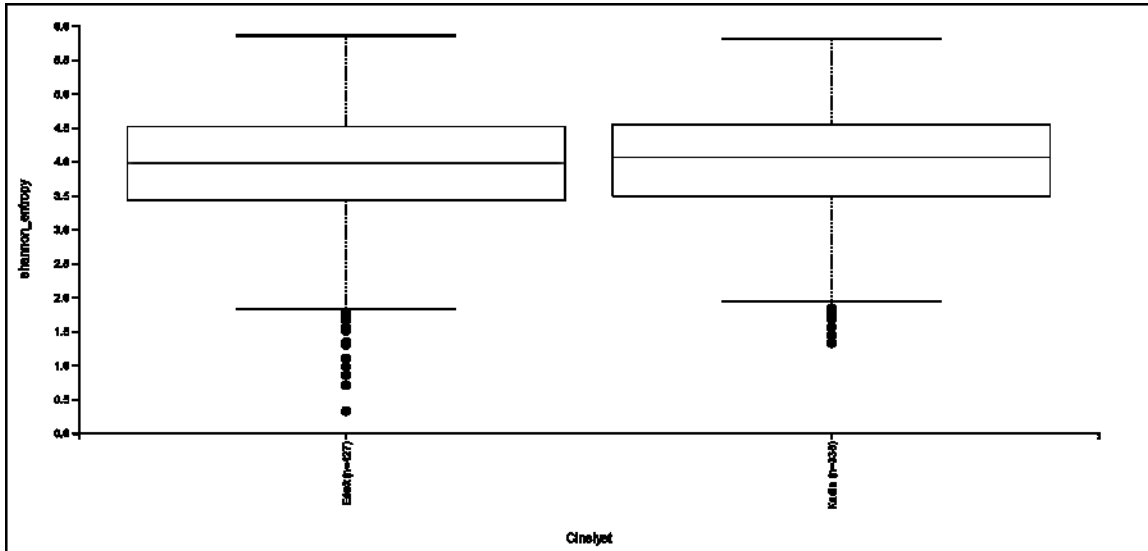
Grup 1	Grup 2	U	p değeri
İleum (n=505)	Rektum (n=260)	64676	0.74



Şekil 20. Teşhis değişkenine bağlı Shannon Entropi değerinin dağılımı

Tablo 19. Teşhis değişkenine bağlı Shannon Entropi değerlerinin Mann Whitney-U analizine ait sonuçları

Grup 1	Grup 2	U	p değeri
CH (n=468)	Kontrol (n=297)	81416	6.31e-5



Şekil 21. Cinsiyet değişkenine bağlı Shannon Entropi değerinin dağılımı

Tablo 20. Cinsiyet değişkenine bağlı Shannon Entropi değerlerinin Mann Whitney-U analizine ait sonuçları

Grup 1	Grup 2	U	p değeri
Erkek (n=427)	Kadın (n=338)	70050	0.49

Tablo 21’de vücut bölgesine bağlı olarak ağırlıklandırılmamış UniFrac Uzaklığındaki değişimin anlamlılığı PERMANOVA testi ile 999 permütasyon sayısı ile ölçülmüştür. Sahte -F değeri 2.92 olarak elde edilmişken $p=0.012$ olarak bulunmuştur. p değerinin 0.05 değerinden küçük olması sebebiyle vücut bölgelerine bağlı ağırlıklandırılmamış UniFrac Uzaklığındaki farkın anlamlı olduğu görülmektedir.

Tablo 21. Vücut bölgesi değişkenine ait ağırlıklandırılmamış UniFrac uzaklığı PERMANOVA sonuçları

Yöntem Adı	PERMANOVA
Test İstatistik Adı	sahte-F
Örneklem Büyüklüğü	765
Grup Sayısı	2
Test İstatistiği	2.92
p-Değeri	0.012
Permütasyon Sayısı	999

Tablo 22’de teşhise bağlı olarak ağırlıklandırılmamış UniFrac Uzaklığındaki değişimin anlamlılığı PERMANOVA testi ile 999 permütasyon sayısı ile ölçülmüştür. Sahte -F değeri 10.16 olarak elde edilmişken $p=0.001$ olarak bulunmuştur. p değerinin 0.05 değerinden küçük olması sebebiyle teşhise bağlı ağırlıklandırılmamış UniFrac Uzaklığındaki farkın anlamlı olduğu görülmektedir.

Tablo 22. Teşhis değişkenine ait ağırlıklandırılmamış UniFrac uzaklığı PERMANOVA sonuçları

Yöntem Adı	PERMANOVA
Test İstatistik Adı	sahte-F
Örneklem Büyüklüğü	765
Grup Sayısı	2
Test İstatistiği	10.16
p-Değeri	0.001
Permütasyon Sayısı	999

Tablo 23’de cinsiyete bağılı olarak ağırlıklandırılmamış UniFrac Uzaklığındaki değışimin anlamlılığı PERMANOVA testi ile 999 permütasyon sayısı ile ölçülmüştür. Sahte -F değeri 1.90 olarak elde edilmişken $p=0.058$ olarak bulunmuştur. p değerinin 0.05 değeriinden büyük olması sebebiyle cinsiyete bağılı ağırlıklandırılmamış UniFrac Uzaklığındaki farkın anlamlı olmadığı görülmektedir.

Tablo 23. Cinsiyet değışkenine ait ağırlıklandırılmamış UniFrac uzaklığı PERMANOVA sonuçları

Yöntem Adı	PERMANOVA
Test İstatistik Adı	sahte-F
Örneklem Büyüklüğü	765
Grup Sayısı	2
Test İstatistiğı	1.90
p-Değeri	0.058
Permütasyon Sayısı	999

Tablo 24’de vücut bölgesine bağılı olarak ağırlıklandırılmış UniFrac Uzaklığındaki değışimin anlamlılığı PERMANOVA testi ile 999 permütasyon sayısı ile ölçülmüştür. Sahte -F değeri 1.69 olarak elde edilmişken $p=0.144$ olarak bulunmuştur. p değerinin 0.05 değeriinden büyük olması sebebiyle vücut bölgesine bağılı ağırlıklandırılmış UniFrac Uzaklığındaki farkın anlamlı olmadığı görülmektedir.

Tablo 24. Vücut bölgesi değışkenine ait ağırlıklandırılmış UniFrac uzaklığı PERMANOVA sonuçları

Yöntem Adı	PERMANOVA
Test İstatistik Adı	sahte-F
Örneklem Büyüklüğü	765
Grup Sayısı	2
Test İstatistiğı	1.69
p-Değeri	0.144
Permütasyon Sayısı	999

Tablo 25’de teşhise bağılı olarak ağırlıklandırılmış UniFrac Uzaklığındaki değışimin anlamlılığı PERMANOVA testi ile 999 permütasyon sayısı ile ölçülmüştür. Sahte -F değeri

16.71 olarak elde edilmişken $p=0.001$ olarak bulunmuştur. p değerinin 0.05 değerinden küçük olması sebebiyle vücut bölgesine bağlı ağırlıklandırılmış UniFrac Uzaklığındaki farkın anlamlı olduğu görülmektedir.

Tablo 25. Teşhis değişkenine ait ağırlıklandırılmış UniFrac uzaklığı PERMANOVA sonuçları

Yöntem Adı	PERMANOVA
Test İstatistik Adı	sahte-F
Örneklem Büyüklüğü	765
Grup Sayısı	2
Test İstatistiği	16.71
p-Değeri	0.001
Permütasyon Sayısı	999

Tablo 26’de cinsiyete bağlı olarak ağırlıklandırılmış UniFrac Uzaklığındaki değişimin anlamlılığı PERMANOVA testi ile 999 permütasyon sayısı ile ölçülmüştür. Sahte -F değeri 1.48 olarak elde edilmişken $p=0.192$ olarak bulunmuştur. p değerinin 0.05 değerinden büyük olması sebebiyle vücut bölgesine bağlı ağırlıklandırılmış UniFrac Uzaklığındaki farkın anlamlı olmadığı görülmektedir.

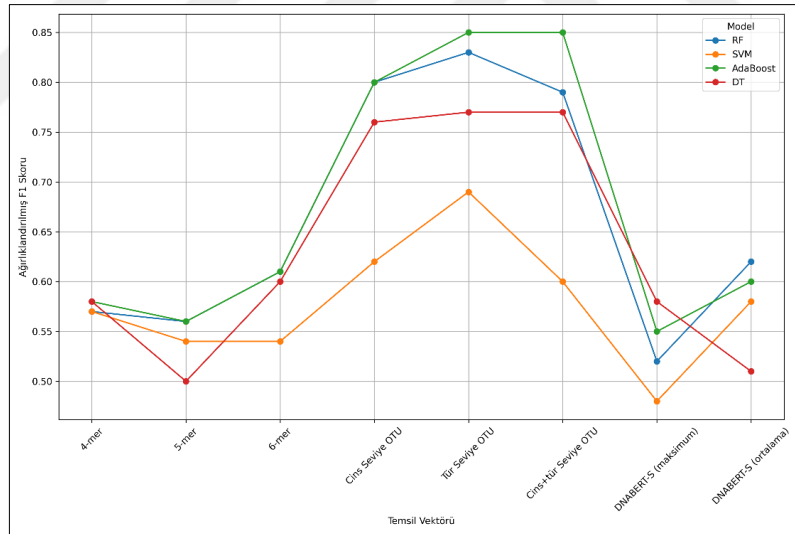
Tablo 26. Cinsiyet değişkenine ait ağırlıklandırılmış UniFrac uzaklığı PERMANOVA sonuçları

Yöntem Adı	PERMANOVA
Test İstatistik Adı	sahte-F
Örneklem Büyüklüğü	765
Grup Sayısı	2
Test İstatistiği	1.48
p-Değeri	0.192
Permütasyon Sayısı	999

4.2. Sınıflayıcılara Ait Sonuçlar

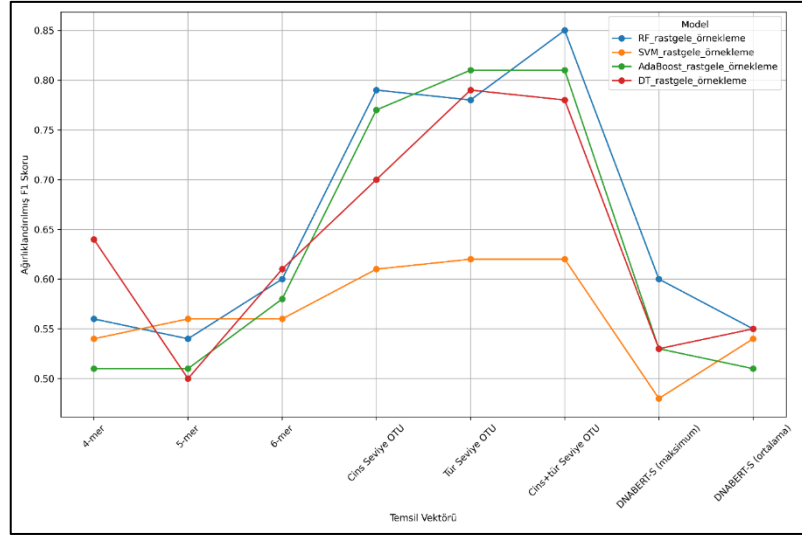
Tez çalışmasında veri seti %80 eğitim, %20 test kümesi oluşturulacak şekilde bölünmüştür. Oluşturulan eğitim verisi üzerinde GridSearchCV algoritması 10 kat çapraz

doğrulama ile birlikte kullanılarak en iyi tahminleyici model belirlenmiştir. Tablo 27’de verilen model sonuçları dikkate alındığında ağırlıklandırılmış F1 skoru açısından en iyi performans sonuçları tür seviye ASV tablolarının AdaBoost algoritması kullanılarak sınıflanmasında, tür seviye ASV ve cins seviye ASV tablolarının birleştirilmesi sonucu oluşan temsil vektörlerinin AdaBoost algoritmasıyla sınıflandırılmasında ve birleştirilen temsil vektörlerinin rastgele örnek seçilerek yerine koyma yaklaşımının rastgele orman algoritması ile birlikte uygulanmasıyla elde edilmiştir. Poligon alan matrisinin ağırlıklandırılmış skor değerleri ile kullanılmasıyla elde edilen en iyi skor PAM 0.73 olarak Adaboost algoritmasının birleştirilmiş ASV tabloları ile birlikte kullanılmasıyla elde edilmiştir. Crohn hastalığının teşhisine ait elde edilen en iyi performans sonuçları Ek 1’de gösterilmiştir. Ek 2’de ise Modelin parametre uzayının GridSearchCV algoritması ile taranması sonucu en iyi modele ait hiper parametre uzayı verilmiştir. Tüm modeller için hesaplanan poligon alan matrisinin şekilsel gösterimi Ek 3’te, modeller üzerinde hesaplanan Jaccard İndeks ve Poligon alan matrisine ait skorlar Ek 4’te gösterilmiştir.



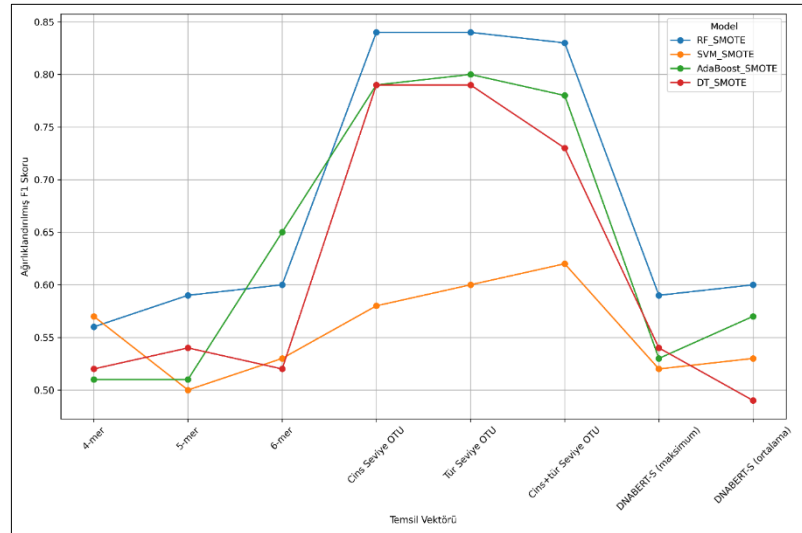
Şekil 22. Orijinal veri kümesi üzerinde eğitilmiş olan modellerin 8 farklı temsil vektörü ile çalıştırılması ile elde edilen ağırlıklandırılmış F1 skorlarına ait grafikler

Şekil 22 incelendiğinde en başarılı modelin birleştirilmiş (cins + tür seviye) OTU temsillerinin ve tür seviye AdaBoost Algoritması ile çalıştırılması ile F1 skor 0.85 olarak elde edilmiştir.



Şekil 23. Rastgele yeniden örnekleme yaklaşımı kullanılarak elde edilen veri kümesi üzerinde eğitilmiş olan modellerin 8 farklı temsil vektörü ile çalıştırılması ile elde edilen ağırlıklandırılmış F1 skorlarına ait grafikler

Şekil 23 incelendiğinde en başarılı modelin birleştirilmiş (cins + tür seviye) OTU temsillerinin rastgele orman algoritması ile çalıştırılması ile F1 skoru 0.85 olarak elde edilmiştir.

