Metagenomik Verilerin Meta-Analiziyle Kolerektal Kanserin Popülasyona Özgü Sınıflandırılması

Population Specific Classification of Colorectal Cancer with Meta-Analysis of Metagenomic Data

Mustafa Temiz
Department of Electrical and Computer Engineering
Abdullah Gul University
Kayseri, Türkiye
mustafa.temiz@agu.edu.tr

Malik Yousef
Department of Information System
Zefat Academic College
Zefat, Israel
malik.yousef@gmail.com

Burcu Bakir-Gungor
Department of Computer Engineering
Abdullah Gul University
Kayseri, Türkiye
burcu.gungor@agu.edu.tr

2023 Innovations in Intelligent Systems and Applications Conference (ASYU) | 979-8-3503-0659-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASYU58738.2023.10296760

***Özetçe*— Yeni nesil dizilemedeki ve "-omik" teknolojilerdeki gelişmeler, insan bağırsak mikrobiyomunu karakterize etmeyi mümkün kılmaktadır. Bu mikroorganizmaların bazıları bağışıklık sistemimizin temel düzenleyicileriyken, mikrobiyotanın modülasyonu çeşitli hastalıklara yol açar. Dünya çapında üçüncü yaygın kanser türü olan kolorektal kanser (KRK), genetik mutasyonlar, çevresel koşullar ve bağırsak mikrobiyotasındaki anomalilerin etkisiyle oluşmaktadır. Bu çalışma, tür seviyesinde metagenomik veri setleri üzerinde çeşitli makine öğrenmesi yöntemleri kullanarak farklı popülasyonlar için meta-analiz gerçekleştirmeyi; bu sayede KRK teşhisine yardımcı olabilecek sınıflandırma modelleri oluşturmayı amaçlamaktadır. Bu çalışmada, 8 farklı ülke ve 9 farklı metagenomik veri seti üzerinde popülasyon içi, popülasyonlar arası ve leave one dataset out (LODO) yöntemi kullanılarak 3 farklı meta-analiz gerçekleştirilmiştir. KRK teşhisine yardımcı model geliştirirken 4 farklı sınıflandırma algoritması (Rastgele Orman (RF), Logitboost, Adaboost ve Karar Agaci (DT)) kullanılmaktadır. Yapılan deneylerde en üstün performans olarak, popülasyonlar arası performans değerlendirmesinde eğitim veri seti için JP ve test veri seti için JPN popülasyonları kullanıldığında Random Forest algoritması ile 0.98 AUC elde etmiştir.**

Anahtar Kelimeler—Kolorektal kanser; metagenomik; sınıflandırma; meta analiz; bağırsak mikrobiyotası

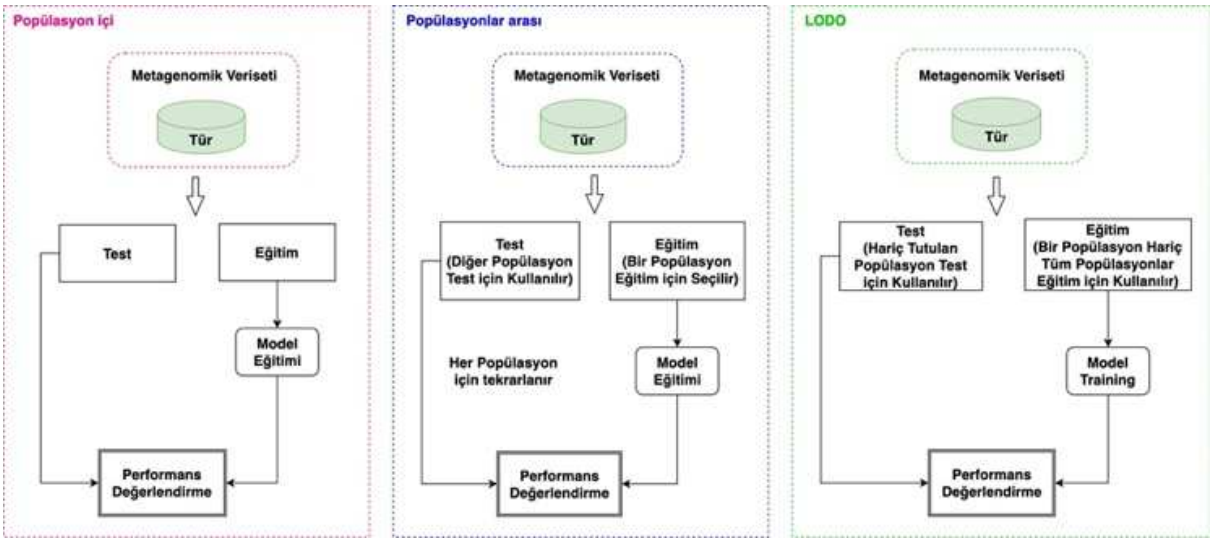
***Abstract*— Advances in next-generation sequencing and "-omics" technologies makes it possible to characterize the human gut microbiome. While some of these microorganisms are important regulators of our immune system, modulation of the microbiota leads to a variety of diseases. Colorectal cancer (CRC), the third most common cancer worldwide, is caused by genetic mutations, environmental conditions, and abnormalities in the gut microbiota. Using various machine learning methods and meta-analysis techniques, this study aims to build a classification model that can help in CRC diagnosis by analyzing metagenomic datasets of different populations obtained at the species level. Using 8 different countries and 9 different metagenomic datasets, 3 different meta-analyzes are performed: within-population, cross-population, and one population is selected for testing and the rest is used as a training dataset (LODO). For CRC classification, 4 different classification algorithms (Random Forest (RF), Logitboost, Adaboost, and Decision Tree (DT)) are used. The best performance among these methods was obtained with the Random Forest algorithm with an AUC of 0.98 by using JP for the training data set and JPN populations for the test data set in the cross-population performance evaluation.**

