GRADUATE SCHOOL OF ENGINEERING & SCIENCE

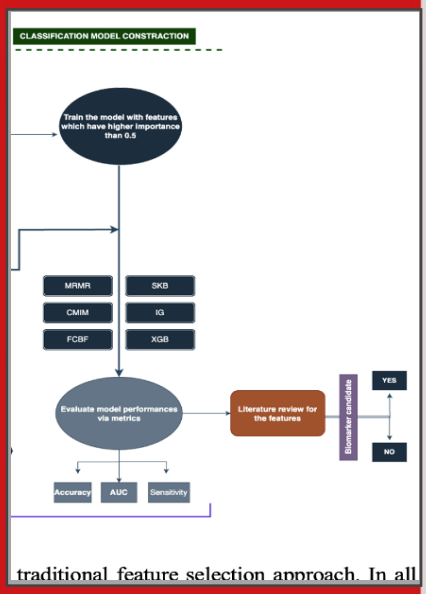
Bioengineering Department

A. Master's Thesis

Beyza Canakcimaksutoglu

beyza.canakcimaksutoglu@agu.edu.tr

0009-0007-3496-6371



traditional feature selection approach. In all

Thesis Advisor

Assoc. Dr. Burcu Bakir Gungor

Prof. Dr. Malik Yousef

burcu.gungor@agu.edu.tr
malik.yousef@gmail.com

22.05.2025

Identifying Potential Taxonomic Biomarkers of Gastrointestinal Cancers from Human Microbiota Using the Grouping-Scoring-Modeling (G-S-M) and Traditional Feature Selection Approaches

abstract Analysis of microbial abundance values holds potential for cancer prediction. This study aims to identify shared microbial biomarkers among gastrointestinal (GI) cancer patients using both tissue and blood samples—an area not previously studied in parallel. This study analyzed blood and tissue samples, focusing on head and neck, esophagus, stomach, colon, and colorectal cancers, processing them individually. By performing decontamination steps, processing non-human genetic codes, determining microorganisms and their abundances at the species level, the TCMA data set was created from the "Cancer Genome Atlas", which collected tissue and blood samples from cancer patients. Traditional feature selection algorithms (CMIM, mRMR, FCBF, IG, XGB, and SKB) reduced the high-dimensional feature space. Classification performance was evaluated using a forest classifier with 100-fold Monte Carlo cross-validation. Moreover, the MicrobiomeGSM model, which was created to decrease the feature size and prediction time via a grouping method, was trained, and the generalizability of the MicrobiomeGSM model was showcased. Traditional feature selection methods and the biological data-based MicrobiomeGSM model were applied, and their performance was compared. In the future, common biomarker candidates may help to understand the possibility of metastasis, and medical doctors can decide their treatment path of patients.

keywords *Taxonomic Biomarker Selections, Machine Learning, Grouping Scoring Modeling, Metagenomics, Feature Selection*

özet Mikrobiyal bolluk değerlerinin analizi, kanser tahmini için bir potansiyel taşır. Bu çalışma, daha önce paralel olarak incelenmemiş bir alan olan hem doku hem de kan örnekleri kullanarak gastrointestinal (GI) kanser hastaları arasında paylaşılan mikrobiyal biyobelirteçleri belirlemeyi amaçlamaktadır. Bu çalışma, baş ve boyun, yemek borusu, mide, kolon ve kolorektal kanserlere odaklanarak kan ve doku örneklerini analiz etti. Dekontaminasyon adımları gerçekleştirilerek, insan olmayan genetik kodlar işlenerek, tür düzeyinde mikroorganizmalar ve bollukları belirlenerek, kanser hastalarından doku ve kan örnekleri toplayan "Kanser Genom Atlası"ndan TCMA veri seti oluşturuldu. Geleneksel özellik seçimi algoritmaları (CMIM, mRMR, FCBF, IG, XGB ve SKB) yüksek boyutlu özellik alanını daralttı. Sınıflandırma performansı, 100-kat Monte Carlo çapraz doğrulaması olan bir Random Forest kullanılarak değerlendirildi. Ayrıca, gruplama yöntemi ile özellik boyutunu ve tahmin süresini azaltmak için oluşturulan MicrobiomeGSM modeli, hem kan hem de dokudan türetilen örnekler kullanılarak eğitildi ve MicrobiomeGSM modelinin genelleştirilebilirliği sergilendi. Geleneksel özellik seçimi yöntemleri ve biyolojik veri tabanlı MicrobiomeGSM modellerinin performansları karşılaştırıldı. Gelecekte, ortak biyobelirteç adayları doktorların metastaz olasılığını anlamasına yardımcı olabilir ve tedavi yollarına buna göre karar verilebilir.

anahtar kelime *Taksonomik Biyobelirteç Seçimleri, Makine Öğrenimi, Gruplama Puanlama Modellemesi, Metagenomik, Özellik Seçimi*