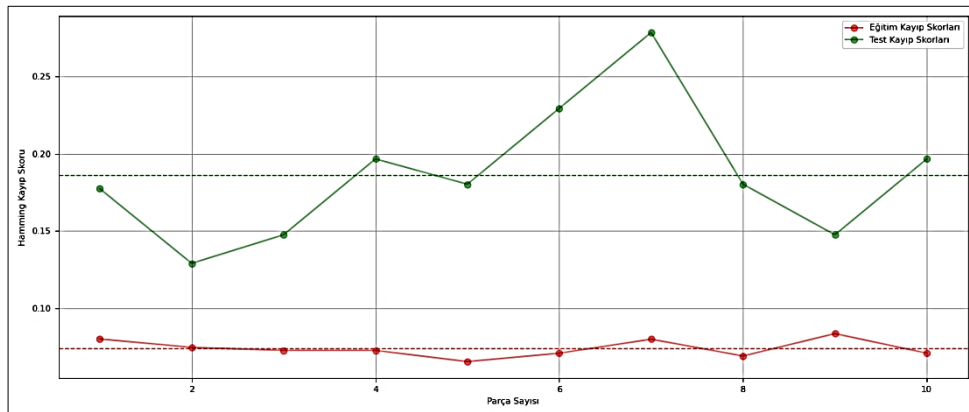


Şekil 24. SMOTE yeniden örnekleme yaklaşımı kullanılarak elde edilen veri kümesi üzerinde eğitilmiş olan modellerin 8 farklı temsil vektörü ile çalıştırılması ile elde edilen ağırlıklandırılmış F1 skorlarına ait grafikler

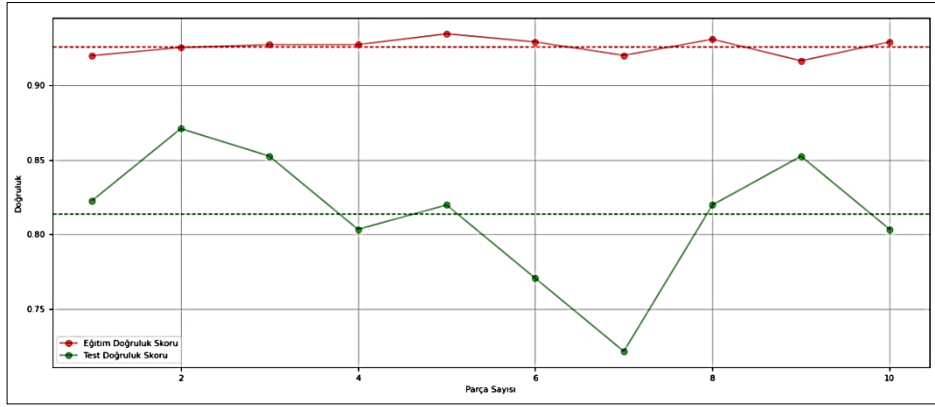
Şekil 24 incelendiğinde en başarılı modelin tür seviye OTU temsillerinin ve cins seviye OTU temsillerinin rastgele orman algoritması ile çalıştırılması ile F1 skoru 0.84 olarak elde edilmiştir.

Ağırlıklandırılmış F1 skorları açısından en iyi sonuçlar orijinal veri kümesi üzerinde birleştirilmiş (tür + cins) seviye OTU temsillerinin ve tür seviye OTU temsillerinin AdaBoost Algoritması ile sınıflandırılmasıyla elde edilmişken, rastgele yeniden örnekleme yaklaşımı kullanılan veri kümesinde eğitilmiş birleştirilmiş (tür + cins) seviye OTU temsillerinin Rastgele Orman Algoritması ile sınıflandırılmasıyla elde edilmiştir. Birçok metriği dikkate alan PAM skoru açısından en iyi sınıflama sonucu birleştirilmiş temsil vektörlerinin AdaBoost algoritması ile birlikte kullanılmasıyla 0.73 olarak hesaplanmıştır.

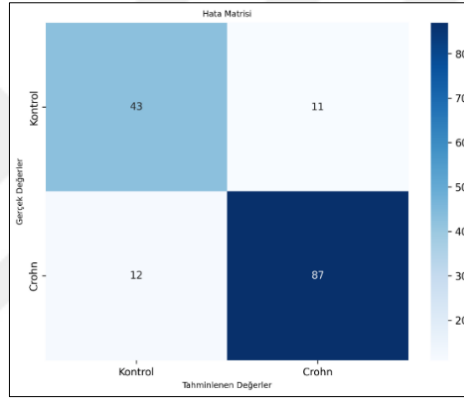
Tür seviye OTU tablolarının AdaBoost Algoritması kullanılması sonucu eğitim verisinde GridSearchCV algoritması kullanılarak en iyi tahminleyici ile elde edilen Hamming kayıp skoru Şekil 24’de gösterilmiştir. Doğruluk skoru Şekil 25’de gösterilmiştir. Tür seviye OTU tablolarının, AdaBoost algoritması kullanılarak test verisi üzerindeki sınıflandırma sonuçlarına ait hata matrisi Şekil 27’te, ROC eğrisi ise Şekil 28’de sunulmuştur.



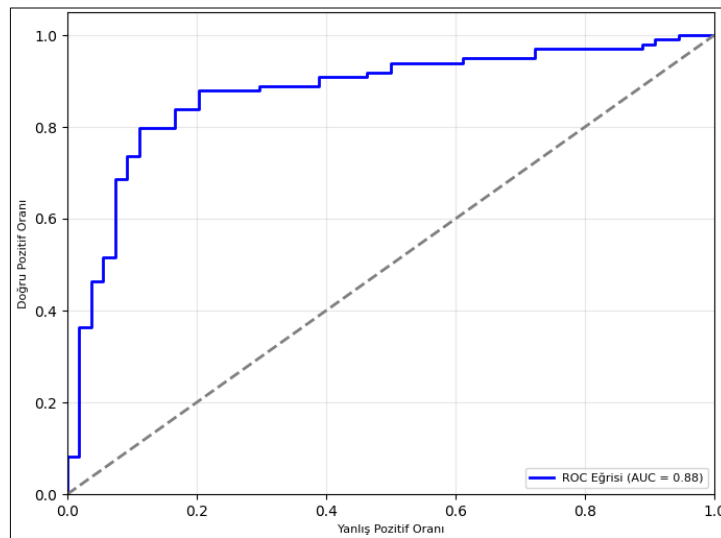
Şekil 25. Eğitim verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait Hamming kayıp skoru



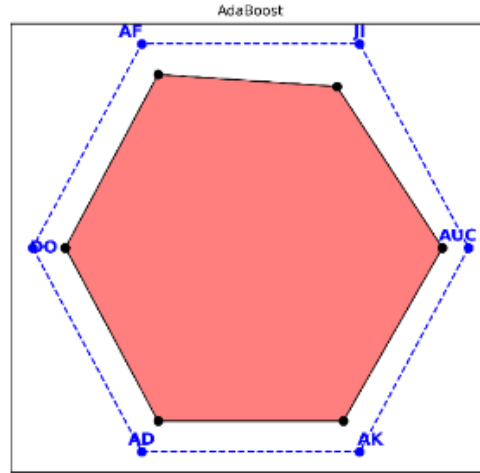
Şekil 26. Eğitim verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait doğruluk skoru



Şekil 27. Test verisi üzerinde oluşturulan tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait hata matrisi



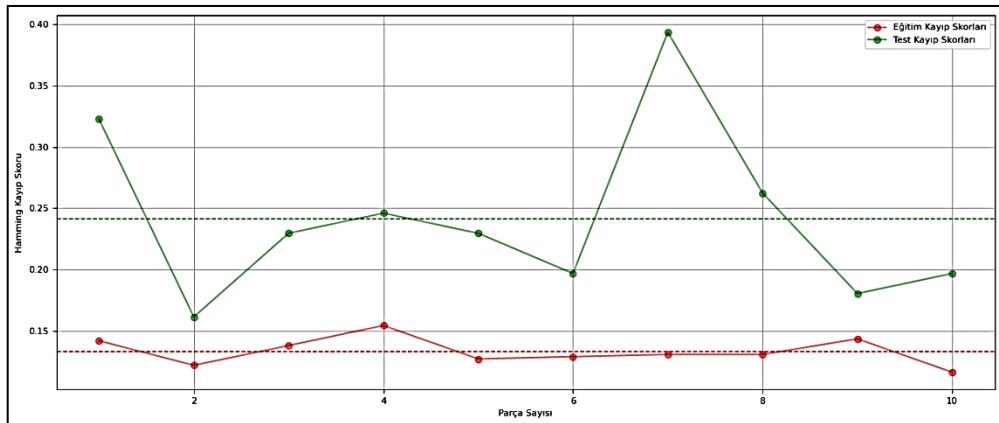
Şekil 28. Test verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait ROC eğrisi



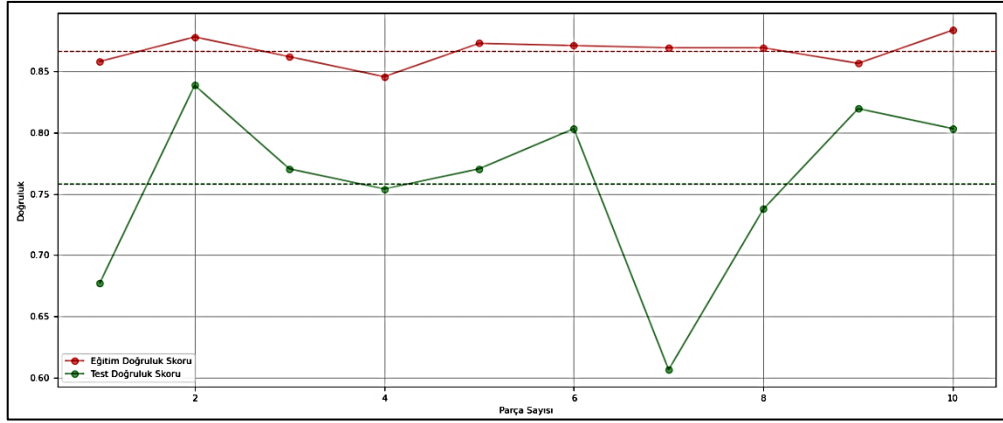
Şekil 29. Tür seviye OTU tabloları AdaBoost algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi

Tür seviye OTU tablolarının AdaBoost algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi skoru 0.71 olarak elde edilmiştir. Elde edilen skorların optimum olan 1 değerine yakın olduğu Şekil 29’da görülmektedir.

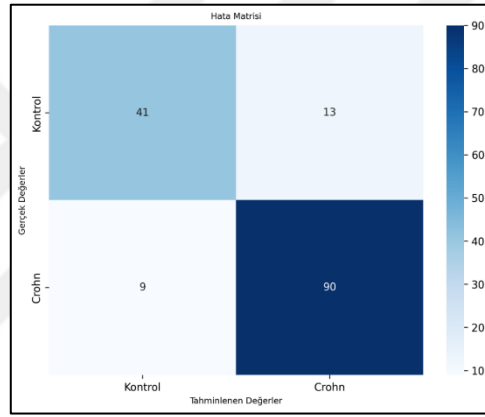
Tür seviyesi OTU tablolarının, cins seviyesi OTU tabloları ile birleştirilmesiyle oluşturulan birleştirilmiş tablolar, AdaBoost algoritması kullanılarak sınıflandırılmıştır. Eğitim verisinde, GridSearchCV algoritmasıyla en iyi tahminleyici seçilmiştir ve bu tahminleyici ile elde edilen Hamming kayıp skoru Şekil 30’da, doğruluk skoru ise Şekil 31’de gösterilmiştir. Birleştirilmiş OTU tablolarının, AdaBoost algoritması kullanılarak test verisi üzerindeki sınıflandırma sonuçlarına ait hata matrisi Şekil 32’de, ROC eğrisi ise Şekil 33’te sunulmuştur.



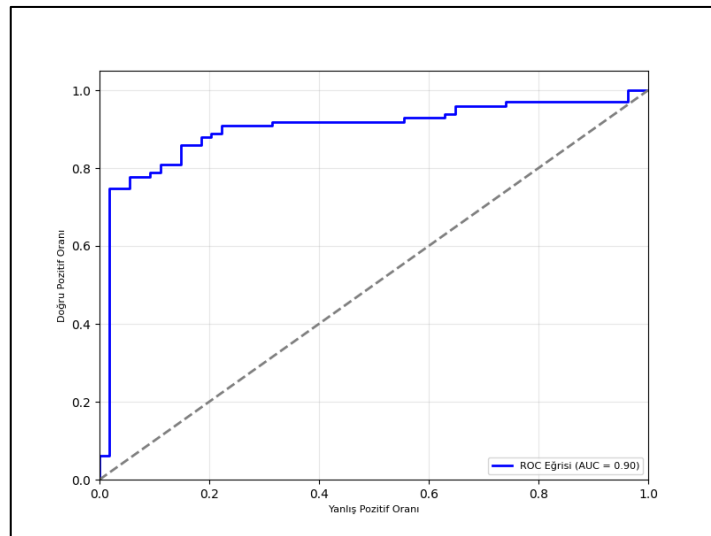
Şekil 30. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait Hamming kayıp skoru



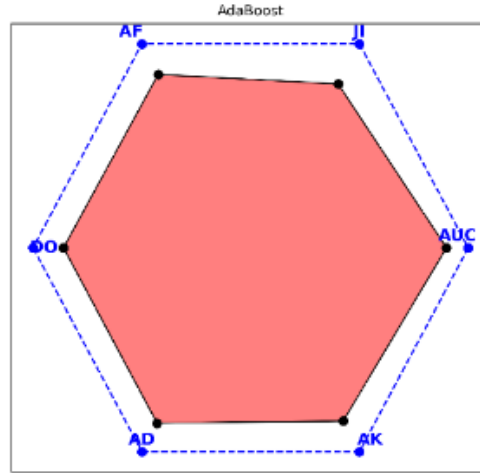
Şekil 31. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait doğruluk skoru



Şekil 32. Test verisi üzerinde oluşturulan birleştirilmiş OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait hata matrisi



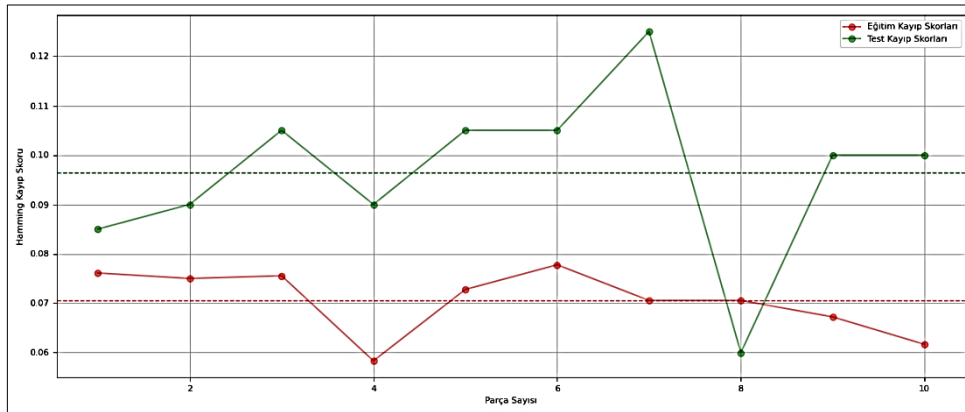
Şekil 33. Test verisi üzerinde tür seviye OTU tablolarının AdaBoost algoritması ile sınıflandırılmasına ait ROC eğrisi



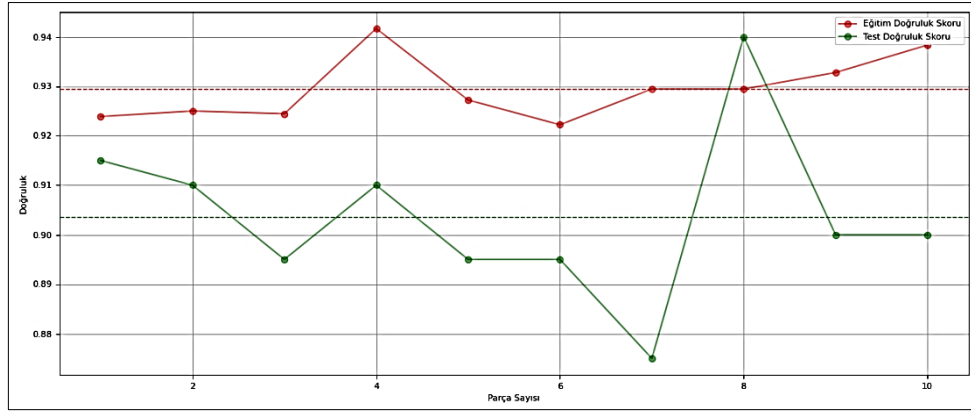
Şekil 34. Birleştirilmiş OTU tabloları AdaBoost algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi

Birleştirilmiş OTU tablolarının AdaBoost algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi skoru 0.73 olarak elde edilmiştir. Elde edilen skorların optimum olan 1 değerine yakın olduğu Şekil 34’te görülmektedir.