Keywords—Colorectal cancer, metagenomic, classification, meta-analysis, gut microbiota

I. GİRİŞ

Kanser, yüksek ölüm oranıyla dünya çapında önde gelen ölüm nedenlerinden birisidir. Nüfusun yaşlanması ve çevresel faktörler nedeniyle kanser nedenli ölüm oranları her geçen gün artmaktadır [1]. Kolorektal kanser (KRK) en yaygın kanser türlerinden biridir ve gelişimi bağırsak mikrobiyotasındaki değişikliklerle ilişkilendirilmektedir [2]. Dünya çapında yılda yaklaşık 900.000 ölümden sorumlu olan KRK, sağlık hizmetleri için büyük bir küresel zorluk oluşturmaktadır [3]. KRK'nın oluşumunda genetik ve çevresel faktörler önemli olmakla birlikte, son yıllardaki çalışmalar, yerleşik mikrobiyal popülasyon ve konakçı bağışıklık tepkileri arasındaki etkileşimlerin bozulmasının

979-8-3503-0659-0/23/$31.00 ©2023 IEEE



Şekil 1 Önerilen Yöntem

KRK'nın en önemli karakteristiği olduğunu bildirmiştir [4].

Dizileme teknolojisi ve biyoinformatik alanındaki gelişmeler sayesinde, daha fazla genomik, epigenomik, transkriptomik, proteomik, radyomik ve metagenomik faktör, KRK karsinogenezinde önemli oyuncular olarak ortaya çıkmaktadır [5]. Multi-omik profilleme ve entegre yaklaşımlar, KRK ile ilişkili mikrobiyal taksonların, metabolitlerin, DNA metilasyonu ile ilişkili gen ekspresyonu seviyesi değişimlerinin bulunabileceği göstermektedir [6]. İnsan bağırsak mikrobiyomunun (gastrointestinal sistemimizdeki mikrobiyal topluluğun kolektif genomlarının) insan –omlarıyla dinamik bir ilişki halinde olması nedeniyle, KRK hastalarında insan bağırsak mikrobiyomu analizleri önem arz etmektedir ve aktif bir araştırma alanıdır [7]. Metagenomik çalışmaların karmaşıklığı nedeniyle, makine öğrenme tekniklerinin bu alandaki uygulamaları çok çeşitli soruları ele almak için popüler hale gelmiştir. Bu bağlamda, mikrobiyom ile hastalık durumları arasında ilişki kurarak, hastalıkların belirlenebileceği belirtilmektedir [8]. Günümüzde metagenomik okuma dizilerinden hastalık tahmini problemi için temel olarak üç tip öznitelik üretilmektedir: i) farklı mikroorganizmaların bolluk değerleri (abundance values), ii) metagenomik örneklerin fonksiyonel anotasyonu, iii) ham okumalardan elde edilen k-mer bolluk değerleri. Vakalar ve kontroller arasında mikrobiyom bileşiminin farklı olması nedeniyle, mikrobiyal bolluk profilleri hastalık tahmininde yaygın olarak kullanılmaktadır. Bu çalışmada ham veriler, mikrobiyal bolluk profillerine dönüştürülerek analizler gerçekleştirilmiştir.

