T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

BİYOİNFORMATİK UYGULAMALARINDA MAKİNE ÖĞRENME YÖNTEMLERİNİN GELİŞTİRİLMESİNE YÖNELİK ÇOK KRİTERLİ YAKLAŞIM

Zeliha GÖRMEZ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Danışman

Prof. Dr. Ahmet SERTBAŞ

II. Danışman

Dr. Hüseyin ŞEKER

İSTANBUL

ÖNSÖZ

Tez çalışmam boyunca yardım ve yönlendirmelerinden ötürü değerli hocalarım Prof. Dr. Ahmet SERTBAŞ ve Dr. Hüseyin ŞEKER'e, ayrıca çalışmam boyunca her durumda benden desteklerini esirgemeyen değerli hocam Doç. Dr. Olcay KURŞUN ve çalışma arkadaşım Dr. Ergün Gümüş'e teşekkür ederim. Ayrıca herzaman yanımda olan çok değerli arkadaşım Dr. Rüya ŞAMLI'ya teşekkür ederim.

Bana desteklerini esirgemen bir diğer grup ise çalışma arkadaşlarımdır. Değerli yöneticim Dr. Mahmut Şamil SAĞIROĞLU'na biyoenformatik konusunda birçok şeyi pratik hayatta uygulamalı öğrettiği için teşekkür ederim. Yine biyolojik konulardaki eksiklikerimi müsamaha ile karşılayan ve sorularıma sabırla cevap veren Dr. Burcu Bakir GÜNGOR'e teşekkür ederim. Son olarak, çalışmalarımın sonuçlarının biyolojik olarak doğrulanmasında yardımlarını esirgemeyen çalışma arkadaşım Betül YÜCETÜRK'e teşekkür ederim.

Doktora sürecim boyunca en büyük destekçilerim olan annem ve babama sonsuzteşekkür ederim. Akademik çalışmalarımda beni her zaman yüreklendiren eşime teşekkür ederim.

Doktora sürecimde hayatıma giren kızım ZEHRA LEYLA'dan, zor zamanlarımda ilk fedakârlığı ondan yaptığım için ve özellikle son dönemlerde annesini bilgisayarla paylaşmak zorunda bıraktığım için özür dilerim. Bu süreçte tüm sabırlarından dolayısevgili kızıma teşekkür ederim.

Mayıs 2014 Zeliha GÖRMEZ

İÇİNDEKİLER

Sayfa No

ŞEKİL LİSTESİ

Sayfa No

TABLO LİSTESİ

Sayfa No

SİMGE VE KISALTMA LİSTESİ

Simgeler	Açıklama
A	**:** AHP de alternatifler
C	**:** Condorcet algoritmasında adaylar, AHP de kriterler kümesi
c	**:** Sınıf bilgisini içeren öznitelik
d	**:** Boyut sayısı
D(.)	**:** Uzaklık fonsiyonu
diff(.)	**:** Relief algoritmasında ödül/ceza fonksiyonu
$\mathbf{f_i(c_1,c_2)}$	**:** Condorcet algoritmasınd tercih fonskiyonu
$\mathbf{f_i(x)}$	**:** PO algoritmasında amaç vektör fonksiyonu
g(.)	**:** Sınıflandırmada model
H(X)	**:** X dağılımındaki entropi
I(X;Y)	**:** X,Y dağılımlarının ortak olasılık fonksiyonu
k	**:** İndirgenmiş öznitelik sayısı
M	**:** MC4 algoritmasında geçiş matrisi
N	**:** Örnek sayısı
p	**:** İstatistiksel anlamlılık
P(.)	**:** Olasılık fonksiyonu
P(.|.)	**:** Koşullu olasılık fonksiyonu
U	**:** MC4 algoritmasında durum uzayı
V	**:** Condercet algoritmasında tercih kümesi
v	**:** MC4 algoritmasında durum vektörü
W	**:** AHP de ağırlıklar
W[A]	**:** Relief algoritmasında A özniteliğinin ağırlığı
$\mathbf{w_i}$	**:** AHP de i. alternatifin önceliği
θ	: Sınıflandırmada model paremetreleri

Kısaltmalar **Açıklama**

A	**:** Adenin
AC	**:** Doğruluk (Accuracy**)**
AHP	**:** Analitik Hiyerarsi Proses
C	**:** Cytosine
CD	**:** Kararlılık Katsayısı
CFS	**:** Correlation Based Feature Selection
ÇG	**:** Çapraz Geçerleme
DNA	**:** Deoksiribonükleik Asit
FFS	**:** Forward Feature Selection
G	**:** Guanine
GWAS	**:** Genome Wide Assosiation Study
HGDP	**:** Human Genome Diversity Project
KNN	**:** K-Nearest Neighbor
Kr	**:** Kromozom
LOO	**:** Leave-One-Out
MC4	**:** Markov Zinciri 4
MI	**:** Mutual Information-Karşılıklı Bilgi
MI-C	**:** MI-Kümeli
MI-D	**:** MI-Ayrık
mRMR	**:** En Küçük Artıklık En Büyük İlgililik (Minimum Redundancy Maximum Relevance)
PC	**:** Temel Bileşen
PO	**:** Pareto Optimallik
POK	**:** Pareto Optimum Küme
RNA	**:** Ribonükleik Asit
Sen	**:** Duyarlılık (Sensitivity)
SNP	**:** Single Nücleotide Polimorfizm
SNR	**:** Signal To Noise Ratio
Spe	**:** Özgüllük (Specificity)
SVM	**:** Support Vector Machine
SVM-RFE	**:** Support Vector Machine-Recursive Feature Elimination
T	**:** Thymin
TBA	**:** Temel Bileşen Analizi
U	**:** Uracil

ÖZET

DOKTORA TEZİ

BİYOİNFORMATİK UYGULAMALARINDA MAKİNE ÖĞRENME YÖNTEMLERİNİN GELİŞTİRİLMESİNE YÖNELİK ÇOK KRİTERLİ YAKLAŞIM

Zeliha GÖRMEZ

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Ahmet SERTBAŞ

II. Danışman: Dr. Hüseyin ŞEKER

Biyo-işaretçi seçimi genom sonrası elde edilen yüksek boyutlu biyolojik verilerin analizlerinin önemli bir parçasıdır ve biyo-işaretçilerin temsil gücü en yüksek alt kümesini seçmeyi hedefler. Ancak, biyolojik verilerin yüksek boyutlu yapıda olması nedeniyle, seçim süreci çok zor bir iştir. Bu süreç, veri setindeki örnek varyasyonlarından (değişintilerden) ve seçim metodundan da bağımsız olmalıdır. Bu çalışmada, çok amaçlı optimizasyon yöntemi olan Pareto Optimallik (PO) ile çok kriterli karar verme yöntemi olan Analitik Hiyerarşi Proses (AHP) yöntemini birleştiren yeni bir karma metot önerilmiştir. Yöntem farklı biyo-işaretçi seçim yöntemleri ile de kullanılabilmektedir. Bu çalışmada önerilen çok kriterli yaklaşımlar çeşitli yüksek boyutlu biyolojik veriler üzerinde test edilmiştir. Alınan sonuçlar PO yönteminin biyolojik verilerde tanımlanmış problem ile ilgili öznitelikleri başarılı bir şekilde seçtiğinigöstermiştir. Ayrıca AHP yönteminin seçilmiş az sayıda biyo-işaretçinin kendi arasında önceliklendirilmesinde kullanılabileceğini de gösterilmiştir.

Mayıs 2014, 165 sayfa.

Anahtar kelimeler: Biyo-işaretçi Keşfi, Öznitelik Seçimi, Çok Kriterli Yaklaşım, Çok Amaçlı Optimizsayon, Pareto Optimallik, Analitik Hiyerarşi Proses

SUMMARY

Ph.D. THESIS

MULTI-CRITERIA APPROACH TO DEVELOPMENT OF MACHINE LEARNING METHODS IN BIOINFORMATICS

Zeliha GÖRMEZ

İstanbul University

Graduate School of Science and Engineering

Department of Computer Engineering

Supervisor:Prof. Dr. Ahmet SERTBAŞ

Co-Supervisor:Dr. Hüseyin ŞEKER

Bio-marker selection is the important part of high dimensional biological data, that is obtain from post-genome, analysis and it aims finding most representative subset of the bio-markers. But selection process is a challenging task due to the high dimensional nature of gene expression data. This should also be independent of sample variations in the dataset. In this paper we present a novel hybrid method that incorporates a multi-objective optimization method, called Pareto Optimal approach (PO) with a multi-criteria decision making method, called Analytical Hierarchy Process (AHP). The method is further supported with different bio-marker selection methods. The multi-criteria approaches proposed in this study were tested on various high-dimensional biological data. The results show that PO method selects the features related to the defined problem in biological data successfully. Furthermore; the results also show that AHP method could be used to prioritize a few selected bio-markers among themselves.

May 2014, 165 pages.

Keywords: Bio-marker discovery, Feature selection, Multi-criteria Approach, Pareto Optimality, Analytical Hierarchy Process

1. GİRİŞ

Makina öğrenme yöntemleri ile biyolojik verilerde gizli olan yapılar ve biyo-işaretçiler keşfedilmeye çalışılmaktadır. Bu amaçla, verilerin analizleri için boyutlarının azaltılması, özelliklerinden bir alt kümenin seçilmesi, verinin sınıflandırılması, kümelenmesi ve yeni durumlarının tahminlenmesi en temel uygulamalardır.

Biyolojik veri analizinde, tanımlanan problem ile ilişkili öznitelik (örneğin mikrodizilerde gen) alt kümelerini seçme işlemi çok önemli ve kritik bir süreçtir. Çünkü biyolojik veri kümleri çok fazla öznitelik içermesine rağmen, bunların küçük bir grubu problem ile ilişkilidir. Biyo-işaretçi seçiminin amaçları şu şekilde özetlenebilir: (i) bir biyolojik veride tanımlanmış bir problem (örneğin hastalığın başlangıç ve ilerleyişinde rolü bulunan genetik ağ, yapı ve mekanizmaları) ile ilişkili öznitelikleri ortaya çıkarmak, (ii) Kestirim modelinin doğruluğunu ve yorumlanabilirliğini geliştirmek için veridenilgisiz özniteliklerin çıkarılmasını sağlamaktır.

Biyo-işaretçi seçimi, post-genom verilerinin (örneğin mikrodizi gen ifadesi) doğasında bulunan yüksek boyut özelliği nedeni ile zorlu bir işlemdir. Buna bağlı olarak, biyo-işaretçilerin genelleştirilebilir ve temsil gücü yüksek bir alt kümesini seçmek de büyük önem arz etmektedir. Bununla birlikte biyolojik verilerdeki küçük örnek boyutları ve yüksek girdi boyutları bu çalışmalarda karşılaşılan temel zorluklardır. Bu yüzden, bu problemlerin üstesinden gelmek için, pek çok öznitelik seçim algoritması sunulmuştur. Sunulmuş birçok çözüm yönteminin bulunmasının dezavantajı, verilen bir veri seti için, farklı seçim yöntemleri uygulandığında farklı öznitelikler yada öznitelik grupları sonucuna ulaşılmasıdır. Aynı zamanda seçim işlemi, veri kümesindeki örnek varyasyonundan da bağımsız olarak yapılabilmelidir. Aynı öznitelik seçim algoritması, birbirinden çok çok az farklı olan iki veri setine uygulansa bile farklı özniteliklerin seçilmesi sonucunu doğurabilir. Konuya ilişkin yapılan birçok çalışmada, öznitelik seçimi ve derecelendirme işleminin veri kümesi varyasyonu ve seçim metodu ile ilişkili olduğu gösterilmiştir[1-3]. Bu çalışmalar, öznitelik seçme ve önceliklendirme işlemiiçin

sadece tek yöntem yada yalnız bir öğrenme kümesi kullanmanın güvenilmez olduğunugöstermiştir.

Bir öznitelik seçim algoritmasının, yanlılık/hata miktarını düşük tutmak amaçlı olarak veri seti varyasyonlarından etkilenmesi beklentisine rağmen, aralarında yüksek benzerlik ve örtüşme bulunan veri setlerinde düşük varyans olması beklenir. Bu iki rakip faktör arasındaki denge, yanlılık-varyans çelişkisi olarak bilinir.İstatistiksel yanlılık, modelin eğitim kümesine uyum derecesini gösterir,ancak genellemeye ilişkin bir ölçü değildir. Diğer taraftan; istatistiksel varyans, bir veri örneğinden diğerine model etkinliğinde meydana gelen sapmadır[4].

Biyoenformatik uygulamalardaki öznitelik seçimi yöntemlerinin yanlılık ve varyansı üzerinde durduğumuz bir vaka (hipertansiyon gen ifadelenme verisi) çalışmasında, öznitelik seçiminin örnek uzayı ve seçim metoduna bağımlılığını gözlemledik. Çalışmadaki amacımız, öznitelik seçim algoritmalarının, eğitim için küçük farklılıkları olan örnekler kullanılsa bile yüksek varyanslara sahip olabileceğini göstermekti. Buamaçla, farklı çapraz doğrulama yöntemleri içinfarklı genlerin seçildiğigösterilmiştir. Örneğin iyi bilinen birini dışarıda bırak (leave-one-out, LOO) çapraz doğrulama yönteminde, her seferinde bir örnek dışarıda tutulduğunda, aynı öznitelik seçim algoritması en açıklayıcı gen olarak farklı genleri seçmektedir. K-kat çapraz geçerlemede de aynı durum sözkonusudur.Her kat dışarıda bırakıldığında veya k farklı değerler aldığında, aynı öznitelik seçim algoritması farklı öznitelikleri seçmiştir. Bu durum, birden fazla öznitelik seçimi hedeflendiği zaman da aynıdır. Şöyle ki; herbir farklı çalıştırma sonucunda, aynı öznitelik seçim yönteminin seçtiği en tahmin edici öznitelik grupları arasında anlamlı bir örtüşme bulunmadığı tespit edilmiştir.

Bu çalışmada, veri setindeki örnek varyasyonunda (değişinti) etkilenmeyen bir seçim yöntemi için, çok amaçlı bir en iyileme yöntemi olan “Pareto Optimal” (PO) yaklaşımı ile çokkriterlibir karar verme yöntemi olan Analitik Hiyerarşi Prosesi (AHP) birleştiren yeni bir karma modelin oluşturulması amaçlanmıştır. Önerilen bu yöntem, farklı öznitelik seçimi ve sınıflandırma yöntemleri ile de desteklenebilmektedir.