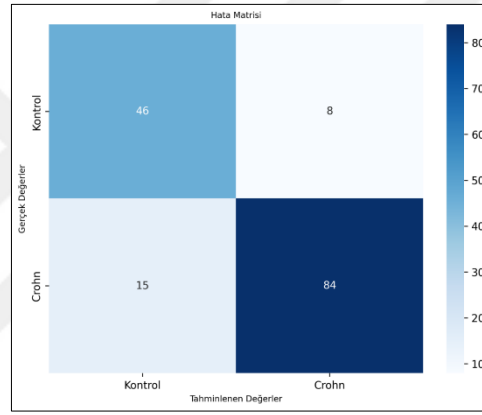
Tür seviyesi OTU tablolarının, cins seviyesi OTU tabloları ile birleştirilmesiyle oluşturulan birleştirilmiş tablolar, Rastgele Orman Algoritması kullanılarak sınıflandırılmıştır. Eğitim verisinde, GridSearchCV algoritmasıyla en iyi tahminleyici seçilmiştir ve bu tahminleyici ile elde edilen Hamming kayıp skoru Şekil 35’te, doğruluk skoru ise Şekil 36’te gösterilmiştir. Birleştirilmiş OTU tablolarının, Rastgele Orman Algoritması kullanılarak test verisi üzerindeki sınıflandırma sonuçlarına ait hata matrisi Şekil 37’de, ROC eğrisi ise Şekil 38’de sunulmuştur.



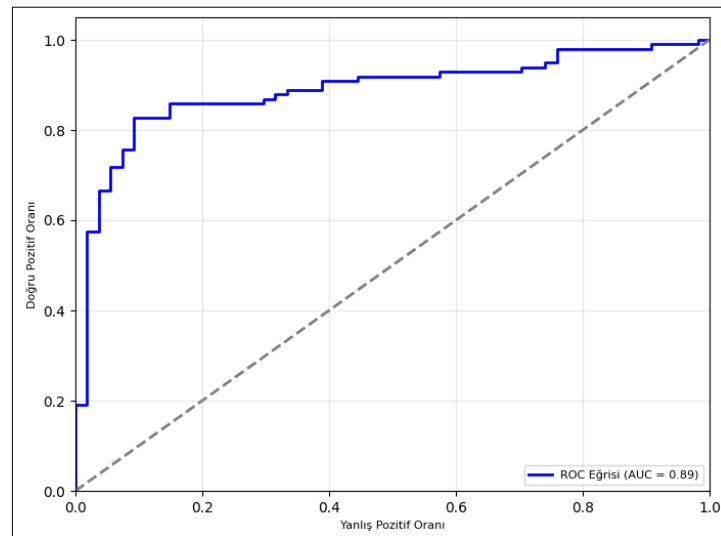
Şekil 35. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının Rastgele Orman algoritması ile sınıflandırılmasına ait Hamming kayıp skoru



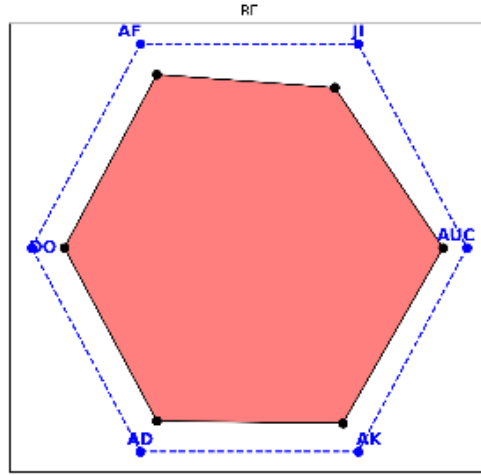
Şekil 36. Eğitim verisi üzerinde birleştirilmiş OTU tablolarının Rastgele Orman algoritması ile sınıflandırılmasına ait doğruluk skoru



Şekil 37. Test verisi üzerinde oluşturulan birleştirilmiş OTU tablolarının Rastgele Orman algoritması ile sınıflandırılmasına ait hata matrisi



Şekil 38. Test verisi üzerinde tür seviye OTU tablolarının Rastgele Orman algoritması ile sınıflandırılmasına ait ROC eğrisi



Şekil 39. Tür seviye OTU tabloları Rastgele orman algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi

Birleştirilmiş OTU tablolarının Rastgele orman algoritması ile birlikte sınıflandırılmasına ait Poligon alan matrisi skoru 0.72 olarak elde edilmiştir. Elde edilen skorların optimum olan 1 değerine yakın olduğu Şekil 39’da görülmektedir.

5. TARTIŞMA ve SONUÇ

Mikrobiyota, kardiyovasküler hastalıklar, kanser, solunum yolu hastalıkları, diyabet, İBH, nörolojik bozukluklar, kronik böbrek hastalıkları ve karaciğer hastalıklarının gelişimiyle ilişkilidir (66). Mikrobiyata ve ülseratif kolit hastalık şiddetinin ilişkisinin makine öğrenmesi yöntemleriyle değerlendirildiği (129, 130) ve Crohn hastalığı teşhisinin yapıldığı (131, 132) çalışmalar mevcuttur. Uygun maliyetli olması bu yöntemden faydalanılma gerekliliğinin önünü açmaktadır (71).

Yapılan tez çalışmasında amaç mikrobiyom verisi kullanılarak Crohn hastalığı teşhisine yardımcı olacak bir karar destek sisteminin geliştirilmesidir. Hastalığın erken teşhisi özellikle pediatrik hastalar için hastalık seyri için oldukça önemlidir. Yetişkin başlangıçlı Crohn hastalığı ile kıyaslandığında, infantil ve çok erken başlangıçlı hastalarda hastalık daha agresif bir fenotip sergilemektedir. Bu gruplar için alternatif hastalık yönetim stratejilerinin belirlenmesi gereklidir (11, 15). Ergen başlangıçlı CH, yetişkin başlangıçlı hastalarla daha çok benzerlik gösterse de, yetersiz beslenme, büyüme bozuklukları, ergenlik gecikmesi ve psikososyal zorluklar gibi özel durumlar mevcuttur (11, 15, 16).

Tez çalışmasının güçlü yönlerinden biri, geniş bir hasta grubundan (765 hasta) elde edilen veri setinin kullanılmasıdır. Ayrıca, farklı temsil vektörleri ve sınıflama modellerinin karşılaştırılması, en etkili yöntemin belirlenmesine olanak tanımıştır. Çalışmada, tür ve cins seviyesinde taksonomik temsil vektörlerinin birleştirilmesi, mikrobiyom verilerinin daha kapsamlı bir şekilde değerlendirilmesine olanak sağlamıştır. Bu birleştirme stratejisi, farklı taksonomik seviyelerdeki bilgilerin entegrasyonu ile sınıflama modellerinin performansını artırmayı hedeflemiştir. Ayrıca, DNABERT-S ön eğitilmiş modeli kullanılarak sekans verilerinden anlamlı özellikler çıkarılmış ve yeniden örnekleme yöntemleriyle veri dengesizliği sorunlarına yönelik çözümler uygulanmıştır.

Literatür incelemesi yapıldığında tez çalışmasında kullanılan veri kümesinin tamamı kullanılan çalışmada CH teşhisi için 3-8 mer temsil vektörleri farklı yeniden örnekleme boyutları kullanılarak rastgele orman algoritması, destek vektör makineleri ve derin sinir ağlar kullanılarak değerlendirilmiştir. Elde edilen en iyi sonuç 6-mer temsil vektörlerinin rastgele orman algoritması kullanılması ile ikili sınıflamada 0.76 mikro-F1 skoru, 0.75 makro-F1 skoru elde edilmiştir (91). Tez çalışmasında tür ve cins seviye OTU tablolarının birleştirildiği çalışmada AdaBoost Algoritması ile mikro-F1 skoru 0.86, makro-F1 skoru 0.84 olarak elde edilmiştir. Fakat çalışma ayrıntılı olarak incelendiğinde

tüm verinin Crohn hastalığı ve kontrol olarak sınıflandırılmadığı kontrol grubu yerine kontrol grubu ve diğer hastalıkların aynı grup içerisinde değerlendirildiği görülmektedir. Bu bağlamda tez çalışması verisi daha kısıtlı kalırken kontrol ve diğer hastalıkların aynı sınıf içerisinde değerlendirilmesinin doğru olmadığı ve model performansını etkileyen bir durum olduğu görülmektedir.

Tez verimizin kullanıldığı başka bir çalışmada Light Gradyan Artırma Modelinin (Light Gradient Boosting, LightGBM) 5-mer temsil vektörleriyle kullanılmasıyla 0.76 F skoru ve 0.82 AUC skoru elde edilmiştir. Sınıflama İBH ve sağlıklı bireyler arasında gerçekleştirilmiştir (133). Çalışmada İBH ve kontrol grubu arasında sınıflamadan bahsederken İskemik Kolit hastalığı İBH sınıfı içerisinde değerlendirilmiştir. Sınıf dengesizliği olan veride bu dengesizlik göz ardı edilmiştir. Ağırlıklandırılmış skorların kullanılması bu bağlamda doğru olabilecek bir senaryo iken ağırlıklandırılmış metrikler kullanılmamıştır. Yaptığımız tez çalışması veri dengesizliğini de dikkate almıştır.

Tez verisinde yapılan başka bir çalışmada CH tahmininde 7-mer temsil vektörleri ile 0.92 AUC skoru elde edilmiştir. Sınıflama Crohn ve kontrol grupları arasında yapılmış çalışmaya sadece pediatrik gruplar değil tüm Crohn hastaları dahil edilmiştir (134). Yapılan tez çalışması ise pediatrik Crohn hastalığı üzerinedir. Ayrıca çocukluk döneminde mikrobiyom çeşitliliğinin yetişkin hastalara göre anlamlı düzeyde farklıdır (135).

Başka bir çalışmada 16s rRNA verisi kullanılarak CH teşhisi için Sıralı Evrişimsel Dikkat Ağı özellik çıkarmak için kullanılırken SVM modeli sınıflama için kullanılmıştır. Çalışmada ortalama doğruluk değeri 0.80 ve kappa değeri 0.76 olarak elde edilmiştir (136). Başka bir çalışmada kolorektal kanser, kolorektal adenomlar, Crohn hastalığı, ülseratif kolit, irritabl bağırsak sendromu, obezite, kardiyovasküler hastalık, akut COVID19 sonrası sendromu ve sağlıklı bireylerden oluşan 2320 kişiden oluşan dışkı genomik veri kümesinde çok sınıflı rastgele orman algoritması kullanılarak alıcı işletim karakteristik eğrisi (AUROC) altındaki alan 0.90-0.99 (IQR, 0.91–0.94), 0.76-0.98 (IQR 0.83–0.95) özgüllük, 0.81 -0,95 (IQR, 0.87–0.93) duyarlılık elde edilmiştir (137).

Başka bir çalışmada 16S rRNA verisi üzerinde seyrek kısmi en küçük kareler ayırıcı analizi (sPLS-DA) yapılarak İBH ve sağlıklı kontroller arasında ve UC, CH arasında sınıflama gerçekleştirilmiştir. Veri kümesi 175 UC ,124 CH,99 kontrol verisinden oluşmaktadır. sPLS-DA modelinin İBH ve kontrol grupları arasında uygulanması ile

ortalama doğruluk 0.908 ve AUC 0.966 ve CH ve UC grupları arasından gerçekleştirilen sınıflamada 0.846 doğruluk ve 0.923 AUC skoru elde edilmiştir (138).

16S rRNA verisi üzerinde 729 İBH hastası ve 700 kontrol grubu üzerinde yapılan çalışmada Doğrusal Ayırıcı Analiz skoru > 3 ve Rastgele Orman algoritması kullanılması sonucu AUC skoru 0.80 olarak elde edilmiştir. Yüksek varyanslı OTUlar rastgele orman algoritması ile kullanılması sonucu 0.82 AUC skoru ve CH ve UC arasında ayrımın test edilebilmesi için 331 CH ve 141 UC verisi üzerinde LDA skoru > 3 ve rastgele orman algoritması kullanılması sonucu AUC skoru > 0.90 elde edilmiştir.

Tezin kısıtlılıkları değerlendirildiğinde CH ve Kontrol grupları üzerinde sınıflama işlemi gerçekleştirilmiştir. Literatürde, üç sınıflı sınıflama çalışmalarının başarı oranının düşük olması, bu tür analizlerin zorluklarını göstermektedir. Ayrıca, elimizdeki 16S rRNA sekanslarının farklı vücut bölgelerinden alınmış olması ve veri kümesindeki yaş grubu değişkenliği, mikrobiyom kompozisyonunda önemli varyasyonlara yol açmaktadır. Bu durum, veri kümesinde filtreleme işlemi gerektirmiş ve modelleme sürecinde bazı verilerin dışlanmasına neden olmuştur.

Ağırlıklandırılmış F1 skoru ve ağırlıklandırılmış AUC skoru ve PAM skoru dikkate alındığında çalışmada en iyi sonuçlar, AdaBoost algoritmasının cins ve tür seviyesindeki ASV (Amplicon Sequence Variant) tablolarının birleştirilmesiyle elde edilmiştir. Bu yaklaşım, farklı taksonomik seviyelerdeki bilgilerin entegre edilmesiyle daha zengin ve ayrıntılı bir veri temsili sağlamış, modelin performansını artırmıştır. Sonuç olarak, AdaBoost Algoritmasının birleştirilmiş ASV temsilleriyle sınıflandırılmasında 0.86 doğruluk, 0.86 ağırlıklandırılmış duyarlılık, 0.85 ağırlıklandırılmış kesinlik, 0.90 ağırlıklandırılmış AUC skoru ve 0.85 ağırlıklandırılmış F1 skoru elde edilmiştir. Bu sonuçlar dikkate alındığında, mikrobiyom verileriyle hastalık teşhisinde yüksek performans sergileyebileceğini ve farklı taksonomik seviyelerin birleştirilmesinin modelleme sürecinde kritik bir avantaj sağladığını göstermektedir.

Bu çalışma, pediatrik Crohn hastalığının teşhisinde mikrobiyom verilerinin potansiyelini ve sınıflama algoritmalarının etkinliğini vurgulamaktadır. Özellikle, AdaBoost algoritmasının cins ve tür seviyesindeki ASV temsilleriyle yüksek doğruluk ve duyarlılık sağlaması, mikrobiyom analizlerinin tanı araçları geliştirmedeki rolünü güçlendirmektedir. Bu yaklaşım, pediatrik hasta gruplarına özgü mikrobiyota dinamiklerinin değerlendirilmesine imkan tanıyarak, hastalığın erken teşhisi ve

bireyselleştirilmiş tedavi yönetimleri için önemli bir adım olarak öne çıkmaktadır. Elde edilen sonuçlar, bu yöntemin pediatrik Crohn hastalığında erken teşhis için umut vadeden bir yaklaşım olduğunu göstermektedir.