Makine öğrenimi (ML) algoritmaları kullanılarak KRK'nın otomatik tespiti birçok araştırmacının ilgisini çekmektedir ve geleneksel makine öğrenimi algoritmaları mikrobiyom verilerinin sınıflandırılmasında yaygın olarak kullanılmaktadır [9]–[13]. KRK sınıflandırması için çeşitli makine öğrenmesi yöntemlerinin kullanıldığı bu çalışmalar, KRK tanı, teşhis ve tedavisinde umut vadetmektedir. Bu çalışmada meta-analiz yöntemleri kullanılarak, farklı popülasyonlar özelinde daha ayrıntılı KRK sınıflandırması gerçekleştirilmiştir. KRK ile ilişkili metagenomik veriler, yapay zekâ yöntemleri ile analiz edilerek KRK teşhisine yardımcı olacak genel bir sınıflandırma modeli oluşturulmuş; popülasyonlara özgü meta-analizler ile 8 farklı popülasyon üzerinde KRK teşhisine yardımcı olacak sınıflandırma modelleri geliştirilmiştir.

Çalışmanın geri kalanında, kullandığımız veri seti ve metodolojimiz, "Materyal ve Yöntem" başlıklı bölümde özetlenmektedir. KRK ile ilişkili metagenomik verilere uygulanan sınıflandırma yöntemleri ve uygulama sonuçları "Uygulama" bölümünde açıklanmaktadır. KRK teşhisi için oluşturduğumuz sınıflandırma model değerlendirmeleri ve ülkelere özgü meta analiz sonuçlarının değerlendirmesi ise "Sonuç" bölümünde ele alınmaktadır.

II. MATERYAL VE YÖNTEM

A. Veriseti

Beghini ve ark. 8 farklı ülkeyi kapsayan 9 farklı veri setinden toplam 1262 metagenomik örneği (600 kontrol ve 662 KRK) bir araya getirmişlerdir [14]. Bu çalışmada her bir örneğin ham mikrobiyom DNA'sı ilgili proje sayfasından indirilmiş taksonomik profilleme için MetaPhlAn3 ve HUMAN3 kullanılmıştır. Thomas ve ark.'nın çalışmasında [11] açıklanan yöntem izlenerek, her bir veri setindeki örnekler için, tür seviyesinde nispi bolluk miktarları belirlenmektedir. Bu çalışmada, yukarıda bahsedilen 8 farklı ülkede toplanmış olan KRK ile ilişkili metagenomik veri setindeki 1262 örneğe ait 917 farklı tür için nispi bolluk değerleri kullanılmaktadır. Meta-analizler için kullanılan popülasyon isimleri ve örnek sayıları Tablo 1. de gösterilmektedir.

Populasyon	Ornek sayisi	Saglikli ornek sayisi	Hasta ornek sayisi
AUT	107	46	61
CHN	128	75	53
DEU	125	60	65
FRA	114	53	61
IND	60	30	30

ITA	106	57	49
JP	80	40	40
JPN	438	187	251
USA	104	52	52

Tablo I. Çalışmada kullanılan veri setine ait bilgiler

B. Yöntem

Bu çalışmada, KRK örneklerini kontrollerden ayırt etmek için, farklı sınıflandırma algoritmaları kullanarak, bir dizi makine öğrenimi modelleri oluşturulmaktadır. İlk olarak tüm ülke verilerini içeren 1262 örnek ve 917 öznitelikten oluşan veri setine makine öğrenmesi performanslarını değerlendirmek amacıyla deneyler gerçekleştirdik. Deneylerimizde' Rastgele Orman (RF), Karar Agaci (DT), Adaboost ve Logitboost sınıflandırma algoritmaları kullanıldı. Ardından popülasyonlar özelinde gerçekleştirilen meta-analizler ile KRK sınıflandırma başarısı incelendi. 3 farklı meta-analiz gerçekleştirildi. Önerilen yöntem Şekil 1'de özetlenmektedir ve yöntemin ayrıntıları aşağıda sunulmuştur.