Bilinmektedir ki verinin farklı alt uzaylarında uzmanlaşmış yerel öğrenicilerin ve ya farklı alanlarda uzmanlaşmış metotların birleştirilmesi daha başarılı modeller ortaya

çıkarmaktadır[4]. Bu gerçeklikten yola çıkarak, bu çalışmada önerilen yöntemde, öznitelik seçiminde veri seti bağımlılığından kaçınmak için yerel öğrenciler/uzmanlar tarafından kullanılmak üzere farklı örnek alt kümeleri ile çoklu öğrenme kümeleri oluşturulmuştur.Öznitelik (örneğin genler) kümelerinin her bir öğrenme kümesini kullanan yerel uzmanlarca üretilmiş sıralama değerleri maksimize etmek istediğimiz kriterlerdir. Tüm kriterler dikkate alınarak, optimum değerli genler PO ile seçilmiştir. Ancak PO, tek başına, özniteliklerin amaca uygun altkümelerini seçmesine rağmen, öncelikleri ile ilgili herhangi bir bilgi sunmamaktadır. Bu problemin üstesinden gelebilmek için PO yöntemi, bir diğer çok kriterli karar verme yöntemi olan AHP ile birleştirilmiştir.AHP, seçilen özniteliklerinyerel uzmanlarca oluşturulmuş sıralama değerlerinden bir konsensüs skor elde etmek için kullanılmıştır. Bu konsensüs skorun sıralaması özniteliklerin nihai önem derecesidir. Bahsedilen sebeplerin sonucu olarak, önerilen öznitelik seçme ve önceliklendirme karmametot, veri seti varyasyonlarından ve seçim metodundan etkilenmemektedir.

Öznitelik seçimin bir diğer bağımlılığı seçim yöntemleridir. Bu çalışmada, veri setinde tanımlanmış problem ile ilişkili özniteliklerin seçilmesindemetot bağımlığını önlemek için ise, farklı amaçlara hizmet eden çeşitli seçim metotlarının çok kriterli karar verme yöntemleri ile birleştirilmesi önerilmiştir. Farklı öğrenme kümelerini kullanan yerel uzmanları birleştirmek için kullanılmasını önerdiğimiz PO yöntemi, metotların birleştirilmesi amacı ile kullanılabilir. Bir veri kümesi üzerinde farklı amaçlara hizmet eden metotların PO ile başarılı bir şekilde birleştirilebileceği gösterilmiştir. Yukarıda bahsedildiği gibi, farklı metotları dikkate alarak öznitelik seçimi başarılı bir şekilde gerçekleştiren PO, öznitelikleri önceliklendirmemektedir. Seçilen özniteliklerin kendi arasında sıralanması gerektiğinde ise PO ile AHP yöntemi birleştirilerek seçilen öznitelikler önceliklendirilmiştir.

Bu çalışmada, öznitelik seçim yöntemlerinin geliştirilmesi için önerilen yeni çok kriterlikarma yaklaşımlar,genel kullanıma açıkolan dört (üç kanser ve bir hipertansiyon gen ifadesi verisi) mikrodizi veri kümesinde ve bir genom veri kümesinde uygulanmıştır. Mikrodizi veri kümelerinde farklı öğrenme kümlerini kullanan yerel uzmanlar birleştirilerek genler hastalıkla ilişki derecelerine göre sıralanmıştır. Ayrıca bu karma yöntemin, SNR, SVM-RFE ve mRMR, Relieff gibi çokça kullanılan tekniklerden

daha iyi sonuç ürettiği gösterilmiştir. Önerin yöntem az sayıda gen ile yüksek sınıflandırma başarısı elde etmiştir. Bu şekilde hastalık etkeni aday gen sayısının azaltımı, potansiyel ilaç adaylarının ve yeni tedavi yöntemlerinin etkin bir şekilde geliştirebilmesine yol açması önemli bir başarıdır.

Tekli nükleotid değişimlerinden (SNP) oluşan genom verisinde ise, farklı amaçlara hizmet eden metotların birleştirilmesi hedeflenmiştir. Farklı metodlarcaseçilen öznitelikler içinden farklı amaçları optimize edecek SNP'leri seçmek amacı ile PO yönteminin öznitelikleri yarıştırıcı özelliği kullanılmıştır. Sonuçlar farklı amaçları gözeten metotların birleştirilmesinin ortak amaca daha iyi hizmet ettiğini göstermiştir.

Çalışmanın geri kalanı şu şekilde organize edilmiştir: Bölüm 2 konu ile ilgili genel bilgileri içermektedir. Bu bölümde öznitelik seçmenin gerekliliği, uygulanan yaklaşımların temel prensipleri, genel alanda ve biyoenformatik alanında uygulanan yöntemler, öznitelik seçme yöntemlerinin geliştirilmesine yönelik bizim önerdiğimiz çok kriterliyaklaşımın temel prensipleri anlatılmıştır. Ayrıca bu bölümde biyolojik veri kümeleri ve öznitelikleri sunulmuştur. Bölüm 3, kullanılan malzeme ve yöntemleri içermektedir. Bu bölümde, öznitelik derecelendirme ve seçme işleminin bazı temel kavramları, veri kümesi çeşitliliğinin öznitelik seçimine etkisi, sınıflandırma yöntemleri, PO, AHP ve seçilen özniteliklerin değerlendirme performansı anlatılmıştır. Ayrıca öznitelik seçim ve derecelendirme yöntemlerinin geliştirilmesine yönelik bu çalışmada önerilen çok kriterli yöntemler de bu bölümde ele alınmıştır. Bölüm 4'te deneysel bulgular verilmiştir. Bu bulguların içerisinde; Pareto Optimal ile öznitelik seçimi sonuçları, AHP tabanlı sıralama sonuçları, seçilen ve sıralanan genlerin sınıflandırma başarıları bulunmaktadır. Son bölümde ise, çalışmanın sonuçları ve ileriki çalışmalar tartışılmıştır

2. GENEL KISIMLAR

2.1BOYUT AZALTMA

Biyolojik veri kümeleri gibi çok yüksek boyutlu verilerin makina öğrenme yöntemleri ile analizinde birçok problemle karşılaşılır. İlk problem analiz için kullanılacak makinaların yüksek donanım gereksinimidir. İkinci problem tüm verinin aynı anda işlenmesi için gerekli zaman kısıtıdır. Bunlardan daha önemli olan ise verinin tüm özniteliklerinin tanımlanan problemle ilişkisi olmadığından makina öğrenme yöntemlerinin başarısının düşmesidir.Dolayısı veri analizinde, veriden ilişkili öznitelik altkümelerinin seçilmesi işlemi en önemli adımlardan biridir. Veri gürültüden arındırıldıktan sonra ancak analize tabi tutulabilir[4]. Verinin analizinden önce ayrı bir işlem olarak boyut azaltılmasını bir kaç madde ile özetlemiştir:

i. Çoğu öğrenme algoritması verinin öznitelik boyutuna (d) ve örneklem sayısına (N) bağlıdır. Dolayısı ile gereken bellek ve işlem sayının azaltılması için verinin boyut sayısını azaltmak gerekir.Bu işlem modelin karmaşıklığını azaltacaktır.
ii. Problem ile ilgisiz bir verinin atılması, onu elde etmek için harcanacak zamanı ve işlem sayısını azaltacaktır.
iii. Küçük veri kümelerinde öğrenme algoritmalarının gürültü, aykırı gözlem gibi veriye bağlı özel nedenlerden daha az etkilendiğinden, bu verilerde daha basit modeller daha güvenilir olur.
iv. Veri kümesinin daha az değişken ile açıklanması durumunda verinin üretim süreci daha iyi anlaşılacağından bilgi çıkarımı kolaylaşacaktır.
v. Veri bilgi kaybetmeksizin bir kaç boyutta görselleştirildiğinde verinin yapısı ve aykırı durumları hakkında gözle analiz mümkün olacaktır.

Boyut azaltmak için iki temel yaklaşım vardır: öznitelik seçimi ve öznitelik çıkarımı. Öznitelik çıkarımı d boyutlu veriden $k < d$ olacakşekilde, k tane yeni özniteliğin üretilmedir. Bu amaçla en iyi bilinen ve sık kullanılan yöntemler temel bileşen analizi – TBA ve doğrusal ayırtaç çözümlemedir[4]. Öznitelik seçimi ise d boyutlu veriden

tanımlanan problem için en çok bilgi taşıyan k tane özniteliği içeren alt kümenin bulunması işlemidir.

Öznitelik seçimi için kullanılan üç ana yaklaşım vardır: filtre, wrapper ve gömülü öznitelik seçim yöntemleri [5]. ttest [6], Fisher oranı ve Relief [7], korelasyon tabanlı öznitelik seçimi (CFS) [8] gibi filtre yöntemleri, basitlikleri ve hızları dolayısıyla post-genom alanında kullanılmaktadır [5]. Bu yöntemler en yüksek skoru sağlayan en küçük olası altkümeyi elde etmeyi amaçlar. Bu yaklaşımlarda genler genellikle amaçlarına uygun bir biçimde (örneğin kestirim başarısı, sınıf etiketi ile korelasyonu gibi) derecelendirilirler. Bununla birlikte, filtre yöntemleri, seçilen kestirim modelinden bağımsızdırlar ve bu sayede herhangi bir kestirim modeline yanlı değildirler. Filtre yöntemleri, genellikle tek değişkenli yöntem olarak sayılır ve öznitelikler arasındaki ilişkileri hesaba katmazlar [5].

SNR, SVM-RFE [9, 10], benzetim tavlaması [11] ve genetik algoritma [8]gibi wrapper yöntemleri, seçilen sınıflandırıcıya bağlı yöntemlerdir. Bu yöntemler ilgili öznitelikleri belirlemek için seçilen sınıflandırıcının başarısını bir değerlendirme kısıtı olarak kullanırlar. Wrapper yöntemler ileriye doğru yada geriye doğru seçim yapabilirler. İleriye doğru seçimde seçilenler kümesi boş iken geriye doğru seçimde tüm öznitelikler seçilenler kümesindedir. İleriye doğru seçim yönteminde her bir adımda seçilenler kümesine yeni öznitelik/öznitelikler eklenir ve kümedeki elaman sayısı artar. Aksine geriye doğru seçim yönteminde her bir adımda kümeden öznitelik çıkarılır ve eleman sayısı azalır. Her ikisinde durma kriteri baştan belirlenmiş bir k öznitelik olabileceği gibi hatayı azaltma konusunda bir farklılık oluşmayana kadar devametme de olabilir [4]. Bu tür yöntemlerin önemli bir dezavantajı yüksek hesaplama maliyetine sahip olmalarıdır. Bunun nedeni, her öznitelik altkümesi için yeni bir sınıflandırıcı eğitimine/değerlendirmesine ihtiyaç duymalarıdır. Yüksek boyutlu bir uzayda veri analizi yapılması gerektiğinde, filtre yaklaşımı, hesaplama maliyetini azaltma açısından en iyi alternatif olarak görülmektedir [12]. Öznitelik seçimi tekniklerinden üçüncüsü olan gömülü yaklaşımlarda filtre ve wrapper yöntemleri birleştirilir.Korelasyon tabanlı öznitelik seçimi ve Taguchi-Genetik Algoritma yöntemlerinin[8] veya Korelasyon katsayıları (correlation coefficients) ve K-en yakın komşuluk yöntemlerinin (K-nearest neighbor-KNN)[12] beraber kullanımı buna örnek olarak verilebilir.

2.2 ÇOK KRİTERLİ ÖZNİTELİK SEÇME

Seçime tabi tutulan özniteliklerin performansı, tekil veya çoğul amaçlara göre hesaplanabilir. Son yıllarda, biyoenformatik çalışmalarında çok amaçlı eniyileme yöntemleri de seçim için kullanılmaya başlanmıştır. Çoğunlukla, öznitelik altkümelerinin performanslarının değerlendirilmesinde iki amaçlı fonksiyonlar kullanılmıştır[13]. Bu noktada ilk amaç, sınıflandırma başarısıdır. Bir altküme, tanımlanan sınıfların (örneğin hasta-kontrol) örneklerini doğru bir şekilde sınıflandırırsa, bu iyi bir altkümedir ve yöntem tarafından seçilen özniteliklerproblemle (örneğin hastalık) ilişkilidir. İkinci amaç, öznitelik altkümesinin boyutu veya model maliyetini en iyilemektir. Birçok çalışmada, araştırmacıların amaçlarından biri, doğruluğu en büyük yada hatayı en küçük kılmak olmuştur. Diğer amaçlar ise eşzamanlı olarak, öznitelik sayısını yada model maliyetini en küçük kılmaktır[11, 14-17].

Çok amaçlı öznitelik seçimi için sıklıkla başvurulan bir yol, çok amaçlı yöntemlerin çeşitli birleştirme teknikleriyle (örneğin ortalama, medyan, minimum, maksimum)[7-12, 14-19]veya daha karmaşık modellerle (Rank Distance Categorization [19], Resampling and Permutation feature Importance[7]) tek amaçlı eniyileme problemine çevrilmesidir.Tek amaca indirgemeye alternatif olarak, yöneylem araştırmalarında popüler bir teknik olan Pareto Optimal (PO) yaklaşımı da öznitelik altkümesi seçimine uygulanabilir.

PO, toplu taşıma [20], sıralama/çizelgeleme [21, 22] ve araç yönlendirme[23, 24] gibi birçok probleme uygulanmıştır.Son yıllarda, PO yaklaşımı gen gibi biyolojik öznitelikler, biyo-işaretçilerin, alt küme seçimiproblemine de uygulanmış ve pek çok gen altkümesinin sınıflandırma performansını değerlendirerek en iyi altkümelerin seçiminde başarılı olmuştur [6, 14-17, 25]. Daha önceki çalışmaların biri genleri filtrelemek için [25], üç amaç fonksiyonu (ortalama eğim-mean slope, eğimin türevi-slope deviation ve geçerli güzergâh- valid trajectories) kullanmıştır. Başka bir çalışmada, fold-change, *p*-değeri ve seçim frekansı gibi tek değişkenli derecelendirme kısıtlarını dikkate alınarak, iki ve üç amaçlı öznitelik seçme işlemi gerçekleştirilmiştir [6]. Ancak bu çalışmaların ortak özelliği, seçilen genlerin önemlerine göre sıralanmamış olmasıdır. Oysaki sıralama,biyo-işaretçi seçiminin önemli çıktılarından biridir.

Analitik Hiyerarşi Prosesi (AHP), yöneylem araştırmalarında yazılım seçimi[26], tedarikçi seçimi [27-29], hasta tercihi [30, 31] kaynak yönetimi ve tahsisi[32, 33] gibi birçok probleme uygulanmış çok amaçlı karar verme yöntemidir. AHP, 1980'de çoklu kriterli karar verme problemini çözebilmek amacıyla Saaty [34] tarafından geliştirilmiştir. Biyoenformatik çalışmalarında, AHP yalnızca bir çalışmada SNP'lerin önceliklendirilmesi amacı ile kullanılmıştır [35]. Bu çalışma, SNP'lerin istatistiksel önem ve biyolojik ilgililiği gibi birçok kriteri AHP yapısı içinde kullanmıştır.Kriterlerin ikili karşılaştırılması için uzman görüşüne başvurulmuştur. Ancak, AHP, hesaplama maliyetinden ötürü yüksek boyutlu verilerdeki tüm öznitelikler için uygun değildir. Bu sebeple, ilk olarak, ilgisiz tüm öznitelikler/genler çıkartılmalı daha sonra geriye kalan az sayıda öznitelik/gen AHP ile sıralanmalıdır.