KAYNAKLAR

1. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* 42: 1118–1125.
2. Yuan Z, Ye J, Zhang BL and L (2016). Unraveling the role of autophagy regulation in Crohn's disease: from genetic mechanisms to potential therapeutics 2.
3. Dolinger M, Torres J, Vermeire S (2024). Crohn's disease. *The Lancet* 403: 1177–1191.
4. Walters TD, Griffiths AM (2009). Mechanisms of growth impairment in pediatric Crohn's disease. *Nature Reviews Gastroenterology and Hepatology* 6: 513–523.
5. Oliveira SB, Monteiro IM (2017). Diagnosis and management of inflammatory bowel disease in children. *BMJ (Clinical research ed)* 357: j2083.
6. Pasternak G, Chrzanowski G, Aebischer D, Myśliwiec A, Dynarowicz K, Bartusik-Aebischer D, Sosna B, Cieślar G, Kawczyk-Krupka A, Filip R (2023). Crohn's Disease: Basic Characteristics of the Disease, Diagnostic Methods, the Role of Biomarkers, and Analysis of Metalloproteinases: A Review. *Life* 13.
7. Carter M, Lim IIP (2024). Surgical management of pediatric Crohn's disease. *Seminars in Pediatric Surgery* 33: 151401.
8. Trindade E (2016). Pediatric Crohn's Disease and Surgery: Another Tool for the Treatment of a Complex Disease. *GE Portuguese Journal of Gastroenterology* 23: 181–182.
9. Aldars-García L, Chaparro M, Gisbert JP (2021). Systematic review: The gut microbiome and its potential clinical application in inflammatory bowel disease. *Microorganisms* (Vol. 9).
10. Baumgart DC, Sandborn WJ (2012). Crohn's disease. *Lancet* (London, England) 380: 1590–1605.
11. Kelsen J, Baldassano RN (2008). Inflammatory bowel disease: the difference between children and adults. *Inflammatory bowel diseases* 14 Suppl 2: 10–12.

12. Silverberg MS, Satsangi J, Ahmad T, Arnott IDR, Bernstein CN, Brant SR, Caprilli R, Colombel JF, Gasche C, Geboes K, et al. (2005). Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Canadian journal of gastroenterology = Journal canadien de gastroenterologie* 19 Suppl A: 5–36.
13. Basoglu IA, Karakoyun B (2023). Crohn's disease: Etiology, pathogenesis and treatment strategies. *Marmara Medical Journal* 36: 249–254.
14. Malmborg P, Grahnquist L, Lindholm J, Montgomery S, Hildebrand H (2013). Increasing incidence of paediatric inflammatory bowel disease in northern Stockholm county, 2002-2007. *Journal of Pediatric Gastroenterology and Nutrition* 57: 29–34.
15. Fontana T, Falco N, Torchia M, Tutino R, Gulotta G (2017). Bowel perforation in crohn's disease: Correlation between CDAI and claven-dindo scores. *Giornale di Chirurgia* 38: 303–312.
16. Ahid SUSM, Inor KYLES M, Oto ROES, Ornung CAAH (2006). Smoking and Inflammatory Bowel Disease : 81: 1462–1471.
17. Roda G, Chien Ng S, Kotze PG, Argollo M, Panaccione R, Spinelli A, Kaser A, Peyrin-Biroulet L, Danese S (2020). Crohn's disease. *Nature reviews Disease primers* 6: 22.
18. Sila S, Trivić I, Pavić AM, Niseteo T, Kolaček S, Hojsak I (2019). Nutritional status and food intake in pediatric patients with inflammatory bowel disease at diagnosis significantly differs from healthy controls. *European Journal of Pediatrics* 178: 1519–1527.
19. Ananthakrishnan AN, Khalili H, Konijeti GG, Higuchi LM, de Silva P, Korzenik JR, Fuchs CS, Willett WC, Richter JM, Chan AT (2013). A prospective study of long-term intake of dietary fiber and risk of Crohn's disease and ulcerative colitis. *Gastroenterology* 145: 970–977.
20. Khalili H, Håkansson N, Chan SS, Chen Y, Lochhead P, Ludvigsson JF, Chan AT, Hart AR, Olén O, Wolk A (2020). Adherence to a Mediterranean diet is associated with a lower risk of later-onset Crohn's disease: Results from two large prospective cohort studies. *Gut* 69: 1637–1644.

21. Mamula P, Grossman AB, Baldassano RN, Kelsen JR, Markowitz JE (2017). Pediatric Inflammatory Bowel Disease. *Pediatric Inflammatory Bowel Disease: Third Edition* 1–754. doi:10.1007/978-3-319-49215-5.
22. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L (2017). Crohn's disease. *The Lancet* 389: 1741–1755.
23. Ng SC, Tang W, Ching JY, Wong MAY, Chow CMO, Hui AJ, Wong TC, Leung VK, Tsang SW, Yu HONHO, et al. (2024). Incidence and Phenotype of Inflammatory Bowel Disease Based on Results From the Asia-Pacific Crohn's and Colitis Epidemiology Study 158–165. doi:10.1053/j.gastro.2013.04.007.
24. Kim HJ, Hann HJ, Hong SN, Kim KH, Ahn IM, Song JY, Lee SH, Ahn HS (2015). Incidence and natural course of inflammatory bowel disease in Korea, 2006-2012: A nationwide population-based study. *Inflammatory Bowel Diseases* 21: 623–630.
25. Park SH, Kim YJ, Rhee KH, Kim YH, Hong SN, Kim KH, Seo SI, Cha JM, Park SY, Jeong SK, et al. (2019). A 30-year trend analysis in the epidemiology of inflammatory bowel disease in the Songpa-Kangdong District of Seoul, Korea in 1986-2015. *Journal of Crohn's and Colitis* 13: 1410–1417.
26. Ananthakrishnan AN, Bernstein CN, Iliopoulos D, Macpherson A, Neurath MF, Ali RAR, Vavricka SR, Fiocchi C (2018). Environmental triggers in IBD: A review of progress and evidence. *Nature Reviews Gastroenterology and Hepatology* 15: 39–49.
27. Orholm M, Binder V, Sørensen TI, Rasmussen LP, Kyvik KO (2000). Concordance of inflammatory bowel disease among Danish twins. Results of a nationwide study. *Scandinavian journal of gastroenterology* 35: 1075–1081.
28. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491: 119–124.
29. Liu JZ, Sommeren S Van, Huang H, Ng SC, Alberts R (2016). Europe PMC Funders Group Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations 47: 979–986.
30. Jia K, Shen J (2024). Transcriptome-wide association studies associated with Crohn's disease: challenges and perspectives. *Cell and Bioscience* 14: 1–19.

31. Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, Andersen V, Andrews JM, Annesse V, Brand S, et al. (2016). Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* (London, England) 387: 156–167.
32. Hoefkens E, Nys K, John JM, Van Steen K, Arijs I, Van der Goten J, Van Assche G, Agostinis P, Rutgeerts P, Vermeire S, et al. (2013). Genetic association and functional role of Crohn disease risk alleles involved in microbial sensing, autophagy, and endoplasmic reticulum (ER) stress. *Autophagy* 9: 2046–2055.
33. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, Albrecht M, Mayr G, De La Vega FM, Briggs J, et al. (2007). A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature Genetics* 39: 207–211.
34. Wildenberg ME, Koelink PJ, Diederens K, Te Velde AA, Wolfkamp SCS, Nuij VJ, Peppelenbosch MP, Nobis M, Sansom OJ, Anderson KI, et al. (2017). The ATG16L1 risk allele associated with Crohn's disease results in a Rac1-dependent defect in dendritic cell migration that is corrected by thiopurines. *Mucosal immunology* 10: 352–360.
35. Uhlig HH, Powrie F (2018). Translating Immunology into Therapeutic Concepts for Inflammatory Bowel Disease. *Annual review of immunology* 36: 755–781.
36. Pazmandi J, Kalinichenko A, Ardy RC, Boztug K (2019). Early-onset inflammatory bowel disease as a model disease to identify key regulators of immune homeostasis mechanisms. *Immunological reviews* 287: 162–185.
37. Denson LA, Jurickova I, Karns R, Shaw KA, Cutler DJ, Okou DT, Dodd A, Quinn K, Mondal K, Aronow BJ, et al. (2018). Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. *Gastroenterology* 154: 2097–2110.
38. Verburgt CM, Ghiboub M, Benninga MA, de Jonge WJ, Van Limbergen JE (2021). Nutritional therapy strategies in pediatric Crohn's disease. *Nutrients* 13: 1–20.
39. Kappelman MD, Moore KR, Allen JK, Cook SF (2013). Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. *Digestive Diseases and Sciences* 58: 519–525.

40. Kelts DG, Grand RJ (1980). Inflammatory bowel disease in children and adolescents. *Current Problems in Pediatrics* 10: 1–40.
41. Van Limbergen J, Haskett J, Griffiths AM, Critch J, Huynh H, Ahmed N, De Bruyn JC, Issenman R, El-Matary W, Walters TD, et al. (2015). Toward enteral nutrition in the treatment of pediatric Crohn disease in Canada: A workshop to identify barriers and enablers. *Canadian Journal of Gastroenterology and Hepatology* 29: 351–356.
42. Allmen D von (2018). *Pediatric Crohn 's Disease* 31: 80–88.
43. MacKner LM, Greenley RN, Szigethy E, Herzer M, Deer K, Hommel KA (2013). Psychosocial issues in pediatric inflammatory bowel disease: Report of the north american society for pediatric gastroenterology, hepatology, and nutrition. *Journal of Pediatric Gastroenterology and Nutrition* 56: 449–458.
44. Ashton JJ, Boukas K, Davies J, Stafford IS, Vallejo AF, Haggarty R, Coelho TAF, Batra A, Afzal NA, Vadgama B, et al. (2021). Ileal Transcriptomic Analysis in Paediatric Crohn's Disease Reveals IL17- and NOD-signalling Expression Signatures in Treatment-naïve Patients and Identifies Epithelial Cells Driving Differentially Expressed Genes. *Journal of Crohn's & colitis* 15: 774–786.
45. Wędrychowicz A, Tomasik P, Kowalska-Duplaga K, Pieczarkowski S, Fyderek K (2021). Plasma elafin, cathelicidin, and α -defensins are increased in paediatric inflammatory Crohn's disease and reflect disease location. *Archives of medical science : AMS* 17: 1114–1117.
46. Clark-Snustad KD, Lee SD (2016). Diagnosing inflammatory bowel disease and differentiating it from potential mimics. *Techniques in Gastrointestinal Endoscopy* 18: 108–115.
47. Maaser C, Sturm A, Vavricka SR, Kucharzik T, Fiorino G, Annese V, Calabrese E, Baumgart DC, Bettenworth D, Borralho Nunes P, et al. (2019). ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 1: Initial diagnosis, monitoring of known IBD, detection of complications. *Journal of Crohn's and Colitis* 13: 144–164.
48. Park SJ, Kim WH (2012). A look into the small bowel in Crohn's disease. *Clinical Endoscopy* 45: 263–268.

49. Allocca M, Fiorino G, Danese S (2013). Cross-sectional imaging modalities in Crohn's disease. *Digestive Diseases* 31: 199–201.
50. Chatu S, Subramanian V, Pollok RCG (2012). Meta-analysis: Diagnostic medical radiation exposure in inflammatory bowel disease. *Alimentary Pharmacology and Therapeutics* 35: 529–539.
51. Horsthuis K, Bipat S, Bennink RJ, Stoker J (2008). Inflammatory bowel disease diagnosed with US, MR, scintigraphy, and CT: Meta-analysis of prospective studies. *Radiology* 247: 64–79.
52. Panés J, Bouzas R, Chaparro M, García-Sánchez V, Gisbert JP, Martínez De Guereñu B, Mendoza JL, Paredes JM, Quiroga S, Ripollés T, et al. (2011). Systematic review: The use of ultrasonography, computed tomography and magnetic resonance imaging for the diagnosis, assessment of activity and abdominal complications of Crohn's disease. *Alimentary Pharmacology and Therapeutics* 34: 125–145.
53. Sahni VA, Ahmad R, Burling D (2008). Which method is best for imaging of perianal fistula? *Abdominal Imaging* 33: 26–30.
54. Allocca M, Fiorino G, Bonifacio C, Furfaro F, Gilardi D, Argollo M, Peyrin-Biroulet L, Danese S (2018). Comparative accuracy of bowel ultrasound versus magnetic resonance enterography in combination with colonoscopy in assessing Crohn's disease and guiding clinical decision-making. *Journal of Crohn's and Colitis* 12: 1280–1287.
55. Magro F, Gionchetti P, Eliakim R, Ardizzone S, Armuzzi A, Barreiro-de Acosta M, Burisch J, Gece KB, Hart AL, Hindryckx P, et al. (2017). Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 1: Definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders. *Journal of Crohn's and Colitis* 11: 649–670.
56. Magro F, Langner C, Driessen A, Ensari A, Geboes K, Mantzaris GJ, Villanacci V, Becheanu G, Nunes PB, Cathomas G, et al. (2013). European consensus on the histopathology of inflammatory bowel disease. *Journal of Crohn's and Colitis* 7: 827–851.
57. Mitsuyama K, Niwa M, Takedatsu H, Yamasaki H, Kuwaki K, Yoshioka S, Yamauchi R, Fukunaga S, Torimura T (2016). Antibody markers in the diagnosis of inflammatory bowel disease. *World Journal of Gastroenterology* 22: 1304–1310.

58. Gu P, Kapur A, Li D, Haritunians T, Vasiliauskas E, Shih DQ, Targan SR, Spiegel BMR, McGovern DPB, Black JT, et al. (2018). Serological, genetic and clinical associations with increased health-care resource utilization in inflammatory bowel disease. *Journal of Digestive Diseases* 19: 15–23.
59. Bernstein CN, Hitchon CA, Walld R, Bolton JM, Sareen J, Walker JR, Graff LA, Patten SB, Singer A, Lix LM, et al. (2019). Increased Burden of Psychiatric Disorders in Inflammatory Bowel Disease. *Inflammatory Bowel Diseases* 25: 360–368.
60. Gomollón F, Dignass A, Annesse V, Tilg H, Van Assche G, Lindsay JO, Peyrin-Biroulet L, Cullen GJ, Daperno M, Kucharzik T, et al. (2017). 3rd European evidence-based consensus on the diagnosis and management of Crohn's disease 2016: Part 1: Diagnosis and medical management. *Journal of Crohn's and Colitis* 11: 3–25.
61. Vermeire S, Van Assche G, Rutgeerts P (2006). Laboratory markers in IBD: Useful, magic, or unnecessary toys? *Gut* 55: 426–431.
62. Carroccio A, Iacono G, Cottone M, Di Prima L, Cartabellotta F, Cavataio F, Scalici C, Montalto G, Di Fede G, Rini GB, et al. (2003). Diagnostic accuracy of fecal calprotectin assay in distinguishing organic causes of chronic diarrhea from irritable bowel syndrome: A prospective study in adults and children. *Clinical Chemistry* 49: 861–867.
63. Goutorbe F, Goutte M, Minet-Quinard R, Boucher AL, Pereira B, Bommelaer G, Buisson A (2015). Endoscopic Factors Influencing Fecal Calprotectin Value in Crohn's Disease. *Journal of Crohn's & colitis* 9: 1113–1119.
64. Guo S, Lu Y, Xu B, Wang W, Xu J, Zhang G (2019). A simple fecal bacterial marker panel for the diagnosis of Crohn's disease. *Frontiers in Microbiology* 10: 1–9.
65. Presti A Lo, Zorzi F, Chierico F Del, Altomare A, Cocca S, Avola A, De Biasio F, Russo A, Cella E, Reddel S, et al. (2019). Fecal and mucosal microbiota profiling in irritable bowel syndrome and inflammatory bowel disease. *Frontiers in Microbiology* 10: 1–14.
66. Hou K, Wu ZX, Chen XY, Wang JQ, Zhang D, Xiao C, Zhu D, Koya JB, Wei L, Li J, et al. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy* 7.