1) Popülasyon içi KRK analizi:

Bu yöntem her bir popülasyonun verisini ayrı ayrı değerlendirmektedir. Her bir popülasyon verisi eğitim için %90 ve test için %10 oranında seçilerek model oluşturulmakta ve yine aynı popülasyon verisi üzerinde test edilmektedir. Deneylerde 10 kat MCCV kullanılmaktadır. Deney 10 kez tekrarlanarak ortalama AUC değerleri ve standart sapma değerleri hesaplanmaktadır.

2) Popülasyonlar arası KRK analizi:

Bu yöntemde model, belli bir popülasyon verisi üzerinde eğitilmekte, geliştirilen model eğitimde kullanılmayan her bir farklı popülasyonun verisi üzerinde ayrı ayrı test edilmektedir. Bu deney her bir popülasyon için tekrarlanmaktadır. Sırasıyla tüm veri kümeleri model eğitimi için kullanılmakta, her seferinde geriye kalan veri kümeleri ayrı ayrı test verisi olarak kullanılmaktadır.

3) Bir popülasyon verisinin test için dışarda bırakılması ile KRK analizi:

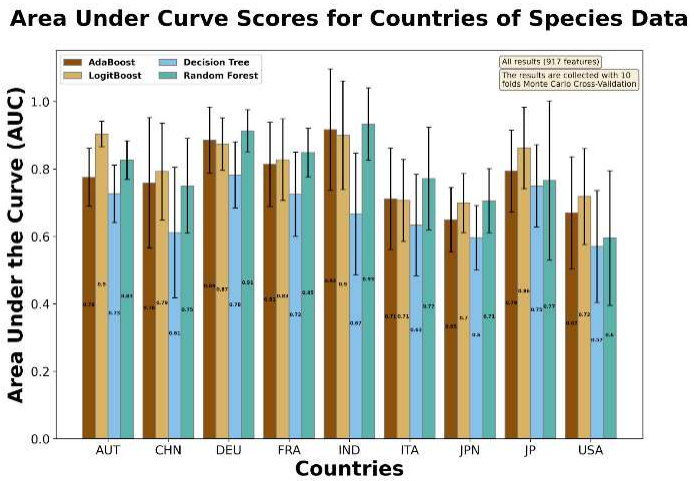
Bu yöntemde belli bir popülasyon verisi test kümesi olarak ayrılırken, test için seçilen bu popülasyon verisi dışında kalan tüm popülasyonların verisi birleştirilerek model eğitiminde kullanılmaktadır. Bu deney her bir popülasyon için tekrarlanmaktadır. Sırasıyla her bir popülasyonun veri kümesi test için kullanılmakta, her seferinde geriye kalan veri kümeleri birleştirilerek eğitim verisi olarak kullanılmaktadır.

III. UYGULAMA

Tüm ülke verileri kullanılarak gerçekleştirilen analizlerde eğitim için %90 ve test için %10 şeklinde ayırarak sınıflandırma performansı değerlendirilmektedir. Ayrıca model doğruluğunu desteklemek için 10 kat MCCV tekniği uygulanarak ortalama performans metrikleri hesaplanmaktadır. Modellerimizi KNIME platformu [15] kullanarak geliştirmekte; H20 ve scikit-learn kütüphanelerinden [16] faydalanmaktayız. Modellerin tahmin performansları doğruluk, F1-Skoru, AUC (Area Under the Receiver Operating Characteristic Curve) ölçüleri kullanılarak değerlendirilmektedir.

Popülasyonlara ait meta-analiz değerlendirme sonuçları aşağıda ayrıntılı bir şekilde paylaşılmaktadır. Öncelikle her bir yöntemin nasıl uygulandığı açıklanmakta ve elde edilen sonuçlar değerlendirilmektedir. Bu sonuçlar KRK tanı ve teşhisinde popülasyon özelinde karar verme aşamasında bilimsel çalışmalara katkı sunmaktadır.