Bu çalışmanın amacı, var olan öznitelik seçim yöntemleri ile birlikte, PO ve AHP'yi birleştirerek daha kararlı, sağlam bir öznitelik seçim yöntemi geliştirmektir. İlk olarak, Pareto Optimal yaklaşımı, derecelendirme tabanlı öznitelik seçim yöntemlerine çok amaçlı bakışla uygulanmış ve ardından seçilen öznitelikler AHP ile önceliklendirilmiştir. Yerel öğrenicilerin birleştirilmesi ile daha başarılı modeller elde edilebileceği yaklaşımı [4] ile çok kriterli karar verme yöntemleri üzerine kurulan bu karma modelin, verinin farklı alt kümelerini kullanan yerel öğrenicileri birleştirerek veri kümesindekivaryasyonun (değişinti) seçim işlemine etkisini en küçük kılması beklenmektedir. Buna ilaveten,farklı metotların birleştirilmesinin daha başarılı modeller oluşturduğu bilindiğinden [4], önerilen yaklaşımın farklı metotların birleştirerek öznitelik seçim işlemindeki metot bağımlılığını da en aza indirgemesi beklenmektedir.

Öncelikle, öznitelik seçimi için, veri kümesi çeşitliliğine bağımlılıktan kaçınmak maksadıyla verinin farklı alt kümelerini kullanan yerel öğreniciler oluşturulmuştur. Yerel öğrencilerin öznitelikler için ürettiği derecelendirme değerlerini dikkate alarak öznitelik seçimi yapabilençok kriterli karar verme yöntemi olan "Pareto-Optimum" yaklaşımı kullanılmıştır. Özniteliklerin çoklu eğitim kümeleri kullanılarak oluşturulan dereceleri, amaç değerler olarak kullanılmıştır.

Öznitelik seçimi veri kümesi çeşitliliğine bağımlı olduğu gibi, özniteliklerin sıralanmasıda veri kümesi çeşitliliğine bağımlıdır. Çünkü derecelendirme yöntemleri, özniteliklerіseçmeden önce onları sıralar ve skorlarına göre bir altküme seçerler. Bunun

yanında, farklı örnek alt kümelerini kullanan yerel uzmanlar/öğreniciler öznitelikler için farklı bir önem sırası oluşturur. Sıralama problemi, çok kriterli bir karar verme problemi olarak tanımlanırsa, farklı sıralama skorları üreten yerel uzamanlar birleştirilerek bir ortak skor elde edilebilir. Bu ortak skorun sıralaması, özniteliklerin yeni bir önem sıralamasıdır. Bu amaçla, AHP ortak skor üretmek ve öznitelikleri önceliklendirmek amacıyla kullanılmıştır.

2.3 BİYOLOJİK VERİ KÜMELERİ

Biyolojik veri kümeleri dizilim verileri, mikro diziler olarak gruplanabilir. Dizilimler saf metinsel verilerdir. Mikro diziler ise genlerin ifade düzeyini (expression) gösteren sayısal ifadelerdir.

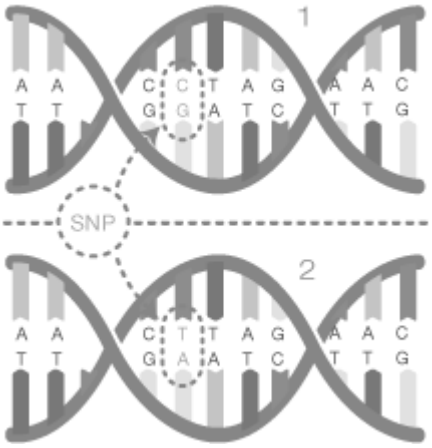
2.3.1 Biyolojik Dizilimler

DNA (Deoksiribonükleik asit), canlıların biyolojik işlevleri için gerekli genetik bilgileri içeren bir nükleik asittir. DNA parçacıkları (gen), canlı hücre için temel yapıtaşı olan proteinlerin yapımından sorumludur. Protein üretimi için kalıp görevi üstlenen genler DNA'ya benzer özelliğe sahip RNA (Ribonükleik asit) dizilimine dönüşür. RNA dizilimleri ise protein dizilimlerinin oluşmasını sağlar.

DNA, genetik bilgiyi taşıyan metin tabanlı en temel biyolojik dizilimdir ve 4 harfli (A, C, G, T) bir alfabesi vardır. **A**denin, **C**ytosine, **G**uanine ve **T**hymin olarak adlandırılan bu harfler ile istenilen uzunlukta kelimeler üretilebilir. Aynı Mors alfabesindeki iki harf (nokta ve çizgi) ve bilgisayar dilindeki 2 harf (0 ve 1) ile oluşturulan kelimeler gibi bu 4 harfli DNA alfabesi ile de yeteri kadar kelime kodlayabiliriz.

Analiz edilecek nükleotid dizilimlerini saf DNA parçaları ve tüm genom dizilimindeki çeşitliliği içeren alt dizilimler olarak düşünülebilir. Genom dizilimindeki çeşitlilik, tek nükleotid polimorfizmi (Single Nukleotid Polimorfizm-SNP) olarak adlandırılır. SNP, DNA diziliminde bir canlıdaki bir baza karşılık, farklı bir canlıda o bazın değişimini ifade eder. Yani bir canlının tüm DNA dizilimin x noktasında A bazı görülürken, diğer bir canlının tüm DNA diziliminin x noktasında T bazının görülmesidir ya da y noktasında G/C bazlarının görülmesidir. Bu durum **Şekil 2.1** gösterilmiştir. Şekilde 2 bireyin çift zincirli DNA'snın 10 bazlık bir görüntüsü şematize edilmiştir. Kalıp zincir

olarak üst iplik alınırsa 1. bireyde 4. bazda C görülürken, 2. birey de aynı bazda T görülmektedir.



Şekil 2.1: Tekli Nükleotid Polimorfizmi-Snp

SNP'lerin görülme sıklığı bin bazda bir olarak düşünülmektedir. Örneğin ~3 milyar baz içeren insan genomunda 3 milyon SNP bulunur. Bir canlının tüm genom dizisini tutmak yerine değişikliklerinin tutulması yerden kazanç sağlayacaktır. Bu mantıkla çeşitli araştırmacılar tarafından farklı amaçlarla birçok SNP verisi oluşturulmuştur. SNP verileri hasta-kontrol guruplarından alınmış örnekler olabileceği gibi, farklı coğrafi bölgelerde yaşayan canlılardan da toplanmış olabilir. Hasta –kontrol guruplarında hastalığa sebep olan biyo-işaretçilerin, coğrafik verilerde ise farklı bölgelerde yaşayan canlıların genetik farklılıklarını gösteren biyo-işaretçilerin keşfedilmesi amacı ile SNP verileri oluşturulabilir. Seçilen biyo-işaretçiler ile hasta-sağlıklı ayrımı için daha hızlı ve insan bağımsız uzman sistemler oluşturarak doğru bir hasta-sağlıklı sınıflandırması yapmak, coğrafik verilerde ise farklı bölgelerde yaşayan canlıların genetik farklılıklarını ortaya koymak bu tarz biyolojik veri analizlerinde örnek hedefler arasında sayılabilir.

Bu veriler tüm genomu temsil etmektedir. Ancak tanımlanan problemi temsil edecek SNP'ler tüm genomda bulanan SNP'lerin çok küçük bir alt kümesidir. Dolayısı ile tanımlanmış problem ile ilişkili SNP'lerin seçilmesi çok önemlidir. Bu çalışma da özellikle bir öznitelik grubundan alt küme seçin yöntemlerin geliştirilmesini hedeflemektedir. Çalışmada önerilen yöntemler çok sayıda öznitelik(SNP) içermesine rağmen çok azının önemli olduğu bu tarz bir veri kümesi ile test edilmiştir.

2.3.2 Mikrodiziler

Bir DNA dizisi olan genler, önce RNA'ya daha sonra proteine dönüşerek canlı hücredeki tüm biyolojik işlevlerde görev alırlar. Biyolojik canlılıktaki işlevlerde herbir genin farklı görevleri vardır. Bir araştırmacı bir işlev için hangi genlerin rol aldığını bulmak isterse o genin ürünü olan proteinlerin hücre içindeki miktarına bakarak bir karar verebilir. Gen ifadelenmesi (gen expresyonu) olarak tanımlanan durum tam da bunun göstergesidir. Bir genin DNA→RNA→Protein dönüşümünün yapılma sıklığı o genin ifadelenme düzeyi olarak anılır. İfadelenme düzeyinin artması (up)bir genin etkinleşmesi, azalması (down) ise o genin baskılanması olarak yorumlanır.

Genlerin ifadelenme düzeyi hücreden hücreye değişebildiği gibi, hücrenin içinde bulunduğu durumdan duruma göre de değişebilir. Örneğin kas hücresinde çok ifadelenen bir gen karaciğer hücresinde hiç ifadelenmiyor olabilir. Diğer taraftan aynı hücrede farklı durumlarda farklı ifadelenen genlerde bulunabilir. Örneğin normalde az ifadelenen bir gen hücre bölünmesi esnasında daha fazla ifadelenebilir. Gen ifadelenme düzeyi canlı vücudu dışındaki çevresel faktörlere göre dedeğişebilir. Örneğin bir bitkide olağan seviyede ifadelelen bir gen, kuraklık zamanında bitkinin çevreye uyumu için daha az yada daha fazla ifadelenebilir.

Genlerin ifadelenme düzeyinin belirlenmesi için mikrodizi teknolojisi kullanılmaktadır. Mikrodizi tekniğinin uygulanması ile aynı anda binlerce genin ifadelenme düzeylerindeki değişikler saptanabilmektedir. Aynı anda birçok geni karşılaştırma imkânı sunması bu tekniğin farklı farklı amaçlar için çokça kullanılmasını sağlamıştır.

Mikrodizi deneyleri çeşitli amaçlar için düzenlenebilir. Örneğin hasta-kontrol gurupları gibi 2 farklı guruptaki canlılara ait genlerin ifade düzeyleri karşılaştırılır. Karşılaştırılan genlerin hasta ve sağlıklı insanları ayırt etme başarısına göre hastalıkla ilgisi olup olmadığına bakılarak hastalığa dair biyo-işaretçiler belirlenebilir.Bu teknik kanserli ve normal hücre arasındaki ifadelenmesi değişmiş genlerin bulunması amacı ile de kullanılabilir. Kanser oluşumunda rol alan genlerin bulunması ile bu genlerin ifadelenmesidüzenleyen etkenler araştırılarak kanserin oluşması engellenebilir. Ayrıca farklı bir kullanımı aynı hücrede farklı zamanlarda alınan gen ifadelenme düzeyleri ile canlının çevresel ortama (örneğin kuraklık, kirlilik) uyumunda ifadelenmesi değişen

genleri bulmaktır. Çevresel koşullara uyumda etkili olan genler üzerinde yapılacak işlemler sonucunda kuraklığa dayanıklı bitkiler yetiştirilebilir.

Oluşturulan deney düzeneği hangi amaçla olursa olsun, ifadelenmesi değerlendirilen binlerce gen içerisinden ancak küçük bir bölümü araştırılan durum ile ilgilidir. Bu sebeple tanımlanan problemle ilişkili genlerin seçilmesi problemin çözülebilmesi açısından çok büyük öneme sahiptir. Ayrıca mikrodizi verisi makina öğrenme yöntemleri ile analiz edilecekse, ilgisiz verilerin varlığı hem analizi güçleştirecek hemde yöntemin başarısını düşürecektir. Makine öğrenme yöntemlerinden öznitelik seçim işlemini geliştirmeye yönelik çok amaçlı yaklaşımların geliştirilmesi hedeflenen bu çalışmada, yüksek sayıda öznitelik taşımasına rağmen çok azının problemle ilişkili olduğu mikrodizi verileri tezin amacı için uygun bir veri olduğundan, çalışmada önerilen metotlar özellikle mikrodizi verileri üzerinde test edilmiştir.

Mikrodizi teknolojisi ile gen ifadelenme verilerinin oluşturulması birçok işlemden oluşmaktadır. Verinin oluşum süreci bu tez kapsamı dışındadır. Dolayısı ile bu çalışmada kullanılan veriler tüm ıslak laboratuvar, görüntü işleme ve kalite değerlendirme süreçlerinden geçmiş analize hazır mikrodizi verileridir. Kullanılan veri herbir satırı bir geni, her birsütunu bir örneği temsil eden bir matris formatındadır.

3. MALZEME VE YÖNTEM

Bu bölümde tez çalışması boyunca kullanılan veri kümeleri ve veri kümelerinin analizinde kullanılan yöntemler hakkında bilgi verilmiştir.

3.1 KULLANILAN VERİ KÜMELERİ

3.1.1Hipertansiyon MikrodiziVeri Kümesi

Hipertansiyon veri seti [36], 22184 gen ve 77'si hipertansiyon hastası ve 82'si kontrol amaçlı olmak üzere 159 erkek bireyden oluşmuştur. Gen skorları (22184 sütun), z-skoruna (sıfır ortalama ve birim varyansı) normalize edilmiştir. Veri kümesiLynn ve arkadaşları[36] tarafından toplanmış, ön işlemleri yapılmış ve genler ifadelenme düzeyleri açısından analiz edilmiştir. Ayrıca öznitelik/gen seçimi için yapay sinir ağları uygulanmış ve SLC4A5, SLC5A10 ve LDOC1 genlerinin tansiyonu yükseltici (yukarı yönde) rol oynarken, BNIP1, APOBEC3F ve LDOC1 genlerinin tansiyonu düşürücü (aşağı yönde) etki yaptığı rapor edilmiştir[36].

Hipertansiyon veri kümesi,bu çalışmada öznitelik seçme işleminin örnek varyasyonuna ve seçim metoduna bağımlılığınınortaya konmasında[3] ve problemin çözümü adına önerilen karmayönteminbaşarısının gösterilmesinde vaka çalışması olarak kullanılmıştır.