67. Purchiaroni F, Tortora A, Gabrielli M, Bertucci F, Gigante G, Ianiro G, Ojetti V, Scarpellini E, Gasbarrini A (2013). The role of intestinal microbiota and the immune system. *European Review for Medical and Pharmacological Sciences* 17: 323–333.
68. Madhogaria B, Bhowmik P, Kundu A (2022). Correlation between human gut microbiome and diseases. *Infectious Medicine* 1: 180–191.
69. Ipci K, Altıntoprak N, Muluk NB, Senturk M, Cingi C (2017). The possible mechanisms of the human microbiome in allergic diseases. *European Archives of Oto-Rhino-Laryngology* 274: 617–626.
70. Berman S, Petriz B, Kajeniene A, Prestes J, Castell L, Franco OL (2015). The microbiota: An exercise immunology perspective. *Exercise Immunology Review* 21: 70–79.
71. Peterson CB, Saha S, Do KA (2024). Analysis of Microbiome Data. *Annual Review of Statistics and Its Application* 11: 483–504.
72. Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, Salzberg SL, Steinegger M (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols* 17: 2815–2839.
73. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology* 41: 1633–1644.
74. Chao, Anne (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics* 11: 265–270.
75. Faith DP (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1–10.
76. Pielou EC (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology* 13: 131–144.
77. Shannon CE (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27: 379–423.

78. Shannon CE, Weaver W (1949). The mathematical theory of communication. The mathematical theory of communication Champaign, IL, US: University of Illinois Press.
79. Deng Y, Umbach AK, Neufeld JD (2024). Nonparametric richness estimators Chao1 and ACE must not be used with amplicon sequence variant data. *ISME Journal* 18: 7–9.
80. Janja Francé, Ioanna Varkitzi, Elena Stanca, Francesco Cozzoli, Sanda Skejić, Nicola Ungaro, Ivano Vascotto, Patricija Mozetič, Živana Ninčević Gladan, Georgia Assimakopoulou, Alexandra Pavlidou, Soultana Zervoudaki, Kalliopi Pagou AB (2021). Large-scale testing of phytoplankton diversity indices for environmental assessment in Mediterranean sub-regions. *Ecological Indicators* 126.
81. Maziarz M, Pfeiffer RM, Wan Y, Gail MH (2018). Using standard microbiome reference groups to simplify beta-diversity analyses and facilitate independent validation. *Bioinformatics* 34: 3249–3257.
82. Jaccard P (1912). the Distribution of the Flora in the Alpine Zone. *New Phytologist* 11: 37–50.
83. Lozupone C, Knight R (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* 71: 8228–8235.
84. Lozupone CA, Hamady M, Kelley ST, Knight R (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* 73: 1576–1585.
85. Chang Q, Luan Y, Sun F (2011). Variance adjusted weighted UniFrac: a powerful beta diversity measure for. *BMC Bioinformatics* 12: 118.
86. Wong RG, Wu JR, Gloor GB (2016). Expanding the UniFrac toolbox. *PLoS ONE* 11: 1–20.
87. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28: 2106–2113.

88. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, et al. (2014). The treatment-naïve microbiome in new-onset Crohn's disease. *Cell host & microbe* 15: 382–392.
89. Aun E, Brauer A, Kisand V, Tenson T RM (2018). A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. *PLoS Comput Biol* 10: E3574–E3581.
90. Ross EM, Moate PJ, Marett LC, Cocks BG, Hayes BJ (2013). Metagenomic predictions: from microbiome to complex health and environmental phenotypes in humans and cattle. *PloS one* 8: e73056.
91. Asgari E, Garakani K, McHardy AC, Mofrad MRK (2018). MicroPheno: predicting environments and host phenotypes from 16S rRNA gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics (Oxford, England)* 34: i32–i42.
92. Salton G, Yu CT (1974). On the Construction of Effective Vocabularies for Information Retrieval. *ACM SIGIR Forum* 9: 48–60.
93. C. -z. Liu, Y. -x. Sheng Z -q. W and Y-QY (2018). Research of Text Classification Based on Improved TF-IDF Algorithm. 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE) IEEE, pp 218–222. doi:10.1088/1742-6596/2171/1/012021.
94. Kawulok J, Deorowicz S (2015). CoMeta: Classification of metagenomes using k-mers. *PLoS ONE* 10: 1–23.
95. Bokulich NA (2024). Integrating sequence composition information into microbial diversity analyses with k-mer frequency counting 1–25.
96. Fuhl W, Zabel S, Nieselt K (2023). Resource saving taxonomy classification with k-mer distributions and machine learning 1–11.
97. Van Etten J, Stephens TG, Bhattacharya D (2023). A k-mer-Based Approach for Phylogenetic Classification of Taxa in Environmental Genomic Data. *Systematic Biology* 72: 1101–1118.

98. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37: 852–857.
99. Hall M, Beiko RG (2018). 16S rRNA Gene Analysis with QIIME2. *Methods in Molecular Biology* 1849: 113–129.
100. Rideout JR, Chase JH, Bolyen E, Ackermann G, González A, Knight R, Caporaso GG (2016). Keemei: Cloud-based validation of tabular bioinformatics file formats in Google Sheets. *GigaScience* 5.
101. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nat Methods* 13: 581–583.
102. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP (n.d.). DADA2: High resolution sample inference from amplicon data. doi:10.1101/024034.
103. Estaki M, Jiang L, Bokulich NA, McDonald D, González A, Kosciulek T, Martino C, Zhu Q, Birmingham A, Vázquez-Baeza Y, et al. (2020). QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols in Bioinformatics* 70: 1–46.
104. Callahan BJ, McMurdie PJ, Holmes SP (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal* 11: 2639–2643.
105. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. (2019, August). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology* 37.
106. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* 6: 1–17.
107. Barupal DK, Fiehn O (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 127: 2825–2830.

108. McKinney W (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 1: 56–61.
109. McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, Morton JT, Nicolaou G, Parks DH, Karst SM, et al. (2024). Greengenes2 unifies microbial data in a single reference tree. *Nature Biotechnology* 42: 715–718.
110. Zhou Z, Ji Y, Li W, Dutta P, Davuluri R V., Liu H (2024). Dnabert-2: Efficient Foundation Model and Benchmark for Multi-Species Genomes. 12th International Conference on Learning Representations, ICLR 2024 1–23.
111. Zhou Z, Wu W, Ho H, Wang J, Shi L, Davuluri R V, Wang Z, Liu H (2024). DNABERT-S: Learning Species-Aware DNA Embedding with Genome Foundation Models 1–21.
112. Chen T, Kornblith S, Norouzi M, Hinton GE (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Icml* 1597–1607.
113. Verma V, Lamb A, Beckham C, Najafi A, Mitliagkas I, Lopez-Paz D, Bengio Y (2019). Manifold mixup: Better representations by interpolating hidden states. 36th International Conference on Machine Learning, ICML 2019 2019-June: 11196–11205.
114. Li B, Zhou H, He J, Wang M, Yang Y, Li L (2020). On the sentence embeddings from pre-trained language models. EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference 9119–9130. doi:10.18653/v1/2020.emnlp-main.733.
115. Zhang D, Li SW, Xiao W, Zhu H, Nallapati R, Arnold AO, Xiang B (2021). Pairwise Supervised Contrastive Learning of Sentence Representations. EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings 5786–5798. doi:10.18653/v1/2021.emnlp-main.467.
116. Wang S, Dai Y, Shen J, Xuan J (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports* 11: 1–11.
117. G Menardi, N. Torelli, “Training and assessing classification rules with imbalanced data,” *Data Mining and Knowledge Discovery*, 28(1), pp.92-122, 2014.

118. Shihong Y, Ping L, Peiyi H (2003). Svm Classification :Its Contents And Challenges. *Appl Math Chin Univ* 18: 332–342.
119. Rath SK, Sahu M, Das SP, Bisoy SK, Sain M (2022). A Comparative Analysis of SVM and ELM Classification on Software Reliability Prediction Model. *Electronics (Switzerland)* 11.
120. Nalepa J, Kawulok M (2019). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review* 52: 857–900.
121. Parmar A, Katariya R, Patel V (2019). A Review on Random Forest: An Ensemble Classifier. *Lecture Notes on Data Engineering and Communications Technologies* 26: 758–763.
122. Belgiu M, Drăgu L (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31.
123. Chowdhury MS (2024). Comparison of accuracy and reliability of random forest, support vector machine, artificial neural network and maximum likelihood method in land use/cover classification of urban setting. *Environmental Challenges* 14.
124. An TK, Kim MH (2010). A new Diverse AdaBoost classifier. *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010* 1: 359–363.
125. Lee CS, Cheang PYS, Moslehpour M (2022). Predictive Analytics in Business Analytics: Decision Tree. *Advances in Decision Sciences* 26: 1–29.
126. Aydemir, O. A New Performance Evaluation Metric for Classifiers: Polygon Area Metric. *J Classif* 38, 16–26 (2021). <https://doi.org/10.1007/s00357-020-09362-5>
127. Bergstra, James, Bengio Y (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13: 281–305.
128. Page ES, Bellman R (1962). Adaptive Control Processes: A Guided Tour. *Journal of the Royal Statistical Society Series A (General)* 125: 161.
129. Lu J, Wang Z, Maimaiti M, Hui W, Abudourexiti A, Gao F (2022). Identification of diagnostic signatures in ulcerative colitis patients via bioinformatic analysis integrated with machine learning. *Human Cell* 35: 179–188.

130. Barberio B, Facchin S, Patuzzi I, Ford AC, Massimi D, Valle G, Sattin E, Simionati B, Bertazzo E, Zingone F, et al. (2022). A specific microbiota signature is associated to various degrees of ulcerative colitis as assessed by a machine learning approach. *Gut Microbes* 14.
131. Kim H, Na JE, Kim S, Kim TO, Park SK, Lee CW, Kim KO, Seo GS, Kim MS, Cha JM, et al. (2024). A Machine Learning-Based Diagnostic Model for Crohn's Disease and Ulcerative Colitis Utilizing Fecal Microbiome Analysis. *Microorganisms* 12.
132. Lee S, Lee I (2024). Comprehensive assessment of machine learning methods for diagnosing gastrointestinal diseases through whole metagenome sequencing data. *Gut Microbes* 16.
133. Unal M, Bostanci E, Ozkul C, Acici K, Asuroglu T, Guzel MS (2023). Crohn's Disease Prediction Using Sequence Based Machine Learning Analysis of Human Microbiome. *Diagnostics* 13.
134. Song, K., Wright, F. A., & Zhou, Y. H. (2020). Systematic Comparisons for Composition Profiles, Taxonomic Levels, and Machine Learning Methods for Microbiome-Based Disease Prediction. *Frontiers in molecular biosciences*, 7, 610845. <https://doi.org/10.3389/fmolb.2020.610845>
135. Radjabzadeh D, Boer CG, Beth SA, van der Wal P, Kieffe-De Jong JC, Jansen MAE, Konstantinov SR, Peppelenbosch MP, Hays JP, Jaddoe VWV, et al. (2020). Diversity, compositional and functional differences between gut microbiota of children and adults. *Scientific Reports* 10: 1–13.
136. Tian H, Tang R (2024). Prediction of Crohn's disease based on deep feature recognition. *Computational Biology and Chemistry* 113: 108231.
137. Su Q, Liu Q, Lau RI, Zhang J, Xu Z, Yeoh YK, Leung TWH, Tang W, Zhang L, Liang JQY, et al. (2022). Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nature Communications* 13: 1–8.
138. Kang SB, Kim H, Kim S, Kim J, Park SK, Lee CW, Kim KO, Seo GS, Kim MS, Cha JM, et al. (2023). Potential Oral Microbial Markers for Differential Diagnosis of Crohn's Disease and Ulcerative Colitis Using Machine Learning Models. *Microorganisms* 11.



Ek 1. Model Sonuçları

Temsil Vektörü	Yeniden Örnekleme	Model	Doğruluk	Hamming Kaybı	Makro-Metrikler				Mikro-Metrikler				Ağırlıklandırılmış Metrikler			
					F1	AUC	Duyarlılık	Kesinlik	F1	AUC	Duyarlılık	Kesinlik	F1	AUC	Duyarlılık	Kesinlik
4-mer	Yok	RF	0.59	0.41	0.48	0.56	0.51	0.52	0.59	0.56	0.59	0.59	0.57	0.56	0.59	0.56
	Rastgele yeniden örnekleme (yerine koyma ile)		0.56	0.44	0.54	0.55	0.54	0.54	0.56	0.55	0.56	0.56	0.56	0.55	0.56	0.58
	Smote		0.55	0.45	0.54	0.56	0.55	0.54	0.55	0.56	0.55	0.55	0.56	0.56	0.55	0.59
	Yok	SVM	0.64	0.36	0.49	0.55	0.52	0.56	0.64	0.55	0.64	0.64	0.57	0.55	0.64	0.59
	Rastgele yeniden örnekleme (yerine koyma ile)		0.65	0.35	0.44	0.56	0.51	0.58	0.65	0.56	0.65	0.65	0.54	0.56	0.65	0.60
	Smote		0.65	0.35	0.49	0.50	0.53	0.58	0.65	0.50	0.65	0.65	0.57	0.50	0.65	0.60
	Yok	AdaBoost	0.58	0.42	0.54	0.55	0.55	0.54	0.58	0.55	0.58	0.58	0.58	0.55	0.58	0.58
	Rastgele yeniden örnekleme (yerine koyma ile)		0.5	0.5	0.48	0.47	0.49	0.49	0.5	0.47	0.5	0.5	0.51	0.47	0.5	0.53
	Smote		0.50	0.50	0.49	0.49	0.51	0.51	0.50	0.49	0.50	0.50	0.51	0.49	0.50	0.55
	Yok	Karar Ağacı	0.58	0.42	0.54	0.53	0.54	0.54	0.58	0.53	0.58	0.58	0.58	0.53	0.58	0.58
	Rastgele yeniden örnekleme (yerine koyma ile)		0.63	0.37	0.61	0.61	0.61	0.61	0.63	0.61	0.63	0.63	0.64	0.61	0.63	0.64
	Smote		0.51	0.49	0.50	0.51	0.52	0.51	0.51	0.51	0.51	0.51	0.52	0.51	0.51	0.56
5mer	Yok	RF	0.60	0.40	0.50	0.54	0.51	0.52	0.60	0.54	0.60	0.60	0.56	0.54	0.60	0.56
	Rastgele yeniden örnekleme (yerine koyma ile)		0.53	0.47	0.51	0.49	0.51	0.51	0.53	0.49	0.53	0.53	0.54	0.49	0.53	0.56
	Smote		0.59	0.49	0.56	0.50	0.56	0.56	0.59	0.50	0.59	0.59	0.59	0.50	0.59	0.60
	Yok	SVM	0.59	0.41	0.46	0.48	0.49	0.48	0.59	0.48	0.59	0.59	0.54	0.48	0.59	0.53
	Rastgele yeniden örnekleme (yerine koyma ile)		0.65	0.35	0.47	0.53	0.52	0.58	0.65	0.53	0.65	0.65	0.56	0.53	0.65	0.60
	Smote		0.51	0.49	0.49	0.49	0.51	0.51	0.49	0.49	0.49	0.49	0.50	0.49	0.49	0.56
	Yok	AdaBoost	0.59	0.41	0.5	0.54	0.51	0.51	0.59	0.54	0.59	0.59	0.56	0.54	0.59	0.55
	Rastgele yeniden örnekleme (yerine koyma ile)		0.5	0.5	0.48	0.44	0.48	0.49	0.5	0.44	0.5	0.5	0.51	0.44	0.5	0.53
	Smote		0.5	0.5	0.46	0.48	0.46	0.46	0.50	0.48	0.50	0.50	0.51	0.48	0.50	0.51
	Yok	Karar Ağacı	0.49	0.51	0.49	0.49	0.51	0.51	0.49	0.49	0.49	0.49	0.5	0.49	0.49	0.55
	Rastgele yeniden örnekleme (yerine koyma ile)		0.48	0.52	0.47	0.47	0.48	0.48	0.48	0.47	0.48	0.48	0.5	0.47	0.48	0.52
	Smote		0.53	0.47	0.52	0.59	0.54	0.53	0.53	0.59	0.53	0.53	0.54	0.59	0.53	0.58
6mer	Yok	RF	0.62	0.38	0.55	0.59	0.56	0.57	0.62	0.59	0.62	0.62	0.61	0.59	0.62	0.60

