GRADUATE SCHOOL OF ENGINEERING & SCIENCE
AGÜ

Electrical and Computer Engineering

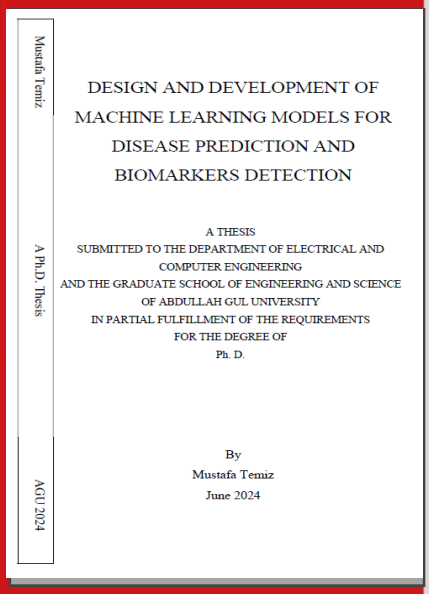
Ph.D. Thesis

Design and Development of Machine Learning Models for Disease Prediction and Biomarkers Detection

Mustafa Temiz

mustafa.temiz@agu.edu.tr

0000-0002-2839-1424



Mustafa Temiz

DESIGN AND DEVELOPMENT OF MACHINE LEARNING MODELS FOR DISEASE PREDICTION AND BIOMARKERS DETECTION

A THESIS
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Ph. D.

By
Mustafa Temiz
June 2024

A Ph.D. Thesis

AGU 2024

abstract In medical science, the prediction of diseases and the identification of biomarkers play an important role in the diagnosis and treatment of various health conditions. The recent proliferation of data mining techniques has accelerated the development of disease prediction systems. In particular, machine learning methods are an effective way to analyze medical data and identify patterns to predict the likelihood of the disease development. Machine learning methods also help to identify biomarkers. Recently, the increasing incidence and mortality rates of inflammatory bowel disease, colorectal cancer and type 2 diabetes have drawn researchers' attention to these research areas. The aim of this thesis is to reduce the number of features and improve the prediction performance of machine learning based on complex biological datasets with a large number of disease-related features, as well as to identify potential biomarkers. In this thesis, three different studies are presented. The first study predicts eleven different cancer subgroups using miRNA data and biological domain knowledge and identifies potential biomarkers for these diseases. The second study predicts three different diseases using metagenomic data and biological domain knowledge and identifies potential biomarkers. The third study uses metagenomic data related to colorectal cancer to conduct global and population-based comprehensive experiments with traditional feature selection methods to identify potential biomarkers. This thesis presents a promising avenue for early disease detection, facilitating expedited treatment protocols, improving human survival rates, and potentially alleviating economic burdens within these critical research domains.

keywords disease prediction, machine learning, identify biomarkers, feature selection, bioinformatics

özet Tıp biliminde, hastalıkların tahmini ve biyobelirteçlerin tanımlanması, çeşitli sağlık koşullarının teşhis ve tedavisinde önemli bir rol oynamaktadır. Veri madenciliği tekniklerinin son zamanlarda yaygınlaşması, hastalık tahmin sistemlerinin gelişimini hızlandırmıştır. Özellikle makine öğrenim yöntemleri, tıbbi verilerin analizinde ve hastalığın ortaya çıkma olasılığını tahmin etmeye yönelik kalıpların belirlenmesinde etkili bir yöntemdir. Makine öğrenim yöntemleri, biyobelirteçlerin tanımlanmasına da yardımcı olmaktadır. Son zamanlarda inflamatuar bağırsak hastalığı, kolorektal kanser ve tip 2 diyabet hastalıkları ile karşılaşma sıklığının artması ve artan ölüm oranları araştırmacıların dikkatini bu araştırma alanlarına çekmektedir. Bu tezin amacı, hastalık ile ilişkili karmaşık ve çok sayıda özellik içeren biyolojik veri setlerinden yola çıkarak özelliklerin sayısını azaltmak ve makine öğrenmesi tahmin performansını artırmaktır ve ayrıca potansiyel biyobelirteçleri tanımlamaktır. Bu tezde üç farklı çalışma tanıtılmaktadır. İlk çalışma miRNA verileri ve biyolojik alan bilgisi kullanılarak on bir farklı kanser alt grubu tahmin edilmekte ve bu hastalıklar için olası biyomarkörler belirlenmektedir. İkinci çalışma da metagenomik veriler ve biyolojik alan bilgisi kullanılarak üç farklı hastalık tahmin edilmekte ve olası biyomarkörler belirlenmektedir. Üçüncü çalışma kolorektal kanser ile ilişkili metagenomik verileri kullanarak geleneksel özellik seçim yöntemleri ile küresel ve popülasyonlara bağlı kapsamlı deneyler gerçekleştirilmekte ve olası biyomarkörler belirlenmektedir. Bu tez, erken hastalık tespiti için umut verici bir yol sunmakta, hızlandırılmış tedavi protokollerine olanak tanımakta, insan sağkalım oranlarını artırmakta ve bu kritik araştırma alanlarında potansiyel olarak ekonomik yükleri azaltmaktadır.

anahtar kelime hastalık tespiti, makine öğrenmesi, biyomarkör/biyobelirteç belirleme, özellik seçimi, biyoinformatik

Thesis Advisor

Doç. Dr. Burcu Bakır Güngör

burcu.gungor@agu.edu.tr

26.06.2024