3.1.2 Kanser Mikrodizi Veri Kümeleri

Bu çalışmada genel kullanıma açık 3 kanser mikrodizi veri kümesi kullanılmıştır. Hasta ve kontrol gruplarındaki bireylerin gen ifade düzeylerinin ölçülmesi ile oluşturulmuş iki sınıfa sahip mikrodizinler için kısa açıklamalar aşağıda verilmiştir. Veri kümeleri birçok çalışma tarafından [6, 10, 37] yöntem karşılaştırması için kullanılmıştır. Bu veri kümelerinin analizinde temel amaç hasta-kontrol grubu arasındaki gen ifade düzeylerinin kullanımı ile hasta-kontrol grubunu birbirinden ayırabilecek en önemli genlerin seçilmesidir. Genler hastalıkta attırıcı yada azaltıcı yönde etkide bulunabilir.

i. Kolon kanseri [38] : Veri kümesi, 62 doku örneği (40'ı tümörlü, 22'si normal kolon dokusu olmak üzere) ve en yüksek minimal yoğunluklu 2000 gen örneği içermektedir.

ii. Duke Meme Kanseri [39] : Veri kümesi 44 örnek ve 7129 gen içermektedir. Örnekler ER+ veya ER- şeklinde sınıflandırılmıştır (ER: estrogenreceptor).

iii. DLBCL [40] : Veri kümesi77 örnek ve 5469 gen ifadesi içermektedir. 58 örnek ilk sınıftan, 19 örnek ikinci sınıftan alınmıştır.

3.1.3 HGDP Veri Kümesi

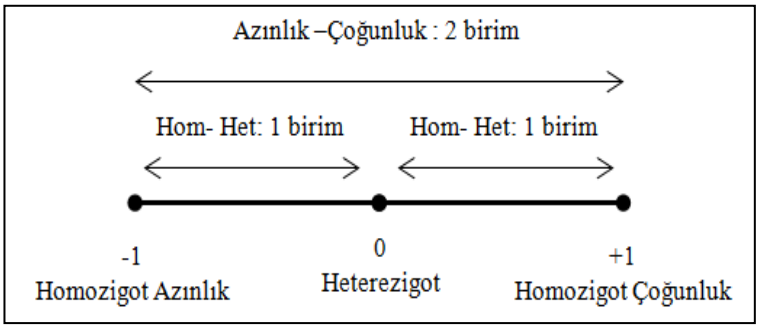
İnsan genomu çeşitliliği (Human Genome Diversity Project - HGDP) projesi farklı milletlerden insanlar üzerindeki evrim tabanlı varyansı tanımlamayı ve popülasyon genetiğinde süregelen mutasyonu izlemeyi amaçlar. Proje, Stanford Üniversitesi'nden Luigi Luca Cavalli-Sforza tarafından yönetilmiş ve dünya çapında pek çok araştırmacı/bağışçı projeye katkıda bulunmuştur [41].Saf veri 52 etnik gruptan seçilen 1064bireyin 660918 SNP verisini içerir. 21 bireyin verisindeki eksiklik dolayısı ile bu çalışma 1043 bireylik veri kümesi kullanılmıştır. İlk 163 SNP mitokondriyal DNA'lardan elde edilmiştir ve bu çalışmanın dışında tutulmuştur. Bireylerin etnik gruplara dağılımı Tablo 3.1'de görülmektedir.

Tablo 3.1:HGDP Kümesindeki Bireylerin Etnik Gruplara Dağılımı

G. No	Etnik Grup	Birey Sayısı	G. No	Etnik Grup	Birey Sayısı	G. No	Etnik Grup	Birey Sayısı
G1	Adygei	17	G18	Japanese	29	G35	Papuan	17
G2	Balochi	25	G19	Kalash	25	G36	Pathan	23
G3	Bantu	20	G20	Karitiana	24	G37	Pima	25
G4	Bedouin	48	G21	Lahu	10	G38	Russian	25
G5	Biaka Pygmies	32	G22	Makrani	25	G39	San	6
G6	Brahui	25	G23	Mandenka	24	G40	Sardinian	28
G7	Burusho	25	G24	Maya	25	G41	She	10
G8	Cambodian	11	G25	Mbuti Pygmies	15	G42	Sindhi	25
G9	Colombian	13	G26	Melanesian	19	G43	Surui	21
G10	Dai	10	G27	Miaozu	10	G44	Tu	10
G11	Daur	9	G28	Mongola	10	G45	Tujia	10
G12	Druze	47	G29	Mozabite	30	G46	Tuscan	8
G13	French	29	G30	Naxi	9	G47	Uygur	10
G14	French Basque	24	G31	North Italian	13	G48	Xibo	9
G15	Han	44	G32	Orcadian	16	G49	Yakut	25
G16	Hazara	24	G33	Oroqen	10	G50	Yizu	10
G17	Hezhen	9	G34	Palestinian	51	G51	Yoruba	24

Ham veri de herbir bireyin herbir SNP değeri biri babadan diğeri anneden gelen 2 adet alel bilgisi içermektedir.Bir pozisyondaki 2 alel bireyin o noktadaki genotip bilgisini oluşturur. Birey anne ve babadan farklı alel miras almış ise o pozisyonda birey heterezigot, aynı aleli almış ise homozigot olarak adlandırılır. Genom boyu ilişkilendirme çalışmalarında (Genome Wide Assosiation Study- GWAS), toplumda daha çok görünen alel çoğunluk, daha az görünen alel ise azınlık alel olarak isimlendirilir. Örneğin bir pozisyonda C ve T alelleri görüldüğünü; C alelinin çoğunluk, T alelinin ise azınlık olduğunu varsayalım. Herbir birey için bir snp pozisyonunda heterezigot-CT, homozigot çoğunluk-CC, homozigot azınlık-TT olmak üzere 3 farklı durum görülebilir.A,C,G,T nükletidlerinden oluşmuş genotip bilgisi içeren HGDP ham verisi şu şekilde sayısallaştırılmıştır: (i) Heterezigot-CT→0, (ii) Homozigot çoğunluk-CC→+1, (iii) homozigot azınlık-TT→-1.

Sayısallaştırılmış genotip bilgisi üzerinden, bireyler arası genomik mesafe hesaplama işlemi **Şekil 3.1**'de gösterilmiştir. Bu durumda heterezigot bir birey her iki homozigot genotipe 1 birim uzaklıkta, homozigot azınlık ve çoğunluk alele sahip bireyler birbirine 2 birim uzaklıktadırlar.



Şekil 3.1: Genotip Bilgisinin Sayısallaştırılması Ve Genomik Mesafeler

Bilindiği gibi, yüksek boyutlu veri daha çok bellek alanı ve daha az işlem hızı anlamına gelmektedir. Dolayısıyla HGDP veri kümesi (1043 x 660755) gibi çok yüksek boyutlu verilerin tek seferde işlenmesi yüksek fiziksel donanım ve zaman gerektirir ki bu durumda verinin tek seferde işlenmesi kısıtlı bellek alanına sahip bir çok bilgisayar için mümkün olamamaktadır. Donanımsal kısıtlar aşılsa bile yüksek boyutta işlem yapmak daha uzun zaman gerektirecektir. Bu problemleri çözmek için veri alt kümelere ayrılmıştır. En doğal ayırma işlemi ise genom verisinin kromozomlara bölünmesidir. Veri 23 kromozoma ayrılıp, SNP'ler kromozomdaki pozisyonlarına göre sıralanmıştır.

Tablo 3.2'de verinin kromozomlara ayrılması ile SNP'lerin kromozom başına düşen sayıları görülmektedir**.**

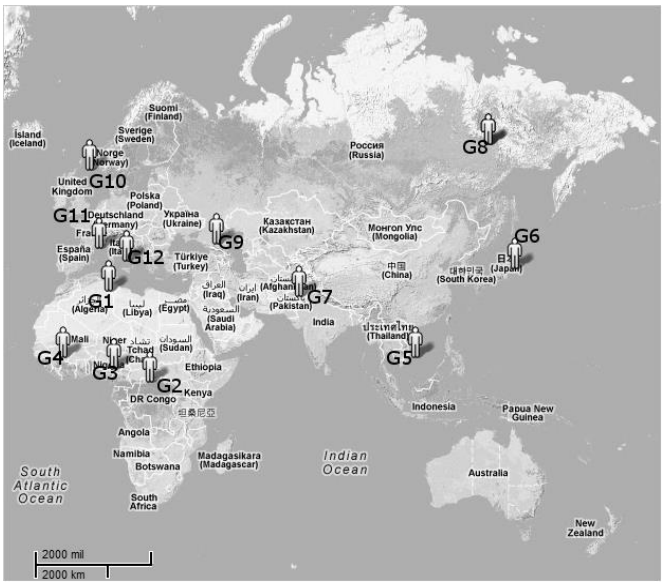
Tablo 3.2: Her Bir Kromozomdaki Snp Sayısı

Kromozom No	SNP Sayısı	Kromozom No	SNP Sayısı
1	49.639	13	25.191
2	53.765	14	21.45
3	44.564	15	19.594
4	39.942	16	19.727
5	40.976	17	16.629
6	43.239	18	20.165
7	35.507	19	10.739
8	37.282	20	16.911
9	31.192	21	9.645
10	34.493	22	9.73
11	32.005	X	16.472
12	31.873	Y	10

Jeo - genomik mesafe ilişkisini incelemek için 51 etnik grup arasında Fransız Bask grubu gibi genetik mirasını soy içi evliklerle korumuş, genetik olarak başka ırklarla az karışmış gruplar tercih edilmiştir. Seçilen etnik gruplar, grupların bulundukları kıtalar, birey sayıları ve coğrafik mesafe hesaplamada kullanılan koordinatları **Tablo 3.3** özetlenmiştir. **Şekil 3.2**ise seçilen etnik grupların dünya üzerindeki dağılımını göstermektedir.

Tablo 3.3: Seçilen Alt Gruplar Ve Coğrafi Koordinatları

Kıta	Etnik Grup		Koordinatlar	Birey Sayısı
Africa	G1	Mozabite	32N, 3E	30
	G2	Biaka_Pygmies	4N, 17E	11
	G3	Yoruba	6-10N, 2-8E	24
	G4	Mandenka	12N, 12W	24
Asia	G5	Cambodia	12N, 105E	11
	G6	Japanese	38N, 138E	29
	G7	Balochi	30-31N, 66-67E	25
	G8	Yakut	62-64N,129-130E	25
Europe	G9	Adygei	44N, 39E	17
	G10	Orcadian	59N, 3W	16
	G11	French_Basque	43N, 0	24
	G12	Sardinian	40N, 9E	28



Şekil 3.2: Seçilen Grupların Coğrafî Dağılımı

Alt küme seçiminden sonra bazı SNP'lerde tüm alt gruplar için alel çeşitliliğinin az olduğu görülmüştür. Bu SNP'lerin grupları ayırd edici bilgi içermemesi nedeni ile veri kümesinden elenmesi gerekmektedir. Bu filtreleme için "Minor Allele Frequency" kullanılmıştır. Bu hesab için herbir alelin tüm bireylerde görülme sayısı hesaplanır. Alel sayısınn toplumdaki birey sayının iki katına (her birey biri anneden diğeri babadan olmak üzere iki alel taşır) bölümü o alelin sıkılığını gösterir. Örneğin A ve T alelinin görüldüğü bir SNP için, eğer azınlık alel T'nin görülme sıklığı %5'ten az ise bu SNP ayrıt edici bilgi içermediği düşünülerek veri kümesinden çıkarılmıştır.

Bu çalışmada seçilen grupların birbiri arasındaki coğrafi uzaklık ile genomik uzaklığın ilişkisi incelenmiştir. Seçilen grupların **Tablo 3.3** de gösterilen koordinatları kullanılarak gruplar arasındaki kuş uçuşu coğrafi uzaklıklar vincenty formulu [42] ile hesaplanmışır Gruplar arası genomik mesafe ise bir gruptaki tüm bireylerin diğer gruptaki tüm bireylere olan uzaklıklarının(İki birey arasındaki genomik mesafe **Şekil 3.1**'de anlatılmıştır.) ortalaması olarak hesaplanmıştır. Bir grubun, grup içi uzaklığının diğer tüm gruplara olan uzaklığından daha az olması öngörülmüştür. Bu nedenle genomik mesafe hesaplanırken grup içi ortalama mesafe de hesaplanmıştır

Seçilen etnik gruplar arası kuş uçuşu coğrafi mesafe (kilometre) ve 21. kromozom için genomik mesafe sırasıyla **Tablo 3.4** ve **Tablo 3.5**'te verilmiştir.**Tablo 3.5**'da görüleceği üzere, öngörüldüğü gibi, grup içi ortalama genomik mesafeler, diğer tüm gruplara olan

genomik mesafeden daima düşüktür.Matris formunda hesaplanan coğrafi ve genomik mesafe verileri arasındaki korelasyon%52'dir. Çalışmanın temel amaçlarına uygun olarak bu korelasyonun arttırılmasına yönelik çok amaçlı yaklaşımlar uygulunmış ve başarılı sonuçlar alınmıştır. Kullanılan yöntemler Bölüm 3.2'de açıklanmış, sonuçlar Bölüm 4'de sunulmuştur.

Tablo 3.4: Seçilen Etnik Grupların İkili Coğrafi Mesafeleri

G. No	1	2	3	4	5	6	7	8	9	10	11	12
1	0	3427	2665	2696	10418	10965	5960	8462	3397	3034	1249	1039
2	3427	0	1400	3314	9709	12379	5958	10721	4930	6340	4636	4068
3	2665	1400	0	1915	10911	12971	6839	10850	5173	5699	3909	3567
4	2696	3314	1915	0	12583	13654	8252	11058	6009	5270	3626	3726
5	10418	9709	10911	12583	0	4361	4459	5985	7176	9873	10296	9666
6	10965	12379	12971	13654	4361	0	6478	2840	7824	8668	10085	9930
7	5960	5958	6839	8252	4459	6478	0	5676	2842	5994	5941	5247
8	8462	10721	10850	11058	5985	2840	5676	0	5795	5883	7432	7433
9	3397	4930	5173	6009	7176	7824	2842	5795	0	3283	3127	2511
10	3034	6340	5699	5270	9873	8668	5994	5883	3283	0	1792	2277
11	1249	4636	3909	3626	10296	10085	5941	7432	3127	1792	0	821
12	1039	4068	3567	3726	9666	9930	5247	7433	2511	2277	821	0

Tablo 3.5: Seçilen Etnik Grupların Ortalama İkili Genomik Mesafeleri

G. No	1	2	3	4	5	6	7	8	9	10	11	12
1	76.01	85.03	83.77	83.79	83.81	84.89	82.47	84.38	80.88	80.86	81.00	81.01
2	85.03	66.29	77.28	77.86	89.59	90.97	90.37	91.34	89.52	90.46	90.31	90.70
3	83.77	77.28	71.92	76.41	88.52	89.78	89.17	89.80	88.37	89.19	89.27	89.85
4	83.79	77.86	76.41	72.77	88.46	89.70	89.19	89.63	88.20	89.01	89.09	89.57
5	83.81	89.59	88.52	88.46	67.33	76.09	82.90	77.23	81.95	83.11	83.60	83.95
6	84.89	90.97	89.78	89.70	76.09	71.51	83.23	76.17	82.35	83.54	84.09	84.51
7	82.47	90.37	89.17	89.19	82.90	83.23	77.19	82.58	79.71	80.01	80.65	80.92
8	84.38	91.34	89.80	89.63	77.23	76.17	82.58	70.81	81.38	82.17	82.84	83.65
9	80.88	89.52	88.37	88.20	81.95	82.35	79.71	81.38	72.87	78.08	78.45	78.65
10	80.86	90.46	89.19	89.01	83.11	83.54	80.01	82.17	78.08	71.37	77.60	77.95
11	81.00	90.31	89.27	89.09	83.60	84.09	80.65	82.84	78.45	77.60	73.92	78.04
12	81.01	90.70	89.85	89.57	83.95	84.51	80.92	83.65	78.65	77.95	78.04	74.41

3.2 VERİ KÜMESİ BÖLÜMLEME YÖNTEMLERİ

Veri setini sürekli olarak parçalara bölen ve her adımda bir katı test için dışarıda bırakan, LOO ve k-kat çapraz geçerleme yöntemlerini kullandık. LOO, her katı yalnızca bir örnek içeren, k-katlı çapraz geçerlemenin istatistiksel olarak daha güvenilir bir versiyonudur.