Ek 1. (Devam)

	Rastgele yeniden örnekleme (yerine koyma ile)	SVM	0.59	0.41	0.59	0.65	0.60	0.59	0.59	0.65	0.59	0.59	0.60	0.65	0.59	0.64
	Smote		0.59	0.41	0.59	0.66	0.62	0.61	0.59	0.66	0.59	0.59	0.60	0.66	0.59	0.66
	Yok		0.61	0.39	0.45	0.53	0.50	0.49	0.61	0.53	0.61	0.61	0.54	0.53	0.61	0.53
	Rastgele yeniden örnekleme (yerine koyma ile)		0.58	0.42	0.5	0.46	0.50	0.51	0.58	0.46	0.58	0.58	0.56	0.46	0.58	0.55
	Smote		0.62	0.38	0.44	0.55	0.50	0.49	0.62	0.55	0.62	0.62	0.53	0.55	0.62	0.54
	Yok	AdaBoost	0.61	0.39	0.58	0.66	0.58	0.58	0.61	0.66	0.61	0.61	0.61	0.66	0.61	0.61
	Rastgele yeniden örnekleme (yerine koyma ile)		0.57	0.43	0.55	0.62	0.56	0.56	0.57	0.62	0.57	0.57	0.58	0.62	0.57	0.6
	Smote		0.64	0.36	0.62	0.70	0.63	0.62	0.64	0.70	0.64	0.64	0.65	0.70	0.64	0.66
	Yok	Karar Ağacı	0.6	0.4	0.55	0.55	0.55	0.56	0.6	0.55	0.6	0.6	0.6	0.55	0.6	0.59
	Rastgele yeniden örnekleme (yerine koyma ile)		0.6	0.4	0.58	0.59	0.58	0.58	0.6	0.59	0.6	0.6	0.61	0.59	0.60	0.62
	Smote		0.52	0.48	0.51	0.52	0.54	0.54	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.58
Cins OTU	Yok	RF	0.8	0.2	0.77	0.89	0.76	0.80	0.80	0.89	0.80	0.80	0.80	0.89	0.80	0.80
	Rastgele yeniden örnekleme (yerine koyma ile)		0.79	0.21	0.77	0.88	0.78	0.77	0.79	0.88	0.79	0.79	0.79	0.88	0.79	0.80
	Smote		0.84	0.16	0.82	0.88	0.83	0.82	0.84	0.88	0.84	0.84	0.84	0.88	0.84	0.84
	Yok	SVM	0.61	0.39	0.60	0.33	0.61	0.60	0.61	0.33	0.61	0.61	0.62	0.33	0.61	0.65
	Rastgele yeniden örnekleme (yerine koyma ile)		0.65	0.35	0.55	0.69	0.56	0.59	0.65	0.69	0.65	0.65	0.61	0.69	0.65	0.62
	Smote		0.63	0.37	0.58	0.69	0.63	0.55	0.63	0.69	0.63	0.63	0.58	0.69	0.63	0.58
	Yok	AdaBoost	0.8	0.2	0.78	0.87	0.78	0.79	0.8	0.87	0.8	0.8	0.8	0.87	0.8	0.8
	Rastgele yeniden örnekleme (yerine koyma ile)		0.76	0.24	0.75	0.84	0.76	0.75	0.76	0.84	0.76	0.76	0.77	0.84	0.76	0.78
	Smote		0.79	0.21	0.77	0.83	0.78	0.77	0.79	0.83	0.79	0.79	0.79	0.83	0.79	0.80
	Yok	Karar Ağacı	0.76	0.24	0.74	0.78	0.74	0.76	0.76	0.78	0.76	0.76	0.76	0.78	0.76	0.76
	Rastgele yeniden örnekleme (yerine koyma ile)		0.7	0.30	0.68	0.70	0.68	0.67	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
	Smote		0.79	0.21	0.77	0.77	0.78	0.77	0.79	0.77	0.79	0.79	0.79	0.77	0.79	0.80
Tur OTU	Yok	RF	0.84	0.16	0.81	0.89	0.80	0.84	0.84	0.89	0.84	0.84	0.83	0.89	0.84	0.84
	Rastgele yeniden örnekleme (yerine koyma ile)		0.78	0.22	0.76	0.87	0.77	0.76	0.78	0.87	0.78	0.78	0.78	0.87	0.78	0.79
	Smote		0.84	0.16	0.82	0.90	0.82	0.82	0.84	0.90	0.84	0.84	0.84	0.90	0.84	0.84
	Yok	SVM	0.69	0.31	0.68	0.75	0.71	0.69	0.69	0.75	0.69	0.69	0.69	0.75	0.69	0.73

Ek 1. (Devam)

	Rastgele yeniden örnekleme (yerine koyma ile)		0.65	0.35	0.56	0.72	0.57	0.59	0.65	0.72	0.65	0.65	0.62	0.72	0.65	0.62
	Smote		0.65	0.35	0.53	0.72	0.55	0.60	0.65	0.72	0.65	0.65	0.60	0.72	0.65	0.62
	Yok		0.85	0.15	0.84	0.88	0.84	0.83	0.85	0.88	0.85	0.85	0.85	0.88	0.85	0.85
	Rastgele yeniden örnekleme (yerine koyma ile)	AdaBoost	0.81	0.19	0.79	0.87	0.79	0.79	0.81	0.87	0.81	0.81	0.81	0.87	0.81	0.81
	Smote		0.80	0.20	0.78	0.84	0.77	0.79	0.80	0.84	0.80	0.80	0.80	0.84	0.80	0.80
	Yok		0.77	0.23	0.75	0.80	0.75	0.75	0.77	0.80	0.77	0.77	0.77	0.80	0.77	0.77
	Rastgele yeniden örnekleme (yerine koyma ile)	Karar Ağacı	0.78	0.22	0.77	0.79	0.77	0.76	0.78	0.79	0.78	0.78	0.79	0.79	0.78	0.79
	Smote		0.78	0.22	0.77	0.74	0.77	0.76	0.78	0.74	0.78	0.78	0.79	0.74	0.78	0.79
Tur+Cins seviye OTU	Yok	RF	0.80	0.20	0.77	0.85	0.76	0.79	0.80	0.85	0.80	0.80	0.79	0.85	0.80	0.79
	Rastgele yeniden örnekleme (yerine koyma ile)		0.85	0.15	0.84	0.89	0.85	0.83	0.85	0.89	0.85	0.85	0.85	0.89	0.85	0.86
	Smote		0.83	0.17	0.82	0.89	0.82	0.81	0.83	0.89	0.83	0.83	0.83	0.89	0.83	0.83
	Yok	SVM	0.68	0.32	0.52	0.72	0.56	0.70	0.68	0.72	0.68	0.68	0.60	0.72	0.68	0.69
	Rastgele yeniden örnekleme (yerine koyma ile)		0.67	0.33	0.56	0.72	0.57	0.62	0.67	0.72	0.67	0.67	0.62	0.72	0.67	0.64
	Smote		0.66	0.34	0.56	0.72	0.57	0.61	0.66	0.72	0.66	0.66	0.62	0.72	0.66	0.63
	Yok	AdaBoost	0.86	0.14	0.84	0.90	0.83	0.85	0.86	0.90	0.86	0.86	0.85	0.90	0.86	0.85
	Rastgele yeniden örnekleme (yerine koyma ile)		0.81	0.19	0.80	0.87	0.81	0.79	0.81	0.87	0.81	0.81	0.81	0.87	0.81	0.82
	Smote		0.78	0.22	0.76	0.84	0.76	0.76	0.78	0.84	0.78	0.78	0.78	0.84	0.78	0.78
	Yok	Karar Ağacı	0.77	0.23	0.74	0.77	0.74	0.75	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
	Rastgele yeniden örnekleme (yerine koyma ile)		0.78	0.22	0.77	0.79	0.78	0.76	0.78	0.79	0.78	0.78	0.78	0.79	0.78	0.79
	Smote		0.73	0.27	0.71	0.66	0.72	0.71	0.73	0.66	0.73	0.73	0.73	0.66	0.73	0.74
DNABERT-S (mamsimum)	Yok	RF	0.58	0.42	0.48	0.62	0.51	0.51	0.58	0.62	0.58	0.58	0.52	0.62	0.58	0.53
	Rastgele yeniden örnekleme (yerine koyma ile)		0.60	0.40	0.59	0.62	0.59	0.59	0.60	0.62	0.60	0.60	0.60	0.62	0.60	0.60
	Smote		0.59	0.41	0.57	0.62	0.57	0.57	0.59	0.62	0.59	0.59	0.59	0.62	0.59	0.59
	Yok	SVM	0.48	0.52	0.47	0.49	0.47	0.48	0.48	0.49	0.48	0.48	0.48	0.49	0.48	0.50
	Rastgele yeniden örnekleme (yerine koyma ile)		0.60	0.40	0.42	0.59	0.51	0.55	0.60	0.59	0.60	0.60	0.48	0.59	0.60	0.56
	Smote		0.58	0.42	0.47	0.55	0.51	0.51	0.58	0.55	0.58	0.58	0.52	0.55	0.58	0.53
	Yok	AdaBoost	0.61	0.39	0.50	0.65	0.54	0.59	0.61	0.65	0.61	0.61	0.55	0.65	0.61	0.59

Ek 1. (Devam)

	Rastgele yeniden örnekleme (yerine koyma ile)	Karar Ağacı	0.53	0.47	0.52	0.54	0.53	0.52	0.53	0.54	0.53	0.53	0.53	0.54	0.53	0.55
	Smote		0.52	0.48	0.51	0.53	0.51	0.51	0.52	0.53	0.52	0.52	0.53	0.53	0.52	0.53
	Yok		0.58	0.42	0.56	0.50	0.56	0.56	0.58	0.50	0.58	0.58	0.58	0.50	0.58	0.58
	Rastgele yeniden örnekleme (yerine koyma ile)		0.52	0.48	0.51	0.52	0.51	0.51	0.52	0.52	0.52	0.52	0.53	0.52	0.52	0.53
	Smote		0.54	0.46	0.53	0.54	0.53	0.53	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.55
DNABERT-S (ortalama)	Yok	RF	0.65	0.35	0.56	0.60	0.56	0.58	0.65	0.60	0.65	0.65	0.62	0.60	0.65	0.62
	Rastgele yeniden örnekleme (yerine koyma ile)		0.54	0.46	0.52	0.54	0.52	0.52	0.54	0.54	0.54	0.54	0.55	0.54	0.54	0.54
	Smote		0.59	0.41	0.56	0.56	0.56	0.56	0.59	0.56	0.59	0.59	0.60	0.56	0.59	0.61
	Yok	SVM	0.58	0.42	0.52	0.53	0.52	0.53	0.58	0.53	0.58	0.58	0.58	0.53	0.58	0.57
	Rastgele yeniden örnekleme (yerine koyma ile)		0.54	0.46	0.51	0.54	0.51	0.51	0.54	0.54	0.54	0.54	0.54	0.54	0.54	0.56
	Smote		0.52	0.48	0.49	0.53	0.52	0.50	0.52	0.53	0.52	0.52	0.53	0.53	0.52	0.55
	Yok	AdaBoost	0.63	0.37	0.54	0.62	0.54	0.55	0.63	0.62	0.63	0.63	0.60	0.62	0.63	0.60
	Rastgele yeniden örnekleme (yerine koyma ile)		0.5	0.5	0.48	0.53	0.49	0.49	0.50	0.53	0.50	0.50	0.51	0.53	0.50	0.54
	Smote		0.56	0.44	0.54	0.56	0.55	0.54	0.56	0.56	0.56	0.56	0.57	0.56	0.56	0.59
	Yok	Karar Ağacı	0.5	0.5	0.48	0.52	0.49	0.49	0.50	0.52	0.50	0.50	0.51	0.52	0.50	0.54
	Rastgele yeniden örnekleme (yerine koyma ile)		0.54	0.46	0.52	0.55	0.53	0.53	0.54	0.55	0.54	0.54	0.55	0.55	0.54	0.58
	Smote		0.48	0.52	0.46	0.49	0.47	0.48	0.48	0.49	0.48	0.48	0.49	0.49	0.48	0.53

Ek 2. Modelin Parametre Uzayının GridSearchCV ile Taranması Sonucu En İyi Modele Ait Hiper Parametre Uzayı

Temsil Vektörü	Yeniden Örneklemeye	Model	En İyi Sonuç Veren Parametre Uzayı	Eğitim Kümesi	Test Kümesi
4-mer	Yok	RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25}	243:Kontrol,369:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	SVM	{'C': 0.25, 'gamma': 0.25, 'kernel': 'poly', 'max_iter': 100}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'C': 0.25, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 50}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'C': 0.25, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 50}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	AdaBoost	{'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 25}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	Karar Ağacı	{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 5}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
5mer	Yok	RF	{'max_depth': 3, 'min_samples_leaf': 10, 'n_estimators': 10}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'max_depth': 5, 'min_samples_leaf': 5, 'n_estimators': 20}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	SVM	{'C': 0.25, 'gamma': 0.5, 'kernel': 'rbf', 'max_iter': 50}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'C': 0.25, 'gamma': 0.25, 'kernel': 'linear', 'max_iter': 50}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'C': 0.25, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 50}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	AdaBoost	{'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 50}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	Karar Ağacı	{'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 5}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
6mer	Yok	RF	{'max_depth': 4, 'min_samples_leaf': 10, 'n_estimators': 10}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 20}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	SVM	{'C': 0.25, 'gamma': 0.5, 'kernel': 'rbf', 'max_iter': 100}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örneklemeye (yerine koyma ile)		{'C': 0.25, 'gamma': 0.25, 'kernel': 'rbf', 'max_iter': 25}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'C': 0.25, 'gamma': 1, 'kernel': 'rbf', 'max_iter': 100}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	AdaBoost	{'algorithm': 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 50}	238:Kontrol,374:CH	54:Kontrol,99:CH

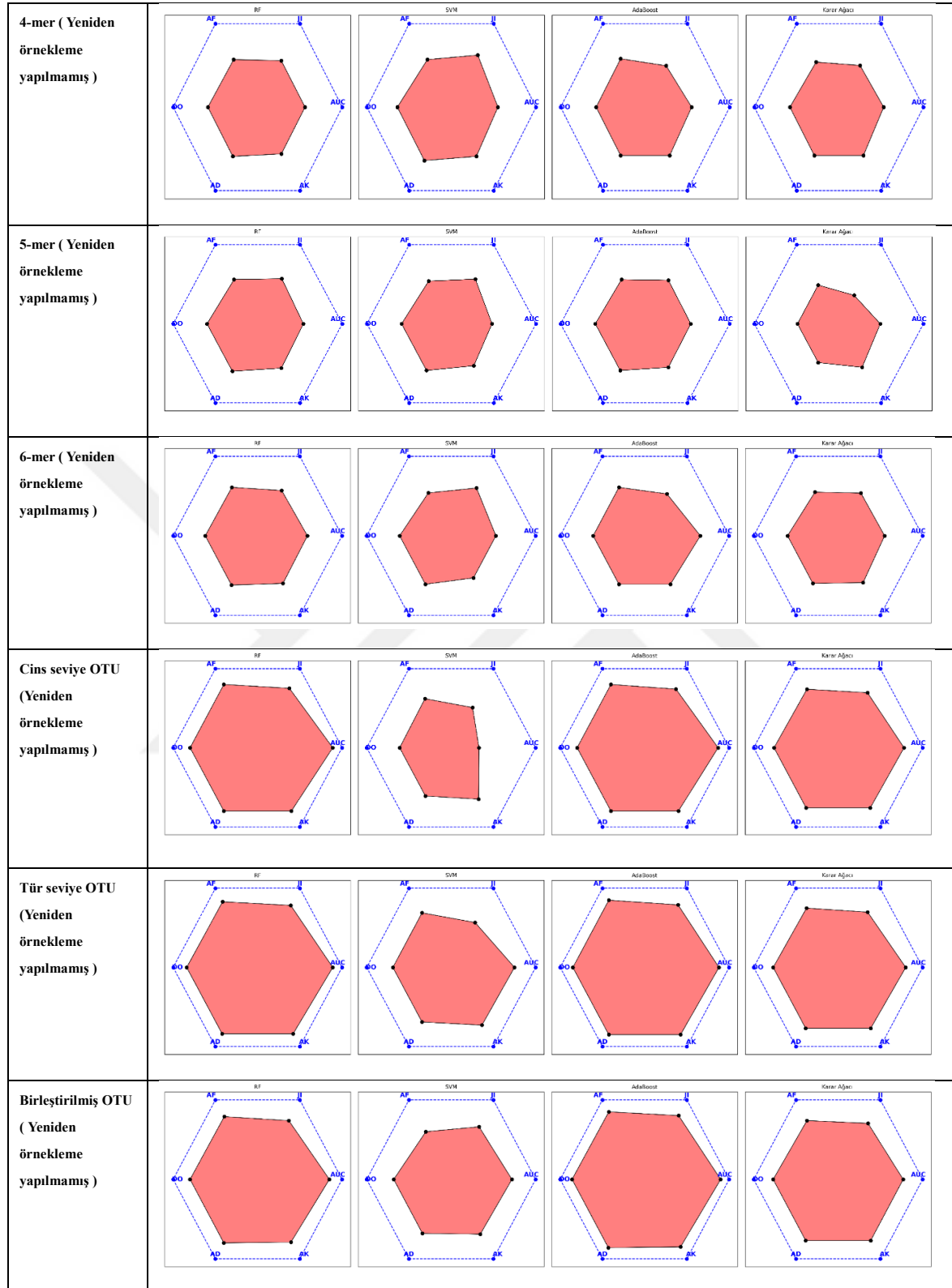
Ek 2. (Devam)