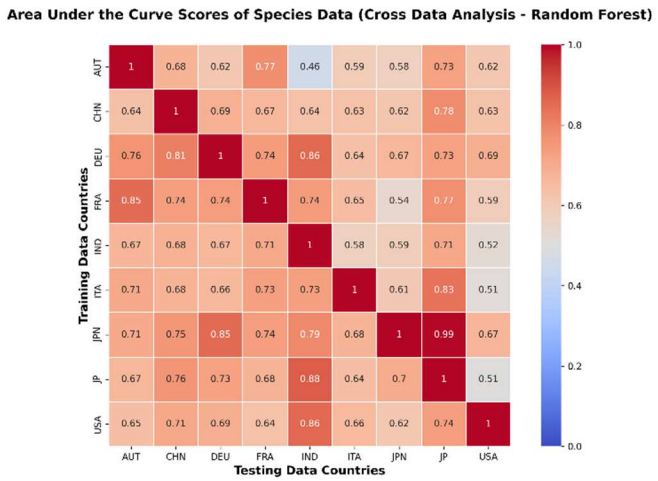
Popülasyon içi KRK analizi: Tür seviyesinde gerçekleştirilen analizlerde, popülasyon içi değerlendirmelerde, %90 örnek eğitim için, %10 örnek test için ayrılmaktadır. 10-kat MCCV uygulandığında ortalama AUC metrikleri ve standart sapma değerleri Şekil 2'de gösterilmektedir. Şekil 2 incelendiğinde tür seviyesinde tüm özniteliklerin (917 tür için nisbi bolluk miktarları) kullanılması ile gerçekleştirilen analizlerde en iyi performans Rastgele Orman algoritması kullanıldığında Hint (IND) popülasyonu için elde edilmiştir (%93 AUC skoru). Bir diğer başarılı sonuç' yine Hint popülasyonu için Adaboost sınıflandırma algoritması ile elde edilmektedir (%92 AUC skoru). Tüm popülasyonlara ait sonuçlar ayrı ayrı değerlendirildiğinde, Rastgele Orman algoritması diğer sınıflandırma algoritmalarından üstün performans göstermektedir. 9 farklı veri seti içinden 5 tanesinde Rastgele Orman algoritması ile diğer algoritmalardan daha yüksek sonuçlar elde edilmektedir. Bu durum, popülasyon içinde, mikrobiyotadaki nisbi bolluk değerlerinden KRK sınıflandırması için Rastgele Orman algoritması ile daha etkili sınıflandırma yapılacağı anlamına gelmektedir. Rastgele Orman algoritmasından sonra ikinci en etkili sonuçlar Logitboost algoritması ile elde edilmiştir. 9 farklı veri seti içerisinden kalan 4 tanesinde Logitboost algoritması üstün performans göstermektedir. Logitboost algoritması ile en yüksek sonuç Avusturya (AUT) popülasyonu için elde edilmiştir (%90 AUC skoru).



Şekil 2. Popülasyon içinde KRK sınıflandırmasında elde edilen AUC skorları

Popülasyonlar arası KRK analizi: Tür seviyesinde popülasyonlar arası gerçekleştirilen KRK analizinde eğitim için bir popülasyon verisi, test için bir başka popülasyon verisi kullanılmaktadır. Bu durum her bir popülasyon için tekrarlanmaktadır. Şekil 3 de popülasyonlar arası analizlerde elde edilen AUC skorları gösterilmektedir. Tüm özniteliklerin kullanılması ile gerçekleştirilen analizlerde en yüksek sonuç, eğitim için Japon (JPN) veri seti, test için Japon (JP) veri seti kullanıldığında elde edilmektedir. Bu

deneyde Rastgele Orman algoritması kullanılarak %99 AUC skoru elde edilmiştir. Bu deneydeki eğitim ve testte kullanılan her iki veri seti de Japon popülasyonuna ait olduğu için diğer farklı popülasyonlar arası deney sonuçlarına göre daha yüksek AUC skoru elde edilmiştir. Popülasyonlar arası deneylerdeki yüksek AUC skorları, değerlendirilen popülasyon çiftindeki KRK hastalarında benzer mikrobiyom profilleri olduğuna işaret eder. Eğitim için Japon (JPN) ve test için Japon (JP) popülasyonları için en yüksek sonuçlar elde edilmesinden dolayı Tablo 2 de ayrıntılı değerlendirme metrikleri gösterilmektedir.