Makina öğrenme yöntemleri ile oluşturulmuş bir kestirim modeli öğrenme ve test olmak üzere en az 2 örneklem kümesine ihtiyaç duyar. Öğrenme kümesi modelin eğitilmesi amacı ile kullanılırken, test kümesi modelin başarısını değerlendirmek için kullanılır.

Bir model değerlendirilirken, kestirdiği değerin gerçek değerinden ne kadar farklı olduğuna bakılır. Modelin hatası, öğrenme kümesindeki hata (yanlılık) ve test kümesindeki hata (varyans) değerleri ile belirlenir.

Yanlılık, örneklemlerdeki değişmeleri dikkate almayarak bir kestirim modelinin ne kadar yanlış olduğunu ölçer. Bir kestiricinin yanlılığı, beklenen değerin(ortalamasının) gerçek değerden ne kadar farklı olduğudur. Varyans, bir kestirim modelinin farklı örneklemlerde, kestirilen değerin beklenen değer çevresinde ne kadar değiştiğini ölçer. Bu model ve öğrenme kümesine bağlıdır.

Bir örneklem için bir kestirim modeli veriye çok iyi oturmuşken- yanlılık düşük-, bir başka örneklem için kötü kestirim yapabilir –varyans yüksek-. Yani algoritma evrensel veriden seçilen öğrenme örneklemine yüksek uyumluluk gösterirken, evrensel veriyi tam olarak temsil edememektedir. Bu durum kestirim modelinin güvenilmezliğini gösterir. Sonuç olarak hem yanlılık hem varyans düşük olduğunda kestirim modelinin başarılı olduğunu söyle biliriz.

Birçok çalışma modelin başarısı için test örneklemi üzerindeki başarısıyı temel almaktadır. Ancak bazı durumlarda veri kümesinden yanlı bir test örneklemi oluşabilir. Örneğin uç noktalardaki durumları içeren örnekler öğrenme kümesinde hiç bulunmazken test kümesinde olabilir. Bu durumda öğrenme kümesinin evrensel veriyi eksik temsilinden dolayı düşük test başarısı elde edilir. Tam tersi durumda, test kümesinde çok sıradan örnekler yer alabilir, uç durumlar bulunmaz ve yüksek geçerleme başarısı elde edilir. Dolayısı ile veri kümesinin bölümlenmesi modelin başarımını etkilemektedir.

Veri kümesi bölümlemenin model başarımına olduğu kadar, öznitelik seçimine de etkisi büyüktür. Uç örneklerin olmadığı örneklemlerde seçilen öznitelikler evrensel uzayı temsilde başarısız kalacaktır. Seçilmiş öznitelikler temel alınarak yapılacak ileri çalışmalarda, temsil gücü düşük öznitelikler zaman ve maliyet kaybına sebep olur.

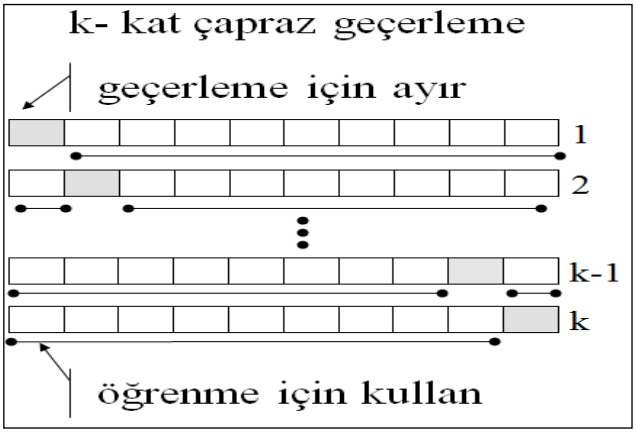
Dolayısı ile yanlı veri bölümleme problemini ortadan kaldırmak için veri seti bölümlemede birçok yol izlenebilir. Aşağıda veri seti bölümlemesi için bazı yöntemler anlatılmıştır.

3.2.1 Rassal Bölümleme

Veri kümesi eşit 2 parçaya ya da yüzdesel olarak 2 parçaya (%60 öğrenme, %40 geçerleme) ayrılır. Rassal olarak ayrılacak bu bölümlemede yanlı ayrımlar oluşabilir. Bu durumda test verisindeki örneklerin durumuna göre model başarısı ve seçilen öznitelikler değişir. Bu rassallığı ortadan kaldırmak için bölümleme, modeli eğitme ve test işlemleri birkaç kez tekrar edilir. Tekrarlar neticesine elde edilen başarının ortalamasına ve standart sapmasına bakılarak model başarımı ölçülür.

3.2.2 K-Kat Çapraz Geçerleme (k-fold cross-validation)

Veri seti bölümlemede oluşacak yanlılık, işlemlerin birkaç kez tekrar edilip ortalaması alınarak giderilmeye çalışılsa da, işlemdeki rassallık nedeni ile yine verideki uç noktaların, gizli-açık grupların öğrenme ve geçerleme kümesinde yer alması garanti edilemez. Bu duruma çözüm olarak veri seti k parçaya bölünür. Bir parça geçerleme, k-1 parça öğrenme olacak şekilde, eğitme ve geçerleme adımları k kez tekrarlanır. Bu şekilde her bir parça geçerleme kümesi olana kadar işlem devam eder. Verideki her bir örnek sadece bir kez geçerleme kümesinde bulunacağından tüm uç noktalar dikkate alınmış olur. Model başarımı için k tekrardan alınan başarının ortalaması ve standart sapması dikkate alınır. **Şekil 3.3**'de k-kat çapraz geçerleme (ÇG) için işlem adımları gösterilmiştir.



Şekil 3.3: K-Kat Çapraz Geçerleme

K-kat ÇG, her örneğin geçerleme de yer almasını garantilemiş olsa da bölümleme işlemindeki rassallık verideki gizli-açık grupların her öğrenme ve geçerleme işleminde temsil edilebileceğini garantilemez. Nitekim farklı k-kat bölümlemeler ile farklı başarım

sonuçları elde edilir. Bu eksikliği ortadan kaldırmak için yine k-kat ÇG işlemi birkaç kez tekrarlanır ve ortalamalar alınarak model değerlendirilir.

k-kat ÇG, etkin nitelik seçimi için de yanlılığı ortadan kaldıracak bir çözümdür. Tüm örneklerin dikkate alındığı seçme işlemlerinin kesişimi, daha tutarlı etkin nitelik kümeleri seçecektir.

3.2.3 Birini Dışarıda Bırak Çapraz Geçerleme

K-kat çapraz geçerlemenin özel bir hali ise gerçekleme için sadece bir örneğin ayrıldığı, N-1 örneğin ise öğrenme için kullanıldığı birini dışarıda bırak (leave one out -LOO) durumudur. Öğrenme ve geçerleme işlemleri verideki tüm örnek sayısı kadar devam eder. Bu yöntemde rassal bölümlemenin önüne geçildiğinden özellikle 2 modelin karşılaştırılmasında en uygun yoldur. Ayrıca bu ayrımın, az sayıda örneğe sahip veri setlerinde öğrenme kümesinin azlığından başarısız olunacak durumlarda kullanılması uygundur.

3.3 ÖZNİTELİK SEÇİM VE ÇIKARIM YÖNTEMLERİ

Biyolojik veriler gibi çok yüksek boyutlu veriler çok sayıda öznitelikten oluşmuşlardır. Ancak bu özniteliklerin çok azı veri kümesi ile ilişkilendirilmiş problem ile alakalıdır. Bu nedenle yüksek boyutlu verilerde ilk işlem boyut azaltılmasıdır. Bu işlem tanımlanmış problemi çözebilecek özniteliklerin seçilmesiile yapılabileceği gibi var olan özniteliklerden bir takım işlemler sonucunda yeni özniteliklerin çıkarımı şeklinde de olabilir. Bu bölümde tek kriterliöznitelik seçim ve çıkarım yöntemleri anlatılcaktır.

3.3.1 Karşılıklı Bilgi -MI

Karşılıklı bilgi, X ve Y gibi rastgele değişkenlerin bağımlılığını değerlendiren bir ölçüm yöntemidir [37]. MI, I(X;Y) ortak olasılık fonksiyonu ve bu dağılımların marjinal olasılıklarının çarpımı arasındaki bağıl entropi olarak bilinir (3.1).

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \tag{3.1}$$

Denklem 3.1 aşağıdaki 3.2-3.5 denklemlerinde görüldüğü gibi genişletildiğinde ayrık ve koşulsal entropi arasındaki ilişki açığa çıkarılabilir.

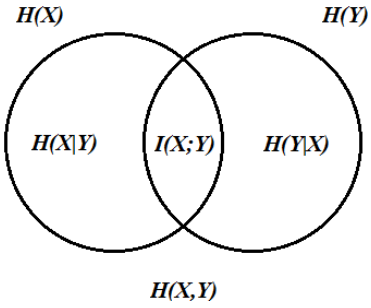
$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x \mid y)}{p(x)} \tag{3.2}$$

$$I(X;Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x \mid y) \tag{3.3}$$

$$I(X;Y) = -\sum_{x \in X} p(x) \log p(x) - \left(-\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x \mid y) \right) \tag{3.4}$$

$$I(X;Y) = H(X) - H(X \mid Y) \tag{3.5}$$

Denklem 3.5, X'in bilinmezliği ie Y biliniyorken X'in bilinmezliğinin farkıolarak düşünülebilir.Burada, H, rastgele değişkenler için entropiyi ifade eder. Bu ilişki **Şekil 3.4**'de gösterilmiştir.



Şekil 3.4: X Ve Y'nin Arasındaki Karışılıklı Bilginin Basit Bir Gösterimi

Şekil 3.4'de gösterildiği üzere, MI, Y bilindiği takdirde, X'ten dolayı oluşan miktarı ölçer, aynı şey, tersi durum için de sözkonusudur. Bir özelliğin aldığı değerler arasındaki ortak bilgi ve karşılık gelen sınıf etiketleri, özelliğin ayırt edici derecesini ölçmek için kullanılabilir.

Öznitelik seçim işleminin veri kümesindeki varyasyondan ve seçim metodundan etkilendiğini ortaya koymak amacı ile yapılan çalışmamızda [3], MI temelli iki farklı öznitelik seçim yöntemi önerilmiştir.

3.3.2 MI-d

MI-d (MI-discrete, MI-ayrık) tamamıyla, özniteliklerin z-skorlarının yuvarlanmış ayrık değerleri kullanılarak hesaplanan karşılıklı bilgidir. Bu yalın MI skorlarından elde edilen karşılıklı bilgiyi kullanılırken seçilen öznitelikler ile sınıf etiketi arasındaki ortak

bağımlılık (joint dependency) hesaba katılmamıştır (yüksek karşılıklı bilgi skoruna sahip değişkenler mutlak anlamda birbirlerini tamamlıyor olmasalar dahi en iyi 10 öznitelik olarak seçilebilirler. Esasında tamamen birbirlerinin kopyası bile olabilirler).

3.3.3 MI-c

MI-c (MI-clustered, MI-kümeli) yöntemi, her öznitelik skorunu kümeler halinde ayrıklaştırmayı amaçlayan bir yöntemdir [43]. Öznitelik gruplarının sınıf etiketleri olan ortak karşılıklı bilgilerini bulmak için, ileri doğru öznitelik seçim (forward selection) yöntemi olarakMI-c'yi kullanabiliriz. Çok değişkenli öznitelik altkümelerinin ayrıklaştırılması ve ortak karşılıklı bilginin hesaplanabilmesi için bir kümeleme algoritması kullanılabilir [44]. Buna göre, öncelikle, her öznitelik (örneğin N değere sahip her sütun) K-kümeye bölünür. Burada K, özniteliklerin ayrık etiketlerinin sayısını ifade eder. Bu tarz ayrık etiketli öznitelikler ve sınıf etiket bilgisi arasındaki MI değeri hesaplanır ve en yüksek skorlu gen seçilir. İkinci özniteliği seçmek için, ilk adımda seçilen öznitelik ve diğer her bir öznitelik, birer birer birleştirilir ve kümelenir. Küme indislerinden (her örnek için bir küme ID'si) oluşan bu dizi ile karşılık gelen sınıf etiket bilgisi arasındaki MI değerleri hesaplanır ve en yüksek değere sahip çift en iyi 2 öznitelik olarak seçilir. Ardından kümeler 3'lüler haline gelir, her 3'lü kümenin indisleri hesaplanır ve sınıf etiketleri ile karşılıklı bilgiyi hesaplamak için kullanılır ve bu işlem 4'lü, 5'li vs şeklinde devam eder.

3.3.4 Korelasyon- CD

Her öznitelik ve sınıf etiket bilgisi arasındaki karesel korelasyon anlamına gelen kararlılık katsayısıdır (coefficient of determination-CD). CD, lineer bağlılığın bir ölçüsüdür. İki sürekli değişken (öznitelik ve sınıf etiketi) arasındaki ilişki korelasyonu, ilişki miktarı ise korelasyon katsayısını (r) göstermektedir. İki değişken arasındaki korelasyon katsayısı -1 ile +1 arasında değerler alırken katsayının işareti ilişkinin yönünü belirler. Aynı yönde değişim gösteren iki değişken arasında pozitif korelasyon, farklı yönde değişim gösteren iki değişken arasında ise negatif korelasyon görülür. Burada önemli olan ilişkinin yönü değil miktarı ise kare yada mutlak değer alınabilir. Korelasyon katsayısı aşağıdaki formül ile hesaplanır:

$$r = \frac{C_{xy}}{C_x C_y} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})^T}{\sqrt{(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})^2)(\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})^2)}} \quad (3.6)$$

3.3.5 mRMR

En küçük artıklık en Büyük ilgililik (minimum Redundancy Maximum Relevance,mRMR) [37] yöntemi, bir hedef sınıflandırma değişkeninin istatistiksel özelliğini en iyi karakterize eden bir öznitelik altkümesi seçimini hedefler. Bu öznitelik alt kümesi seçimi, özniteliklerin mümkün mertebe karşılıklı olarak birbirine benzememesi (redundancy) fakat sınıf etiketiyle mümkün mertebe benzerlik göstermesi (relevance) şeklindeki bir kısıta tâbidir. Bu iki terim, ikili tek değişkenli karşılıklı bilgi (bir öznitelik ile sınıf etiketi arasında ya da bir öznitelik ile diğeri arasında) kullanılarak hesaplanır. Her adımdaki seçim, iki terim arasındaki farkı iyilik skoru olarak kullanır. Her adımda tüm öznitelikleri deneyerek en iyi kümeyi seçmeyi hedefleyen bu metotbir ileri yönlü öznitelik seçim metodudur.