6mer	Rastgele yeniden örnekleme (yerine koyma ile)	Karar Ağacı	{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
Cins Seviye OTU	Yok	RF	{'max_depth': 5, 'min_samples_leaf': 5, 'n_estimators': 15}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 15}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 20}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		{'C': 1, 'gamma': 0.5, 'kernel': 'rbf', 'max_iter': 50}	238:Kontrol,374:CH	54:Kontrol,99:CH
Cins Seviye OTU	Rastgele yeniden örnekleme (yerine koyma ile)	SVM	{'C': 2, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 500}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'C': 2, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 300}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		algorithm: 'SAMME.R', 'learning_rate': 0.1, 'n_estimators': 50	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)	AdaBoost	algorithm: 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		"{'criterion': 'entropy', 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 5}"	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)	Karar Ağacı	"{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}"	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		{'max_depth': 5, 'min_samples_leaf': 5, 'n_estimators': 20}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)	RF	{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 20}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
Tür Seviye OTU	SMOTE		{'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		{'C': 16, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 200}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)	SVM	{'C': 16, 'gamma': 8, 'kernel': 'rbf', 'max_iter': 400}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'C': 2, 'gamma': 1, 'kernel': 'rbf', 'max_iter': 400}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		{'algorithm': 'SAMME', 'learning_rate': 1, 'n_estimators': 75}	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)	AdaBoost	{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75}	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok		"{'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 2, 'min_samples_split': 5}"	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)	Karar Ağacı	"{'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5}"	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 10}	1000:Kontrol,1000:CH	54:Kontrol,99:CH

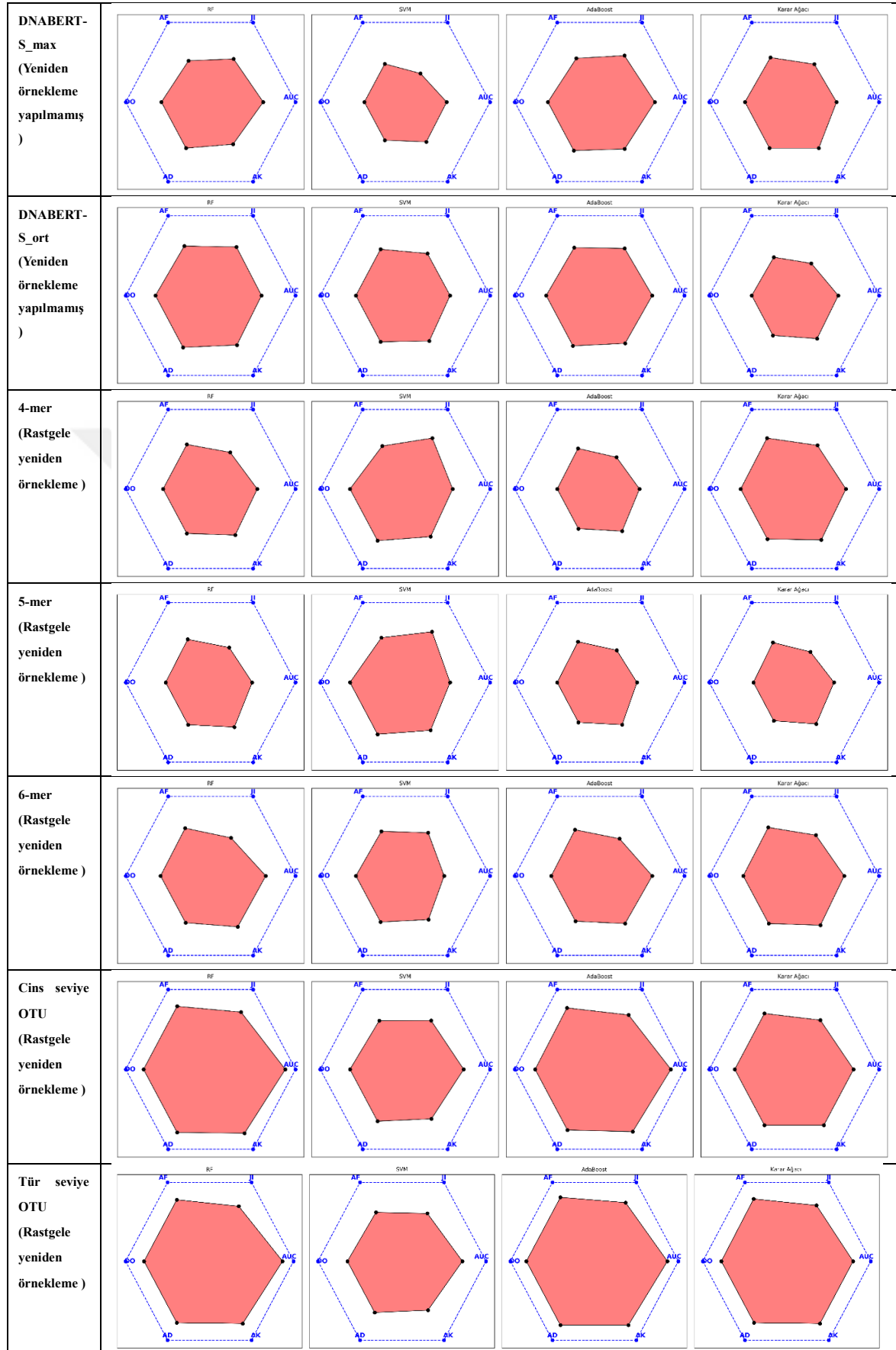
Ek 2. (Devam)

Tur seviye OTU+ Cins seviye OTU tablolarının birlikte kullanılması	Yok	RF	{ 'max_depth': 5, 'min_samples_leaf': 10, 'n_estimators': 10 }	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)		{ 'max_depth': 5, 'min_samples_leaf': 1, 'n_estimators': 25 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{ 'max_depth': 5, 'min_samples_leaf': 5, 'n_estimators': 25 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	SVM	{ 'C': 2, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 200 }	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)		{ 'C': 2, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 500 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{ 'C': 2, 'gamma': 2, 'kernel': 'rbf', 'max_iter': 300 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	AdaBoost	{ 'algorithm': 'SAMME.R', 'learning_rate': 0.5, 'n_estimators': 75 }	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)		{ 'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{ 'algorithm': 'SAMME.R', 'learning_rate': 1, 'n_estimators': 75 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	Yok	Karar Ağacı	{ 'criterion': 'entropy', 'max_depth': 3, 'min_samples_leaf': 6, 'min_samples_split': 5 }	238:Kontrol,374:CH	54:Kontrol,99:CH
	Rastgele yeniden örnekleme (yerine koyma ile)		{ 'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH
	SMOTE		{ 'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5 }	1000:Kontrol,1000:CH	54:Kontrol,99:CH

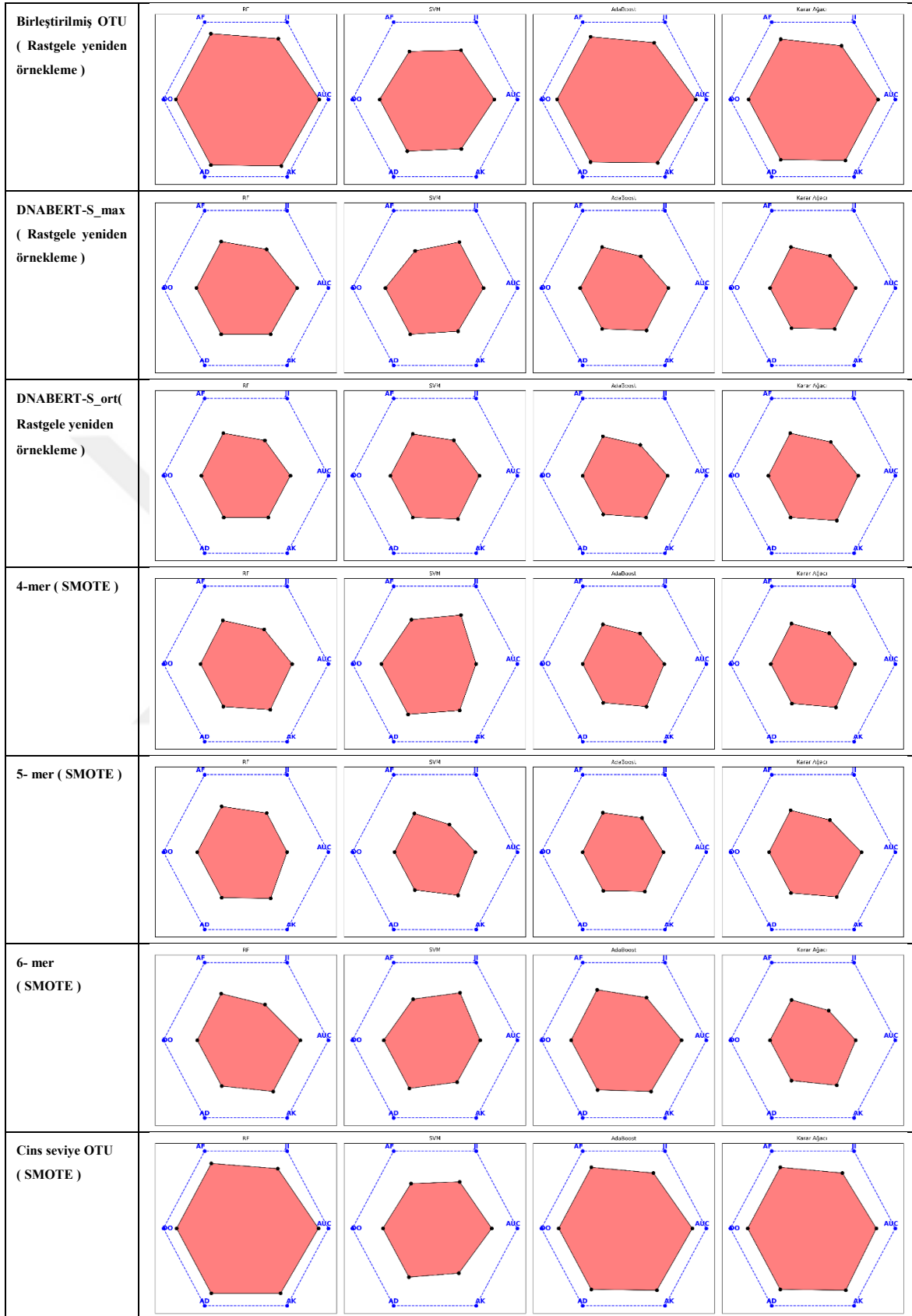
Ek 3. Test Verisi Üzerinde Poligon Alan Metriğine Ait Gösterim



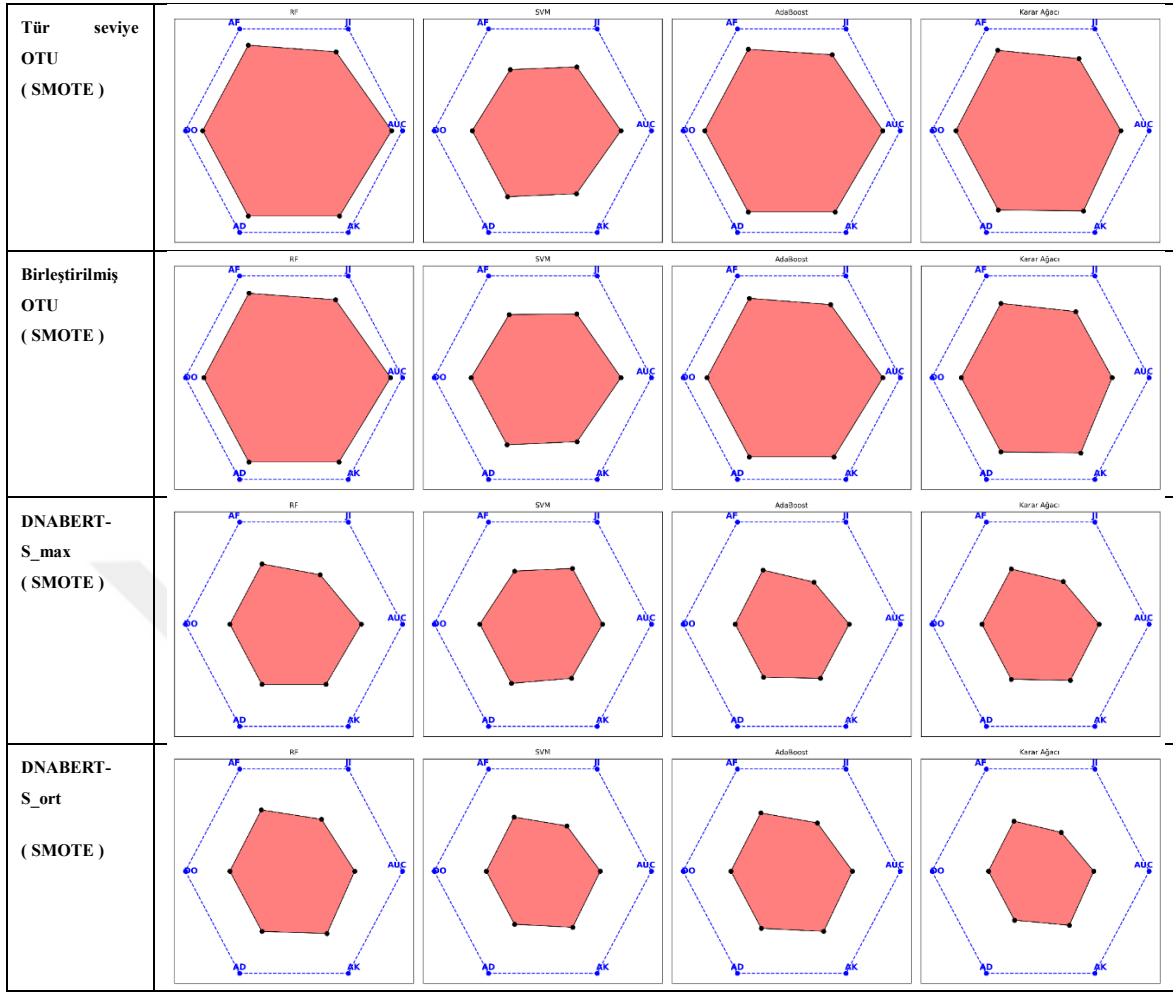
Ek 3. (Devam)