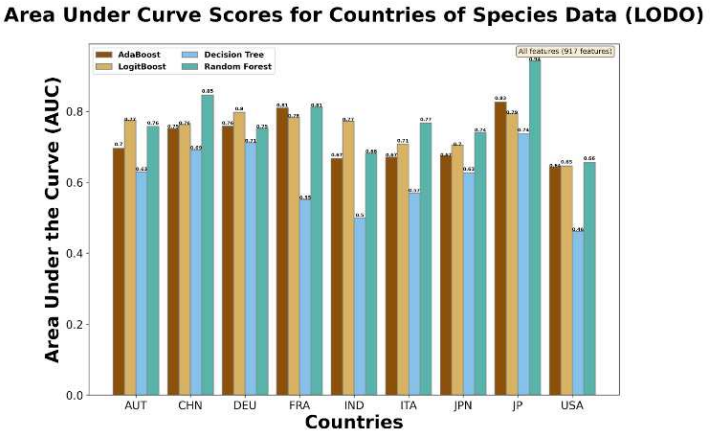
Şekil 3. Popülasyonlar arası KRK sınıflandırmasında elde edilen AUC skorları

Tablo 2 de popülasyonlar arası KRK değerlendirme yöntemi ile eğitim veri seti için Japon (JPN) test veri seti için Japon (JP) popülasyonları ile kullanılarak elde edilen analiz sonuçları gösterilmektedir.

Model	Acc	Sensitivity	Sensitivty	AUC
Adaboost	0.89	0.70	0.92	0.90
DT	0.75	0.70	0.80	0.75
Logitboost	0.75	0.60	0.88	0.86
RF	0.96	0.80	0.98	0.99

Tablo II. Popülasyonlar arası KRK sınıflandırması ile ilgili değerlendirme metrikleri

Bir popülasyon verisinin test için dışarda bırakılması (Leave One Dataset Out, LODO) yöntemiyle KRK sınıflandırması sonuçları: Bu deneylerde, tür seviyesinde gerçekleştirilen KRK sınıflandırmasında bir popülasyon test edilmek için dışarıda bırakılmaktadır. Diğer 8 popülasyon eğitim için kullanıldıktan sonra dışarda bırakılan popülasyon üzerinde test edilerek model değerlendirilmektedir. Her bir popülasyon için bu yöntem tekrarlanmaktadır. Şekil 4'te test verisi için ayrılmış ülkeler x ekseninde, elde edilen AUC skorları ise y ekseninde gösterilmektedir. Şekil 4 incelendiğinde en yüksek sonuç Rastgele Orman algoritması ile, test için Japon (JP) popülasyonunun, eğitim için kalan diğer popülasyon verilerinin topluca kullanılmasıyla elde edilmiştir (%94 AUC skoru). Bir sonraki başarılı sonuç ise Cin (CHN) popülasyonu için Rastgele Orman algoritması kullanılarak elde edilmiştir (%85 AUC skoru). Farklı sınıflandırma algoritmalarının performansları kıyaslandığında, 9 farklı veri seti içerisinden 5 veri setinde Rastgele Orman algoritması daha yüksek AUC skoru üretmektedir. 9 KRK veri setinden 3 tanesinde ise LODO analizlerinde Logitboost algoritması diğer algoritmalardan üstün sonuç göstermektedir. Bu durum KRK sınıflandırması için Rastgele Orman algoritmasının diğer algoritmalardan daha yüksek performans metrikleri ürettiğini göstermektedir.



Şekil 4. Bir popülasyonun test için dışarda bırakılması (LODO) yöntemi ile gerçekleştirilen KRK sınıflandırmasında elde edilen AUC skorları

IV. SONUÇ

Bu çalışmada KRK tanısı için tür seviyesinde nisbi bolluk verilerini öznitelik olarak kullanan ve makine öğrenmesi yöntemleri yardımıyla popülasyona özgü meta analizler gerçekleştiren bir yapay zekâ modeli önerilmektedir. 1262 adet kontrol ve vaka örneği için 917 türün nisbi bolluk değerlerini içeren veri setleri farklı sınıflandırma modelleri kullanılarak analiz edilmiştir.

Gerçekleştirilen meta analiz sonuçları 3 ayrı başlık altında sunulmuştur: i) Popülasyon içi, ii) popülasyonlar arası, iii) LODO analizleri. 8 farklı popülasyon ve 9 ayrı veri seti üzerindeki bu 3 ayrı meta-analizde elde edilen sonuçlar içerisinde en başarılı sonuç, Rastgele Orman algoritması ile popülasyonlar arası analizde elde edilmiştir. Bu deneyde eğitim için Japon (JPN) verisi ve test için Japon (JP) verisi kullanılarak %99 AUC skoru elde edilmiştir.