Bu yöntem,öznitelik seçim yöntemlerinin veri kümesindeki varyasyondan ve seçim metodundan etkilendiğini ortaya koymak amacı ile yapılan çalışmamızda [3], öznitelik seçim yöntemi olarak kullanılmıştır.

3.3.6 Relief

Öznitelik seçimi için çokça kullanılan Relief [45] algoritması, öznitelikleri farklı sınıflardaki komşu örnekleri ayırma başarısına göre ağırlıklandırır. Bu amaçla Relief, verilen bir örnek için aynı sınıftan en yakın bir komşusunu (*nearest hit* olarak isimlendirilir) ve diğer sınıftan en yakın bir komşusunu (*nearest miss* olarak isimlendirilir) arar. Relief algoritmasına göre, A özniteliğinin ağırlığı (W[A]) aşağıda verilen olasılıkların farkı ile belirlenir:

W[A] = P(A'nın farklı değeri | farklı sınıf)– P(A'nın farklı değeri | aynı sınıf)

Temel mantık, iyi bir öznitelik farklı sınıflardan gelen örnekleri ayırd edebilmeli ve aynı sınıftan gelen örnekler için aynı değere sahip olmalıdır[46]. Yukarıdaki ağırlıklandırmaya göre, bir öznitelik, farklı sınıftan iki en yakın komşu farklı değere sahip ise + puanla ödüllendirilirken, aynı sınıftan iki en yakın komşu farklı değerlere sahipse – puan ile cezalandırılır.

Relief iki sınıflı problemler için uygundur ve en yakın bir komşuya bakarak karar verir. Relief algoritasının temel çalışma prensibi sözde kod olarak şu şekildedir:

Girdi:Herbir örnek için özniteliklerin ve sınıf bilgisinin bulunduğu vektör
Çıktı: özniteliklerin ağırlıklarını gösteren W vektörü

1. *tüm ağılıklara sıfır ata W[A]=0*
2. *for i=1:m*
3. *R_i 'ye rastgele bir örnek ata*
4. *R_i ilefarklı (miss-M) ve aynı(hit-H) sınıftan en yakın komşularını bul*
5. *for A=1:a*
6. *W[A]=W[A]-diff(A, R_i ,H)/m + diff(A, R_i ,M)/m*
7. *end*
8. *end*

$diff(A, I_1, I_2)$fonksiyonu, öznitelik A'nın sayısal yada kategorik veri içermesine göre farklı formulasyona sahiptir. Kategorik veriler için:

$$diff(A, I_1, I_2) = \begin{cases} 0; değer(A, I_1) = değer(A, I_2) \\ 1; otherwise \end{cases} \quad (3.7)$$

ve sayısal veriler için:

$$diff(A, I_1, I_2) = \frac{|değer(A,I_1) - değer(A,I_2)|}{\max(A) - \min(A)} \quad (3.8)$$

şeklinde tanımlanmıştır.

3.3.7Temel Bileşen Analizi

Temel Bileşen Analizinin ana kullanım amacı tüm öznitelikleri ağırlıklandırarak, veri kümesini farklı bir uzaya çeviren yeni öznitelikler çıkarmaktır. Yöntem veri kümesindeki varyansı (değişinti) koruyarak, veriyi yeni bir boyuta taşıyabilen verinin temel bileşenlerini bulmayı amaçlar. Burada, temel bileşenlerin oluşturulmasındaki amaç, verideki varyansı en iyi temsil eden uzayın bulunmasıdır. TBA ile verinin yeni uzaya çevrilmesi sonucunda verideki gizli örüntülerin ortaya çıkarılması sağlanabilir. Novembre ve diğ.[47], ilk iki temel bileşenin, tüm Avrupa'daki milletlerin genetik yerleşimini gösteren bir izdüşüm oluşturmak için kullanılabileceğini göstermiştir.

TBA, her ne kadar boyut azaltmak için kullanılsa da özniteliklerin seçimi içinde kullanılabilir. Analiz sonucunda ortaya çıkan temel bileşenler herbir öznitelik için bir

ağırlığa sahiptir. Öznitelik seçim yöntemlerinin de ilk işi zaten öznitelikleri belli bir amaca göre (örneğin sınıf bilgisi ile olan korelasyon, karşılıklı bilgi vs.) ağırlıklandırmaktır. Bu ağırlıklar yardımıyla öznitelikler derecelendirilmiş olur.Bu amaçla öznitelikler herbir temel bileşende farklı ağırlıklandırılarak özniteliklerin yeni boyutlara katkıları hesaplanır. Yeni boyuta katkısı çok olan öznitelik en değerli olandır şeklinde düşünülerek öznitelik seçim işlemi uygulanabilir.Çalışma [48] TBA ile seçilmiş az sayıda SNP'in toplumların ayırd edilmesinde başarılı olarak kullanılabileceğini göstermiştir.

Boyut azaltma için kullanılan TBA gibi izdüşüm yöntemlerinin temel amacı *d* boyutlu uzayda tanımlanan veriden en az bilgi kaybı ile $k<d$ boyutluyeni bir uzay tanımlamaktır. İşlem aslında aralarında korelasyon olan çok sayıdaki değişkenden oluşan veriyi, aralarında korelasyon bulunmayan az sayıda değişkene indirgemektir. Eğer verideki değişkenler birbiri ile korelasyonları yüksek ise *k, d* değerinden oldukça düşük olacaktır. Ancak boyutlar arasındaki korelasyon düşük ise *k, d* kadar büyük olacaktır [4].TBA sınıf bilgisi kullanmayan gözetimsiz bir öğrenme algoritması olduğundan ölçüt olarak verideki değişintiyi kullanır.Yani elde edilmek istenen, verideki değişintiyi en iyi koruyan *k* boyutlu yeni bir uzaydır.Yeni uzaydaki herbir boyut, temel bileşenler olarak adlandırılır. Temel bileşenler birbiri ile korelasyonları yoktur.

Problemin Tanımı:

$X=[x_1,x_2,...,x_N]$, *U* uzayında,*d* boyutlu ve N örnek içeren bir veri kümesi olsun.*z*, *x*'in yine *d* boyutlu *U'* uzayına izdüşürülmüş halini göstersin.*w*, *x*'i *z*'ye dönüştüren vektörü göstersin.*x* 'in *w* yönüne izdüşümü iç çarpım olarak aşağıdaki şekilde yazılabilir:

$$z = w^T x \tag{3.9}$$

w_1, verideki değişintiyi (varyans) koruyan en iyi vektördür ve verinin birinci temel bileşenidir. Birinci temel bileşen yönüne tüm örneklerin izdüşümü aşağıdaki gibi yazılır:

$$z_1 = w_1^T x \tag{3.10}$$

Gözlenen tüm değerlerin ortalamadan uzaklıklarının karelerinin ortalaması toplam varyansı verir. Yapılan işlem aşağıdaki denklemde gösterilmiştir.

$$Var(z_1) = w_1^T C w_1 \tag{3.11}$$

Burada *C*, *dxd* boyutlu kovaryans matrisi olup

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(x_i - \bar{X})^T \tag{3.12}$$

şeklinde tanımlıdır.

Buradaki temel işlem, izdüşüm sonrası, verideki toplam varyansı en büyük kılmak için en uygun $w_1$vektörününün yönünün bulunmasıdır. Bu durum bir optimizasyon problemine dönüşmüştür. Toplam varyans ($w_1^T C w_1$) ifadesininw_1'e göre türevinin alınması problemin çözümü için yeterlidir (Bishop, 2006). Burada baz vektörünün boyu değil yönü önemli olduğundan, vektörün uzunluğu (normu$\|w_1\| = 1$) 1 olmalı koşulu vardır.w_1'in orijine göre uzaklığı (uzunluğu),

$$uzaklık = \sum_{i=1}^{d}\left(w_1^i - 0\right)^2 = w_1^T w_1 = 1 \tag{3.13}$$

şeklinde bulunur. λ_1 Lagrange çarpanınen iyileme problemimize eklenmesi ile kısıtı denklemde ifade etmiş oluruz:

$$w_1^T C w_1 + \lambda_1(1 - w_1^T w_1) \tag{3.14}$$

En iyileme problemimizin çözümü yukarıdaki ifadeninw_1'e göre türevinin sıfıra eşitlenmesidir. Bu işlem sonunda aşağıdaki eşitlik elde edilir:

$$C w_1 = \lambda_1 w_1 \tag{3.15}$$

Elde edilen ifade özdeğer-özvektör problemidir. *C* kovaryans matrisinin en büyük değerli özdeğere karşılık gelen özvektör, veriyi yeni uzaya izdüşürülürken en yüksek varyansın sağlanacağı yönü gösteren baz vektörüdür. Bu şekilde *d* adet özvektör hesaplanır. Herbir özvektör verinin temel bileşenleridir.

Temel bileşenler toplam değişintiyi temsil etmedeki başarılarına göre sıralandığında, verideki değişintiyi en iyi açıklayan bileşen birinci temel bileşendir. Herbir temel bileşenin değişintiye katkısı λ kadardır. Bu durumda *k* temel bileşen ile temsil edilen yeni uzayın verideki değişintiyi açıklama oranı aşağıdaki gibi formülize edilebilir. Bu açıdan bakıldığında, belirli sayıda temel bileşen yerine değişintinin %90'nını temsil eden *k* tane bileşende alınabilir.

$$\frac{\lambda_1 + \lambda_1 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_k + \ldots \lambda_d} \tag{3.16}$$

Seçilen k tane temel bileşen ($w^T = \{w_1, w_2, \ldots, w_k\}$) ile verinin yeni uzaya izdüşümü alınır. Bunun için verinin önce merkezlenmesi yani tüm örneklerden ortalamanın çıkarılması gerekir. Daha sonra merkezlenmiş veri tüm baz vektörler ile iç çarpıma tabi tutularak yeni uzaya dönüştürülür (*m, X* verisin ortalamasını ifade eder)

$$z = w^T(x - m) \tag{3.17}$$

3.4 ÇOK KRİTERLİ ÖZNİTELİK DERECELENDİRME VE SEÇİM YÖNTEMLERİ

Öznitelikler tek bir amaç gözetilerek derecelendirilebilir ve bunun sonucunda oluşan sıra dikkate alınarak en iyi-k tane öznitelik seçilebilir. Ancak birçok seçim işleminde tek bir amaç yerine birden çok amaca uygun özniteliklerin seçilmesi istenebilir. Bu durum çok kriterli karar verme problemi olarak karşımıza çıkar. Problemin çözümü için tüm amaçları dikkate alıp, farklı amaçlardan gelen derecelendirme değerlerini birleştirerek bir konsensüs skor üreten yöntemler önerilmiştir. Bu yöntemlerde nihai derecelendirme dikkate alınarak en iyi-k öznitelik seçilir. Buna karşılık bir nihai skor üretmeksizin öznitelikleri tüm amaç uzaylarında yarıştırarak seçim yapan yöntemlerde önerilmiştir. Bu bölümde çok kriterli öznitelik derecelendirme ve seçim yöntemleri ele alınmıştır.

Çalışmamız boyunca PO yöntemi, temel çok kriterli öznitelik seçim yöntemi olarak kullanılmıştır.AHP yöntemi ise PO tarafında seçilen ancak bir sıralaması olmayan özniteliklerin önceliklendirilmesi amacı ile kullanılmıştır. Bu amaçla, PO ve AHP yöntemini birlikte kullanan karma bir metot önerilmiştir. Önerilen karma yöntem ile

farklı yerel öğreniciler ve ya farklı amaçlara hizmet eden yöntemler birleştirilerek öznitelik seçme ve önceliklendirmeişlemi yapılmıştır. Önerilen karma yöntem, mikrodizi veri kümelerinde hastalık ilişkili genlerin seçilmesi ve sıralanması için kullanılmıştır. HGDP veri kümesinde ise etnik grupları ayırd edebilecek en az sayıda SNP alt kümesi seçme amacı ile kullanılmıştır.

3.4.1 Pareto Optimallik (PO)

Birçok durumda, çok kriterli karar problemlerinde tek bir en uygun çözüm bulunamaz. Bunun yerine, en az bir amaç için, diğer çözümlerden daha uygun değerlere sahip alternatif çözümler kümesi bulunur. Bu alternatif çözümler,"Pareto Optimal Çözümler" olarak bilinir [13].

Çok amaçlı optimizasyon problemleri, *m* adet parametre (karar değerleri) ve *n* adet amaç içeren bir *f* vektör fonksiyonunun en büyük ya da en küçük kılınması olarak tanımlanabilir. Fonksiyon şu şekilde verilir:

$$\text{En uygun hale getir (optimizasyon)} \quad y = (f_1(x), f_2(x), ..., f_n(x)) \tag{3.18}$$

$$\text{Bağlı olunan kısıtlar} \quad \begin{aligned} x &= (x_1, x_2, ..., x_m,) \in X \\ y &= (y_1, y_2, ..., y_n,) \in Y \end{aligned} \tag{3.19}$$

Burada, *x* karar vektörü; *y* ise amaç vektörü olarak bilinir. *X* parametre uzayını ve *Y* de amaç uzayını temsil eder [13]. Çözüm kümesi, başka herhangi bir karar vektörü tarafından kapsanmayan tüm karar vektörlerini içerir. PO kümesindeki her karar vektörü en az bir en uygun amaç değeri içermektedir. Bu bastırılmayan/kapsanamayan çözümler "Pareto Optimal Küme"olarak bilinir.