Ek 3. (Devam)



Ek 3. (Devam)



Ek 4. Ağırlıklandırılmış Test Verisi Sonuçları ve Ağırlıklandırılmış Metrikler Üzerinden Hesaplanan Poligon Alan Metrik Sonuçları

Model	Yeniden Örnekleme	Temsil Vektörü	Doğruluk (DO)	Ağırlıklandırılmış F1 Skor (AF)	Ağırlıklandırılmış AUC (AUC)	Ağırlıklandırılmış Duyarlılık (AD)	Ağırlıklandırılmış kesinlik (AK)	Jaccard İndeks (JI)	PAM
AdaBoost	Yok	Tur+Cins seviye OTU	0.86	0.85	0.9	0.86	0.85	0.80	0.73
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Tur+Cins seviye OTU	0.85	0.85	0.89	0.85	0.86	0.79	0.72
AdaBoost	Yok	Tur Seviye OTU	0.85	0.85	0.88	0.85	0.85	0.79	0.71
RF	SMOTE	Tur Seviye OTU	0.84	0.84	0.9	0.84	0.84	0.78	0.70
RF	Yok	Tur Seviye OTU	0.84	0.83	0.89	0.84	0.84	0.79	0.70
RF	SMOTE	Cins Seviye OTU	0.84	0.84	0.88	0.84	0.84	0.77	0.70
RF	SMOTE	Tur+Cins seviye OTU	0.83	0.83	0.89	0.83	0.83	0.77	0.69
AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Tur+Cins seviye OTU	0.81	0.81	0.87	0.81	0.82	0.73	0.65
AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Tur Seviye OTU	0.81	0.81	0.87	0.81	0.81	0.74	0.65
RF	Yok	Cins Seviye OTU	0.8	0.8	0.89	0.8	0.8	0.75	0.65
AdaBoost	Yok	Cins Seviye OTU	0.8	0.8	0.87	0.8	0.8	0.74	0.64
AdaBoost	SMOTE	Tur Seviye OTU	0.8	0.8	0.84	0.8	0.8	0.75	0.64
RF	Yok	Tur+Cins seviye OTU	0.8	0.79	0.85	0.8	0.79	0.74	0.63
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Cins Seviye OTU	0.79	0.79	0.88	0.79	0.8	0.72	0.63
AdaBoost	SMOTE	Cins Seviye OTU	0.79	0.79	0.83	0.79	0.8	0.72	0.62
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Tur Seviye OTU	0.78	0.78	0.87	0.78	0.79	0.70	0.61
AdaBoost	SMOTE	Tur+Cins seviye OTU	0.78	0.78	0.84	0.78	0.78	0.72	0.61
.Karar Ağacı	SMOTE	Cins Seviye OTU	0.79	0.79	0.77	0.79	0.8	0.72	0.60
.Karar Ağacı	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Tur Seviye OTU	0.78	0.79	0.79	0.78	0.79	0.71	0.60
.Karar Ağacı	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	Tur+Cins seviye OTU	0.78	0.78	0.79	0.78	0.79	0.69	0.59

Ek 4. (Devam)

.Karar Ağacı	SMOTE	Tur Seviye OTU	0.78	0.79	0.74	0.78	0.79	0.71	0.58
AdaBoost	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	Cins Seviye OTU	0.76	0.77	0.84	0.76	0.78	0.68	0.58
.Karar Ağacı	Yok	Tur Seviye OTU	0.77	0.75	0.8	0.77	0.77	0.70	0.58
.Karar Ağacı	Yok	Tur+Cins seviye OTU	0.77	0.74	0.77	0.77	0.77	0.71	0.57
.Karar Ağacı	Yok	Cins Seviye OTU	0.76	0.74	0.78	0.76	0.76	0.70	0.56
.Karar Ağacı	SMOTE	Tur+Cins seviye OTU	0.73	0.73	0.66	0.73	0.74	0.65	0.50
.Karar Ağacı	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	Cins Seviye OTU	0.7	0.7	0.7	0.7	0.7	0.62	0.47
SVM	Yok	Tur Seviye OTU	0.69	0.69	0.75	0.69	0.73	0.57	0.47
SVM	Yok	Tur+Cins seviye OTU	0.68	0.6	0.72	0.68	0.69	0.66	0.45
SVM	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	Tur+Cins seviye OTU	0.67	0.62	0.72	0.67	0.64	0.64	0.43
SVM	SMOTE	Tur+Cins seviye OTU	0.66	0.62	0.72	0.66	0.63	0.63	0.43
SVM	SMOTE	Tur Seviye OTU	0.65	0.6	0.72	0.65	0.62	0.63	0.41
SVM	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	Tur Seviye OTU	0.65	0.62	0.72	0.65	0.62	0.61	0.41
AdaBoost	SMOTE	6mer	0.64	0.65	0.7	0.64	0.66	0.55	0.41
SVM	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	Cins Seviye OTU	0.65	0.61	0.69	0.65	0.62	0.61	0.41
RF	Yok	DNABERT-S ort	0.65	0.62	0.6	0.65	0.62	0.61	0.39
SVM	SMOTE	Cins Seviye OTU	0.63	0.58	0.69	0.63	0.58	0.60	0.38
.Karar Ağacı	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	4mer	0.63	0.64	0.61	0.63	0.64	0.55	0.38
AdaBoost	Yok	DNABERT-S ort	0.63	0.6	0.62	0.63	0.6	0.59	0.37
SVM	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	4mer	0.65	0.54	0.56	0.65	0.6	0.64	0.37
AdaBoost	Yok	6mer	0.61	0.61	0.66	0.61	0.61	0.53	0.36
SVM	Rastgele yeniden örneklemeye (yeniden yerine koyma ile)	5mer	0.65	0.56	0.53	0.65	0.6	0.64	0.36

Ek 4. (Devam)

SVM	Yok	4mer	0.64	0.57	0.55	0.64	0.59	0.62	0.36
RF	Yok	6mer	0.62	0.61	0.59	0.62	0.6	0.57	0.36
SVM	SMOTE	4mer	0.65	0.57	0.5	0.65	0.6	0.63	0.36
AdaBoost	Yok	DNABERT-S_max	0.61	0.55	0.65	0.61	0.59	0.59	0.36
RF	SMOTE	6mer	0.59	0.6	0.66	0.59	0.66	0.46	0.35
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	6mer	0.59	0.6	0.65	0.59	0.64	0.48	0.35
.Karar Ağacı	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	6mer	0.6	0.61	0.59	0.6	0.62	0.51	0.35
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_max	0.6	0.6	0.62	0.6	0.6	0.5	0.34
SVM	SMOTE	6mer	0.62	0.53	0.55	0.62	0.54	0.61	0.33
RF	SMOTE	DNABERT-S_max	0.59	0.59	0.62	0.59	0.59	0.48	0.33
RF	SMOTE	DNABERT-S_ort	0.59	0.6	0.56	0.59	0.61	0.51	0.33
RF	Yok	5mer	0.6	0.56	0.54	0.6	0.56	0.57	0.33
.Karar Ağacı	Yok	6mer	0.6	0.55	0.55	0.6	0.59	0.54	0.33
RF	Yok	4mer	0.59	0.57	0.56	0.59	0.56	0.56	0
SVM	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_max	0.6	0.48	0.59	0.6	0.56	0.59	0.32
SVM	Yok	6mer	0.61	0.54	0.53	0.61	0.53	0.60	0.32
AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	6mer	0.57	0.58	0.62	0.57	0.6	0.47	0.32
AdaBoost	Yok	5mer	0.59	0.56	0.54	0.59	0.55	0.55	0.32
RF	SMOTE	5mer	0.59	0.59	0.5	0.59	0.6	0.50	0.32
RF	Yok	DNABERT-S_max	0.58	0.52	0.62	0.58	0.53	0.54	0.32
SVM	Yok	DNABERT-S_ort	0.58	0.58	0.53	0.58	0.57	0.53	0.31
AdaBoost	Yok	4mer	0.58	0.58	0.55	0.58	0.58	0.50	0.31
SVM	Yok	Cins Seviye OTU	0.61	0.62	0.33	0.61	0.65	0.51	0.31
.Karar Ağacı	Yok	4mer	0.58	0.54	0.53	0.58	0.58	0.5	0.30
AdaBoost	SMOTE	DNABERT-S_ort	0.56	0.57	0.56	0.56	0.59	0.47	0.30
SVM	SMOTE	DNABERT-S_max	0.58	0.52	0.55	0.58	0.53	0.55	0.30
SVM	Yok	5mer	0.59	0.54	0.48	0.59	0.53	0.57	0.30
.Karar Ağacı	Yok	DNABERT-S_max	0.58	0.56	0.5	0.58	0.58	0.48	0.30

Ek 4. (Devam)

SVM	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	6mer	0.58	0.56	0.46	0.58	0.55	0.54	0.30
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	4mer	0.56	0.56	0.55	0.56	0.58	0.46	0.30
RF	SMOTE	4mer	0.55	0.56	0.56	0.55	0.59	0.44	0.29
.Karar Ağacı	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_ort	0.54	0.55	0.55	0.54	0.58	0.44	0.28
.Karar Ağacı	SMOTE	5mer	0.53	0.54	0.59	0.53	0.58	0.42	0.28
SVM	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_ort	0.54	0.54	0.54	0.54	0.56	0.46	0.28
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_ort	0.54	0.55	0.54	0.54	0.54	0.46	0.28
.Karar Ağacı	SMOTE	DNABERT-S_max	0.54	0.54	0.54	0.54	0.55	0.42	0.27
SVM	SMOTE	DNABERT-S_ort	0.52	0.53	0.53	0.52	0.55	0.44	0.27
RF	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	5mer	0.53	0.54	0.49	0.53	0.56	0.44	0.26
AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_max	0.53	0.53	0.54	0.53	0.55	0.41	0.26
.Karar Ağacı	SMOTE	6mer	0.52	0.52	0.52	0.52	0.58	0.38	0.26
AdaBoost	SMOTE	DNABERT-S_max	0.52	0.53	0.53	0.52	0.53	0.41	0.26
.Karar Ağacı	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_max	0.52	0.53	0.52	0.52	0.53	0.42	0.26
.Karar Ağacı	SMOTE	4mer	0.51	0.52	0.51	0.51	0.56	0.40	0.25
AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	DNABERT-S_ort	0.5	0.51	0.53	0.5	0.54	0.40	0.25
.Karar Ağacı	Yok	DNABERT-S_ort	0.5	0.48	0.52	0.5	0.54	0.40	0.24
AdaBoost	SMOTE	5mer	0.5	0.51	0.48	0.5	0.51	0.44	0.24
AdaBoost	SMOTE	4mer	0.5	0.51	0.49	0.5	0.55	0.39	0.24

Ek 4. (Devam)

AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	4mer	0.5	0.51	0.47	0.5	0.53	0.40	0.23
AdaBoost	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	5mer	0.5	0.51	0.44	0.5	0.53	0.40	0.23
SVM	SMOTE	5mer	0.49	0.5	0.49	0.49	0.56	0.36	0.23
.Karar Ağacı	Yok	5mer	0.49	0.49	0.49	0.49	0.55	0.36	0.23
.Karar Ağacı	SMOTE	DNABERT-S_ort	0.48	0.49	0.49	0.48	0.53	0.38	0.23
.Karar Ağacı	Rastgele yeniden örnekleme (yeniden yerine koyma ile)	5mer	0.48	0.5	0.47	0.48	0.52	0.38	0.22
SVM	Yok	DNABERT-S_max	0.48	0.48	0.49	0.48	0.5	0.36	0.22

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Sinem ÖZDEMİR

Uyruğu : T.C.

Doğum Tarihi :

Telefon (İş) :

E-Posta :

Yazışma Adresi (İş) :

Bilişimi Anabilim Dalı

EĞİTİM BİLGİLERİ

Derece	Mezun Olduğu Kurumun Adı	Mezuniyet Yılı
Doktora/Uzmanlık	KTÜ Biyoistatistik ve Tıp Bilişimi	-
Yüksek Lisans	KTÜ Bilgisayar Mühendisliği	2018
Lisans	YTÜ Matematik Mühendisliği	2014
Lise	Cengizhan Anadolu Lisesi	2009

AKADEMİK/MESLEKİ DENEYİMİ

Görevi	Kurum	Süre (Yıl -Yıl)
1 Araştırma Görevlisi	KTÜ	2020-2025

YABANCI DİL

İngilizce

UZMANLIK ALANI

Biyoenformatik

YAYINLAR

- 1 Kurt B, Gürlek B, Keskin S, Özdemir S, Karadeniz Ö, Kırkbir İB, Kurt T, Ünsal S, . Kart C, Baki N, Turhan K. Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. Med Biol Eng Comput. 2023 Jul;61(7):1649-1660. doi: 10.1007/s11517-023-02800-7. Epub 2023 Feb 27. PMID: 36848010; PMCID: PMC9969040.
- 2 Kurt B, Gürlek B, Keskin S, Kurt T, Kasap ZA, Kırkbir İB, Kart C, Özdemir S, Baki . N, Ünsal S, Turhan K. A prospective study for gestational diabetes mellitus: Analysis of risk factors in Turkish women for early prediction. Int J Gynaecol Obstet. 2023 May;161(2):525-535. doi: 10.1002/ijgo.14531. Epub 2022 Nov 11. PMID: 36306416.

BİLDİRİLER

1. Özdemir, S., & Dizdaroğlu, B., (2015). PDEs-Based Gaussian Noise Removal from Color Images . 9th International Conference on Electrical and Electronics Engineering (ELECO) (pp.246-250). Bursa, Turkey
2. Ünsal, S., Özdemir, S., Özdiñ, I., Bayraklı, A., Albayrak, M., Turhan, K., ... Doğan, T.(2022). A Protein Representation Model for Low-Data . 15th International Symposium on Health Informatics and Bioinformatics (pp.105-106). Mersin, Turkey
3. Turhan, K., Ünsal, S., Cüvitoğlu, A., Özdemir, S., Ünsal, Ü., Kasap, B., ... Karaca, C.(2022). Machine Learning-Based New Gene Discovery and Drug Repositioning Platform for Improving Immunotherapy Efficacy in Cancer . International Molecular Immunology & Immunogenetics Congress V (pp.48). İzmir, Turkey.
4. Turhan, K., Özdemir, S., Zihni, N. B., Kalaycı, M. E., & Ünsal, S., (2023). A Study on Immunoreceptor Tyrosine-Based Motifs . 14. Tıp Bilişimi Kongresi (pp.124-127), İzmir, Turkey.
5. Siyah, B., Kalaycı, M. E., Özdemir, S., Berber, T., & Turhan, K. (2024). Integration of non-negative matrix factorization to identification of immune system checkpoints. 15. Tıp Bilişimi Kongresi (pp.158-171), Trabzon, Turkey.
6. Özdemir, S., & Turhan, K. (2024). Utilizing pretrained GPT-2 model for analyzing drug reviews. 15. Tıp Bilişimi Kongresi (p. 306), Trabzon, Turkey.Integration of Non-Negative Matrix Factorization to Identification of Immune System Checkpoints (pp.158-171). Trabzon, Turkey

ÖDÜLLER / TEŞVİKLER/ BURSLAR

1. 100/2000 YÖK Doktora Bursu, 2018-2022
2. (Bursiyer) Gestasyonel Diyabet Tanısında Klinik Karar Destek Modeli, TÜBİTAK Projesi, 2018-2021
3. (Bursiyer) Kanserde İmmünoterapi Etkinliğini Artırmak İçin Makine Öğrenmesine Dayalı Gen Keşfi ve İlaç Yeniden Konumlandırma Platformunun Geliştirilmesi, TÜBİTAK Projesi,2020-2023

HOBİLER

1. Müzik dinlemek
2. Yeni bilgiler edinmek