Önerilen yöntemin başarısı, KRK sınıflandırma çalışmaları içerisinde başarı oranı yüksek çalışmalar arasındadır. Ancak öznitelik sayısının fazla olması, hesaplama maliyetini artırmaktadır. İleriki çalışmalarda bu öznitelikleri çeşitli öznitelik seçim algoritmaları uygulanarak daha az sayıya düşürerek yüksek tahmin başarısı elde etmek için model tasarlanması planlanmaktadır. Ayrıca farklı tür metagenomik verilere de uygulanacak bu yöntem ile KRK sınıflandırma için farklı öznitelik türleri değerlendirilecektir.

TEŞEKKÜR

Bu çalışmada gerçekleştirilen analizlerin bir kısmı TUBITAK ULAKBIM TRUBA iş istasyonları kullanılarak elde edilmiştir.

KAYNAKÇA

[1] A. Dokht Khosravi, S. Seyed-Mohammadi, A. Teimoori, and A. Asarehzadegan Dezfuli, "The role of microbiota in colorectal

cancer," *Folia Microbiol*, vol. 67, no. 5, pp. 683–691, Oct. 2022, doi: 10.1007/s12223-022-00978-1.
[2] M. A. R. Escalona, J. de F. Poloni, M. J. Krause, and M. Dorn, "Meta-analyses of host metagenomes from colorectal cancer patients reveal strong relationship between colorectal cancer-associated species," *Mol. Omics*, vol. 19, no. 5, pp. 429–444, Jun. 2023, doi: 10.1039/D3MO00021D.
[3] S. Sakata and D. W. Larson, "Targeted Therapy for Colorectal Cancer," *Surgical Oncology Clinics*, vol. 31, no. 2, pp. 255–264, Apr. 2022, doi: 10.1016/j.soc.2021.11.006.
[4] Y. Cheng, Z. Ling, and L. Li, "The Intestinal Microbiota and Colorectal Cancer," *Frontiers in Immunology*, vol. 11, 2020, Accessed: Jun. 18, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fimmu.2020.615056
[5] E. Sardo *et al.*, "Multi-Omic Approaches in Colorectal Cancer beyond Genomic Data," *Journal of Personalized Medicine*, vol. 12, no. 2, Art. no. 2, Feb. 2022, doi: 10.3390/jpm12020128.
[6] J. Zhang *et al.*, "Expansion of Colorectal Cancer Biomarkers Based on Gut Bacteria and Viruses," *Cancers*, vol. 14, no. 19, Art. no. 19, Jan. 2022, doi: 10.3390/cancers14194662.
[7] S. M. Ahmad Kendong, R. A. Raja Ali, K. N. M. Nawawi, H. F. Ahmad, and N. M. Mokhtar, "Gut Dysbiosis and Intestinal Barrier Dysfunction: Potential Explanation for Early-Onset Colorectal Cancer," *Frontiers in Cellular and Infection Microbiology*, vol. 11, 2021, Accessed: Jun. 15, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcimb.2021.744606
[8] L. J. Marcos-Zambrano *et al.*, "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment," *Frontiers in Microbiology*, vol. 12, 2021, Accessed: Nov. 01, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2021.634511
[9] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, and N. Segata, "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights," *PLOS Computational Biology*, vol. 12, no. 7, p. e1004977, Jul. 2016, doi: 10.1371/journal.pcbi.1004977.
[10] J. Wirbel *et al.*, "Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer," *Nat Med*, vol. 25, no. 4, Art. no. 4, Apr. 2019, doi: 10.1038/s41591-019-0406-6.
[11] A. M. Thomas *et al.*, "Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation," *Nat Med*, vol. 25, no. 4, Art. no. 4, Apr. 2019, doi: 10.1038/s41591-019-0405-7.
[12] N. T. Baxter, M. T. Ruffin, M. A. M. Rogers, and P. D. Schloss, "Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions," *Genome Medicine*, vol. 8, no. 1, p. 37, Apr. 2016, doi: 10.1186/s13073-016-0290-3.
[13] A. Kishk *et al.*, "A Hybrid Machine Learning Approach for the Phenotypic Classification of Metagenomic Colon Cancer Reads Based on Kmer Frequency and Biomarker Profiling," in *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*, Dec. 2018, pp. 118–121. doi: 10.1109/CIBEC.2018.8641805.
[14] F. Beghini *et al.*, "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3," *eLife*, vol. 10, p. e65088, May 2021, doi: 10.7554/eLife.65088.
[15] M. R. Berthold *et al.*, "KNIME - the Konstanz information miner: version 2.0 and beyond," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 26–31, Nov. 2009, doi: 10.1145/1656274.1656280.
[16] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*.