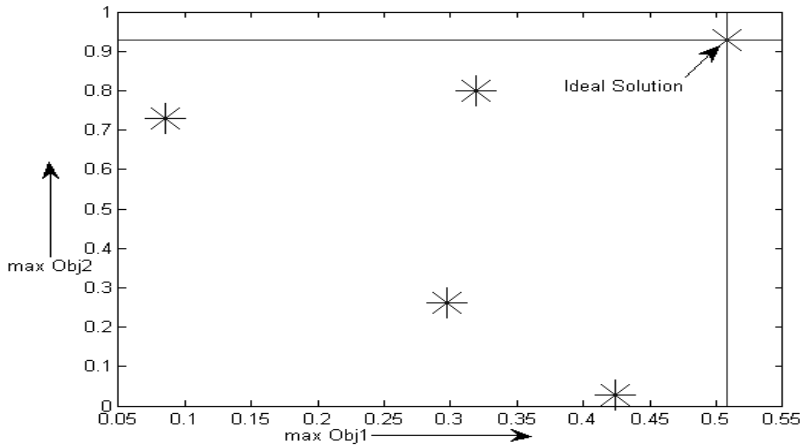
En iyileme problemini tüm amaçların en küçük kılınması bir problemi ve $a, b \in X$olmak üzere iki adet karar vektörü olduğunu varsayalım. Aşağıdaki kısıt(3.20) altında "*a, b*'ye *baskındır ya da b'yi kapsar.*", diğer bir deyişle "*b, a tarafından bastırılmıştır/kapsanmıştır.*" denir.

$$\begin{aligned} &\forall_i \in \{1,2,...,n\} : f_i(a) \leq f_i(b) \wedge \\ &\exists_j \in \{1,2,...,n\} : f_j(a) < f_j(b) \end{aligned} \tag{3.20}$$

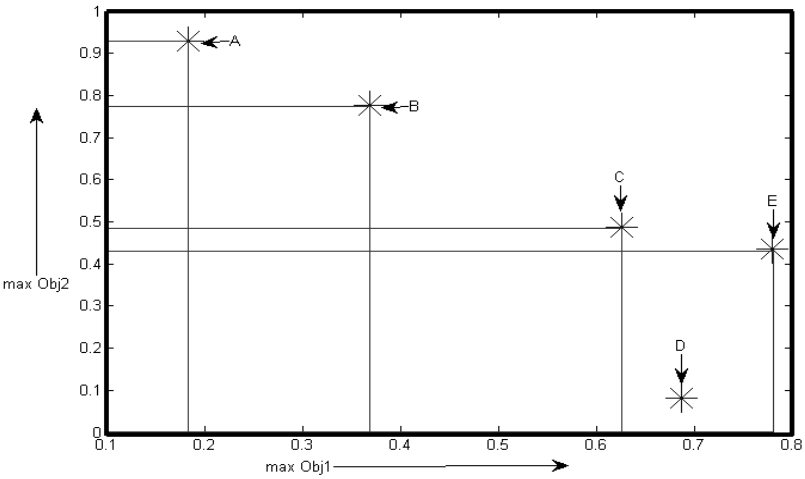
Şekil 3.5'de görüldüğü üzere, iki amacın en büyük kılınması problemi için, eğer bir çözüm, her amaç için en büyük değerlere sahipse, bu çözüm en uygun/ideal çözümdür.

İdeal çözüm, tüm amaçlar için diğer tüm çözümleri kapsar. Fakat bir çözüm, eğer en az bir ve en fazla *n-1* amaç için en büyük değere sahipse bu, alternatif bir çözüm olur.

Örneğin, **Şekil 3.6**'de en büyük kılınmış iki amaç ve beş çözüm görülmektedir. Dört çözüm (A, B, C ve E) herhangi bir nokta tarafından kapsanmadığından Pareto-Optimal çözümlerdir. Fakat çözüm D, Pareto-Optimal bir çözüm değildir, çünkü E çözümü iki amaç için de daha yüksek değerlere sahiptir. Buna göre; D, E tarafından bastırılmıştır/kapsanmıştır. Diğer bir deyişle, E, D'yi bastırır/kapsar. Eğer E çözümü mevcut olmasaydı, D çözümü Pareto-Optimal bir çözüm olacaktı. Çünkü birinci amaç için en büyük değere sahiptir.

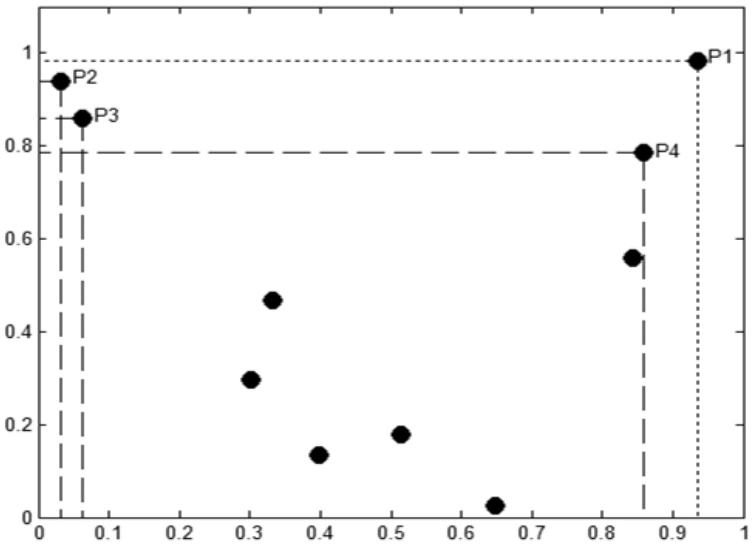


Şekil 3.5: İdeal Çözüm



Şekil 3.6: Pareto Optimal Çözümler

Diğer bir örnek **Şekil 3.7**'da 10 tane çözümü bulunan iki kriterli maksimizasyon problemi görülmektedir. P1 çözümü, diğer tüm çözümleri kapsadığı için ideal çözüm olarak ifade edilmiştir. Diğer taraftan, P1 çözümü mevcut olmasaydı, P2, P3 ve P4 çözümleri, herhangi bir çözüm tarafından kapsanmadığından pareto optimal çözümlerdir. Etiketlenmeyen 6 nokta, P4 tarafından kapsandıklarından 1. Seviye optimal çözüm kümesinde yer almamışlardır.



Şekil 3.7: İdeal Çözüm Ve 1. Seviye Pareto Optimal Çözümler[49]

3.4.2 Analitik Hiyerarşi Prosesi (AHP)

Saaty [34], AHP'yi, çok kriterli karar verme problemlerini çözmek için önermiştir. AHP, kriterlerin ağırlıkları yardımıyla, alternatiflerin göreceli derece/önem veya ağırlıklarını belirlemek için kullanılır. AHP, dört adımdan oluşur:

i) Çok kriterli karar verme problemi tanımlanır. Amaç, kriterler ve alternatifler belirlenir.
ii) Amaç, kriter/altkriterler ve alternatifler kullanılarak bir hiyerarşi ağacı oluşturulur.
iii) Hiyerarşi ağacındaki her eleman için ikili karşılaştırma matrisi oluşturulur.
iv) Karşılaştırma matrisleri kullanılarak nihaive yerel öncelikler/ağırlıklarelde edilir.

AHP'nin bir girdisi; amaç, kriterler/alt kriterler ve alternatiflerden oluşan çok seviyeli bir yapıdır. Bir diğer girdisi ise, kriterler/alt kriterler ikili karşılaştırma matrisidir. Saaty, kriterler arasında karşılaştırma yapabilmek için, göreceli önem derecesi skalasını önermiştir. Alt çıktılar, her kriter için alternatiflerin yerel öncelikleridir. Nihai çıktı ise

alternatiflerin genel öncelikleridir. AHP'de karşılaştırma matrislerinden öncelikler çıkarmak için pek çok yöntem bulunmaktadır: Eklemeli (additive) Normalizasyon, Özvektör, Ağırlıklı en küçük kareler, Logaritmik En Küçük Kareler, Logaritmik Amaç Programlama ve Bulanık Tercih-Programlama [50].

Alternatiflerin derecelendirme değerlerinin oranı, her bir kriter için, alternatiflerin ikili karşılaştırma matrisini oluşturmada kullanılır. Örnek bir karşılaştırma matrisi**Tablo 3.6**'de verilmiştir (r_i *i*. alternatifin derecesini göstermektedir). Bu kullanım,ikili karşılaştırma matrisinin tutarlılığını garantilemektedir. Eğer matris, aşağıdaki özelliğe sahipse karşılıklı (reciprocal) matris olarak isimlendirilir [51].

$$A=(a_{ij})=a_{ij}=r_i/r_j, \forall i,\ j\in\{1,2,...,n\} \text{ ve } a_{ij}>0, \forall i,\ j \tag{3.21}$$

Dolayısı ile A matrisi çarpmaya göre karşılıklılık özelliğini sağlar: $a,b\in X, a_{ij}\times a_{ji}=1$

Tablo 3.6: İkili Karşılaştırma Matrisi

C_1	A_1	A_2	…	A_n
A_1	1	r_1/r_2	…	r_1/r_n
A_2	r_2/r_1	1	…	r_2/r_n
…	…	…	1	…
A_n	r_n/r_1	r_n/r_2	…	1

Bu çalışmada hiyerarşideki herbir karşılaştırma matrisinden öncelikler türetmek amacıyla eklemeli en iyileme yöntemi ("Yaklaşıklık Yöntemi" olarak da isimlendirilir) kullanılmıştır. Bu yöntemde A ($n\times n$), matrisinden öncelikleri/ağırlıkları elde etmek için aşağıdaki adımlar uygulanır.

i) Sütunu normalize et; A matrisindeki her değeri, kendi sütununun toplam değerine böl.
ii) Normalize edilmiş sütundaki tüm elemanların toplamını hesapla.
iii)Toplamı, sütunun eleman sayısına böl.

Bu işlem, aşağıdaki denklem ile formülize edilebilir (w_i, A matrisindeki sütunun önceliğini göstermektedir.)

$$w_i = (1/n)\sum_{j=1}^{n}(a_{ij} / \sum_{j=1}^{n} a_{ij}), \ i=1, 2, ...,n. \quad (3.22)$$

Alternatiflerin nihai önceliği, ikili karşılaştırma matrislerinden elde edilen karar verme matrisi kullanılarak hesaplanır. Bir karar verme matrisi **Şekil 3.8**'de görülmektedir. A, C ve W sırasıyla alternatifleri, kriterleri ve ağırlıkları temsil etmektedir. N ve M kriterlerin ve alternatiflerin sayısını göstermektedir. w_j, c_j. kriterin ağırlığını, ve a_{ij} , *i.* alternatifin, *j.* kriterdeki önceliğini temsil etmektedir [52].

	Criterion				
	C_1	C_2	C_3	...	C_N
Alt.	W_1	W_2	W_3	...	W_N
A_1	a_{11}	a_{12}	a_{13}	...	a_{1N}
A_2	a_{21}	a_{22}	a_{23}	...	a_{2N}
A_3	a_{31}	a_{32}	a_{33}	...	a_{3N}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_M	a_{M1}	a_{M2}	a_{M3}	...	a_{MN}

Şekil 3.8: Ahp Karar Matrisi [52]

Alternatiflerin nihai öncelikleri aşağıdaki formül kullanılarak hesaplanır (S_i, *i.* alternatifin AHP skorunu işaret etmektedir.):

$$S_i = \sum_{j=1}^{N} a_{ij} w_j \ , i=1, 2, \ldots,M. \quad (3.23)$$

3.4.3 Condorcet Derecelendirme

Condorcet derecelendirme işleminde, belirli bir adaylar kümesi varken, seçmenlerin tüm adayları sıralaması istenilir (en çok istenilen adayı 1. sıraya en az istenilen adayı sonuncu sıraya konulur). Daha sonra adayların ikili tercih karşılaştırması yapılarak adaylar tercih puanlarına göre sıralanır. Bu işlemler sonucunda en çok tercih edilen aday seçimi kazanır.

- ✓ C aday kümesi olmak üzere $C = \{c_1, c_2, \ldots, c_n\}$,
- ✓ V seçmen tercihleri kümesi (n aday için sıralama) olmak üzere$V = \{>_i \ \ldots \ >_n\}$,
- ✓ $f_i(c_1, c_2)$, seçmen i'nin c_1 adayını c_2 adayına tercih ettiğini göstersin yani $c_1 >_i c_2$:

- $f_V(c_1, c_2) = |\{>_i \in V |\ c_1 >_i c_2\}|$, c_1 adayının c_2 adayına tercih edilme sayısını gösterir.

- Eğer $f_V(c_1, c_2) \geq f_V(c_2, c_1)$ ise, c_1 adayı c_2 adayından daha çok tercih edilmiştir.

- Burada $\forall\ c_1 \neq c_2\ \in C: f_V(c_1, c_2) + f_V(c_2, c_1) = v$ olmalıdır.

Örnek verecek olursak, C=3 aday (x,y,z) ve V=25 seçmen olsun. Tüm seçmenlerin aday sıralaması **Tablo 3.7**deki gibi olsun. Seçmenlerin sıralamalarına bakarak ikili tercih matrisi (matrix of pairwise preferences) oluşturulur **Tablo 3.8**. Herbir adayın diğer herbir adaya göre kaç kez tercih edildiği hesaplanır. Örneğin x-y adaylarının karşılaştırması için: x adayı (z>x>y, x>z >y tercih profillerine göre) 9+6=15 kez y adayına tercih edilmiştir. y adayı ise 10 kez x adayına tercih edilmiştir. $f_V(x, y) + f_V(y, x) = v$; 15+10=25. İkili tercih matrisindeki toplam sutunun sıralaması bize adayların sıralamasını vercektir.Tercih toplamlarına bakıldığında z adayının sıralamalarda 31 kez diğer adaylardan önce yer aldığını , y adayının ise yalnız 21 kez önce yer aldığını buluruz. Dolayısı ile z condorcet galibi, y ise condorcet mağlubudur. Adayların enaz birinin diğerlerini yenememesi durumu literatürde condercet paradoksu olarak bilinir[53].

Tablo 3.7: Seçmen Tercihleri

9 seçmen	2 seçmen	6 seçmen	8 seçmen
z	Y	x	Y
x	X	z	Z
y	Z	y	X

Tablo 3.8: İkili Tercih Matrisi

	x	y	z	Toplam
x	-	15	8	23
y	10	-	11	21
z	17	14	-	31

Condorcet derecelendirm çok amaçlı özellik seçiminde, farklı metotlarla sıralanmış özelliklerin yeniden sıralanması [54] ve kümeleme algoritmalarının sıralanması amacı ile kullanılmıştır [55].

3.4.4 MC4 Derecelendirme

Markov Zinciri (Markov Chain-MC), durumlar arasındaki geçişleri ve geçişlerin olasılıklarını ifade etmede kullanılan bir yöntemdir. Sonlu yada sayılabilir sayıda durumun geçiş olasılıklarını gösteren matrise geçiş matrisi (transition matrix) denir ve bu matris ile gelecekteki bir durum geçmiş durumun analizi ile tahmin edilebilir.

$\{X_d: \text{d=1,2, ...}\}$, d durumdan oluşan bir markov zinciri ve U tüm durum uzayı olsun. $M = (m_{i,j})$olarak tanımlanan M matrisi U üzerinde tanımlı bir markov matrisidir ve bir markov zincirinin geçiş matrisi olarak tanımlanır. $i, j \epsilon\, U$için tanımlı $m_{i,j}$sayıları için aşağıdaki iki durum geçerlidir. Matris üzerindeki herbir eleman bir olasılık gösterdiğinden negatif bir değer alamaz. Ayrıca bir satırdaki tüm olasılıkların toplamı 1 olmalıdır.

$$0 \leq\ M_{i,j} \leq 1 \forall i, j \epsilon\, U\ \ ve\ \ \sum_{j=1}^{d} m_{i,j} = 1\ \forall j \epsilon\, U\,, (i = 1,2, \dots, d) \tag{3.24}$$

Yukarıdaki şartlar altında geçiş matrisi *M*aşağıdaki formda gösterilir:

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,d} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,d} \\ \vdots & \vdots & \cdots & \vdots \\ m_{d,1} & m_{d,2} & \cdots & m_{d,d} \end{bmatrix}$$

Geçiş matrisinin satırları, bir durumdan diğer tüm durumlara geçiş olasılıklarını gösteren durum vektörleridir(v). i. satırdaki v_i olasılık vektörü i. durumdan diğer tüm durumlara geçişi simgeler. Durum vektörünün elemanları olasılıksal bir değer içermektedir. Dolayısı ile tüm değerler [0-1] aralığındadır ve vektor toplamı 1 dir. $v^0\ ve\ v^n$, sırasıyla geçiş matrisinin başlangıç ve denge konumundaki durum vektörlerini simgeler. Durum vektörü aşağıdaki bağıntı ile ifade edilir.

$$v^{n+1} = v^n * M \tag{3.25}$$

Geçiş matrisi üzerinde herhangi bir durumdan diğer tüm durumlara geçiş mümkün ise bu markov zinciri ergodiktir. Markov modeli ile tanımlanmış bir sürecin denge durumuna (steady-state) ulaşabilmesi için mutlaka ergodik olması gerekir. M^n, M ergodik geçiş matrisinin denge durumu;v_i, M matrisinin olasılık vektörünü göstermek üzere, denge durumu için iki kural geçerlidir:

i. n nin yeterince büyük değerleri için, v_i^n olasılık vektörlerinin değerleri aynıdır ve değişmez.

ii. $v_i^{n+1} = v_i^n * M \; ve \; v_i^{n+1} = v_i^n \; olduğundan \;\; v = v * M$ eşitliğini sağlayan bir v denge vektörü vardır. Bu durumda v vektörü denge durumu koşullarını içeren olasılıkları kapsar.

MC4, markov zinciri tabanlı bir algoritmadır. Bu özel algoritmanın amacı, bir web sayfasının yüksek derecelendirilme şansını arttırmak için arama motorlarının yönlendirilmesi (search engine spamming) ile mücadele etmekti. Spamlerden etkilenen arama motorları genellikle yüksek derecelendirilmiş ilgisiz sayfalar getirirler. MC4 algoritmasının amacı, arama listelerinnin az bir kısmında sahte bir şekilde yüksek derecelendirilmiş sayfaları önemsemeyen ortak bir derecelendirme oluşturmaktır [56]. Bu algoritma mikrodizi veri kümelerininin analiz sonuçlarının birleştirilmesi (meta analiz) amacı ile kullanılmıştır [56].

MC4, farklı sıralayıcılar tarafından oluşturulmuş listeleri birleştirerek ortak bir tercih sıralaması oluştur. Listelerin en azından bir tanesinde sıralanmış öğelerin ikili karşılaştırmasına dayandığından dolayı MC4 algoritması ilişkisel yaklaşım olarak tanımlanabilir. Algoritma bu ikili karşılaştırmaların sonuçlarını kullanarak ergodik geçiş matrisini oluşturur. M geçiş matrisinin sütün bazında toplanması ile ayrı ayrı listelerden ortak bir derecelendirme skoru elde edilir.Algoritmanın adımları aşağıda tanımlanmıştır[56]:

i) En az bir top-k listesinde bulunan tüm elemanların birleşimini içeren bir U kümesi oluşturulur.

ii) U'daki her i-j çifti için, eğer i ve j'nin ikisinide sıralamış listelerin çoğunluğu (%50den fazlası) j'yi i'den yukarıda sıralamış ise $m_{i,j}^* = 1$, değilse 0'dır ve j i'ye tercih edilmiştir. i ve j elamanları hiç bir listede karşılaştırılmamışsa $m_{i,j}^* = m_{j,i}^* = 0.5$

iii) Geçiş matrisi M = $\{ m_{ij} \}$ aşağıdaki şekilde tanımlanır: $:i \neq j$ için $m_{i,j} = m_{i,j}^*/|U|, \; m_{ii} = 1 - \Sigma_{j \neq i} m_{ij}$

iv) M geçiş matrisinin her elemanı 1− ε ile çarpılır, ardından her elemana ε / |U| eklenerek matris ergodik hale getirilir.

Condercet bölümündeki seçmen tercihleri (**Tablo 3.7**) üzerinden MC4 algoritması i-iv adımları için çalıştırılacak olursak aşağıdaki geçiş matrisi elde edilir.

i. U=$\{x,y,z\}$ en az bir top-k liste bulunan tüm elemanlar kümesi

ii. $m^*_{i,j}$ lerin hesaplanması

a. x>y tercihleri, tüm tercihlerin %50'sinden büyüktür(15/25→%60). Dolayısı ile x, y'ye tercih edilmiştir. $m^*_{x,y} = 0, m^*_{y,x} = 1$olur.

b. x>z tercihleri, tüm tercihlerin %50'sinden küçüktür (8/25→%32). Dolayısı ile z, x'ye tercih edilmiştir. $m^*_{x,z} = 1, m^*_{z,x} = 0$olur.

c. y>z tercihleri, tüm tercihlerin %50'sinden küçüktür (11/25→%44). Dolayısı ile z, y'ye tercih edilmiştir. $m^*_{y,z} = 1, m^*_{z,y} = 0$ dir.

iii. Köşegen dışındaki değerler elaman sayısınına (3 aday) bölünür ve matrisin köşegeni, satır toplamlarının birden çıkarılması ile hesaplanır.

iv. Geçiş matrisinininözdeğer-özverktörler hesaplanır. Değeri bir olana özdeğere karşılık gelen özvektör adayların önceliklerini gösterir.

Tablo 3.9: Geçiş Matrisi

	x	y	z	→		x	y	z
X	-	0	1		x	2/3	0	1/3
Y	1	-	1		y	1/3	1/3	1/3
Z	0	0	-		z	0	0	1

3.5 SINIFLANDIRMA YÖNTEMLERİ

Sınıflandırma, X girdi ve Y çıktı değişkenleri için, girdiden çıktıya bir eşleme öğrenen, gözetimli öğrenme uygulamasıdır. Parametrelere göre tanımlanan bir model ile girdi kullanarak çıktı değeri kestirilmeye çalışılır. Model$g(.)$, parametreleri iseθ ile gösterilir. $Y = g(x|\theta)$ile gösterdiğimiz çıktı 0 ya da 1 değerini alabilen sınıf etiketidir [4].

Sınıflandırma modellerine ait formüllerde kullanılacak veri setinin yapısı aşağıdaki gibidir.

$$X = \left\{ x_t^d , c_t \right\}_{t=1}^{N}, d \in R \tag{3.26}$$

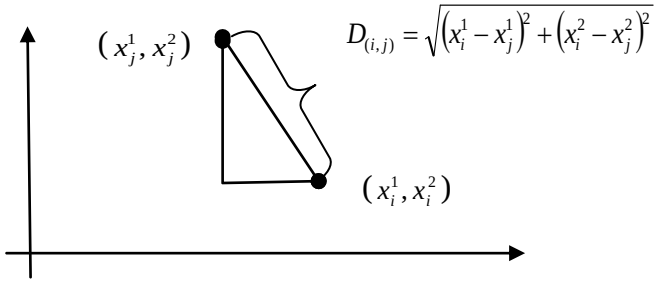
Burada (X) veri setini, (*x)* sınıflandırmada esas alınan öznitelikler vektörünü, (*d)* x öznitelik vektörünün boyutunu ve öznitelik sayısını, (*c*) sınıf bilgisini içeren öznitelikleri, (*N/n)* örnek sayısını göstermektedir.

3.5.1 K-En Yakın Komşu

Cover ve Hart[57]tarafından önerilen bu yöntemde temel işlem, sınıfı tahmin edilecek yeni örneğin, sınıfları bilinen diğer örneklere olan uzaklığına bakarak bu örneği sınıflandırmaktır. Sınıflandırma için en yakın (uzaklığı en az olan) k tane komşuya bakılır. Komşular en çok hangi sınıfa ait ise yeni örneğimiz o sınıftandır. 2 sınıflı bir problemde her sınıftan eşit sayıda komşu olmasını engellemek için k komşu 1,3,5,7 gibi tek sayılardan seçilir. Uzaklık hesabı için en çok bilinen Öklid ve Manhattan uzaklıklarıdır. Bu iki uzaklık hesabı için aşağıdaki denklemler kullanılır.

i. **Öklid uzaklığı:** 2 örneğe ait tüm özniteliklerin farklarının karelerinin toplamlarının karekökü olarak hesaplanır. x_i ve x_j örneklerinin Öklid uzaklık hesabı aşağıdaki formül ile hesaplanır.

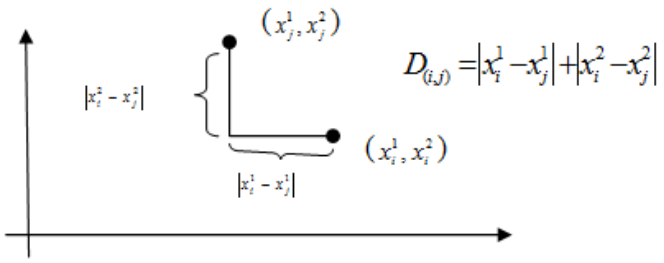
$$D_{(i,j)} = \sqrt{\sum_{k=1}^{d} \left(x_i^k - x_j^k \right)^2} \tag{3.27}$$



Şekil 3.9: Öklit Uzaklığı

ii. **Manhattan uzaklığı:** 2 örneğe ait tüm özniteliklerin farklarının mutlak değerinin toplamları olarak hesaplanır. x_i ve x_j örneklerinin manhattan uzaklık hesabı aşağıdaki formül ile hesaplanır.

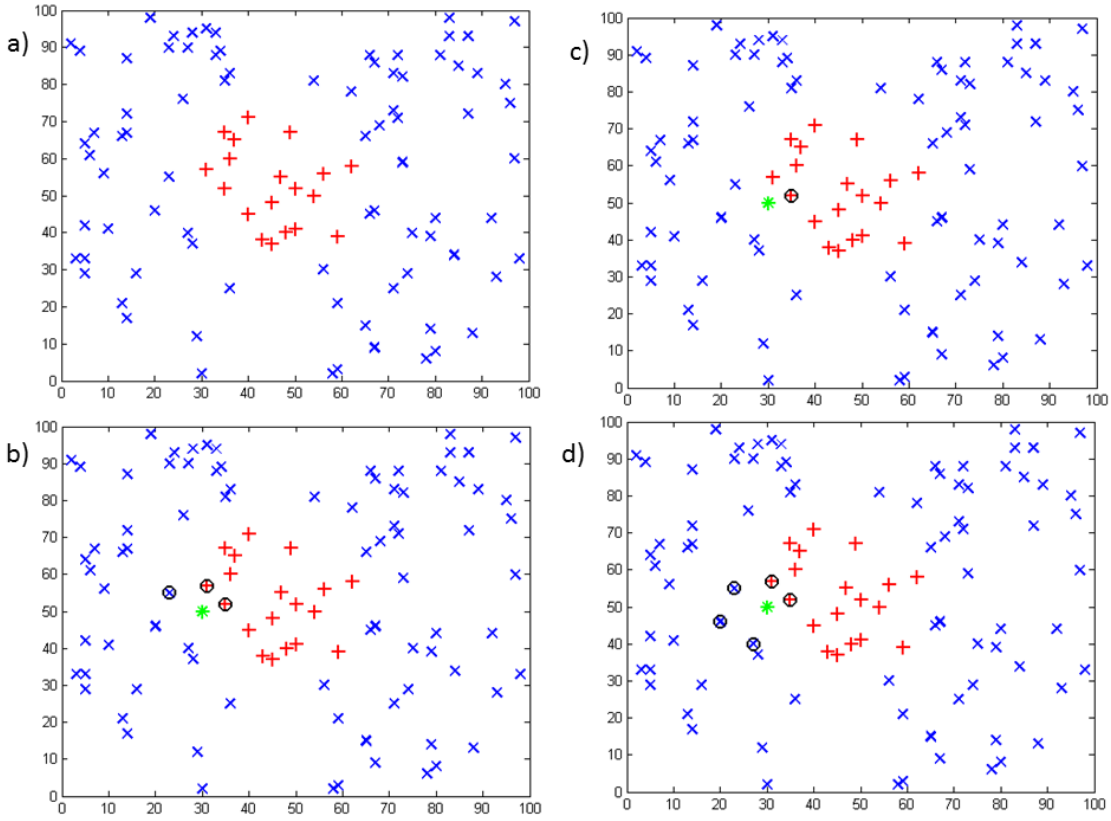
$$D_{(i,j)}= \sum_{k=1}^{d}\left|x_i^k - x_j^k\right| \tag{3.28}$$



Şekil 3.10: Manhattan Uzaklığı

Knn yönteminin 2 sınıflı,2 boyutlu bir problemde bir test örneğine uygulanışı aşağıda anlatılmıştır. **Şekil 3.11a**'da örneklerin 2 boyutlu uzayda dağılımını göstermektedir. Tüm şekillerde kırmızı noktalar pozitif örnekleri, mavi noktalar negatif örnekleri, yeşil nokta test örneğini temsil etmektedir. **Şekil 3.11**'deki**b, c, d**şekillerinde en yakın 1,3,5 komşuluğa göre sınıflandırma işlemi gösterilmiştir. Test örneğine en yakın komşular, siyah daire içine alınmıştır. Buna göre:

- ✓ k=1 komşuluk için en yakın komşu pozitif sınıfta olduğundan test örneği **pozitif** olarak sınıflandırılır.
- ✓ k=3 komşuluk için en yakın 2 komşu pozitif, 1 komşu negatiftir. En fazla örneğe sahip sınıf test örneğinin sınıfı olarak atanır ve test **pozitif** olarak sınıflandırılır.
- ✓ k=5 komşuluk için en yakın 2 komşu pozitif, 3 komşu negatiftir. En fazla örneğe sahip sınıf test örneğinin sınıfı olarak atanır ve test **negatif** olarak sınıflandırılır.



Şekil 3.11: 2 Sınıflı, 2 Boyutlu Bir Problem Üzerinde Knn Uygulaması

3.5.2 Naive Bayes Sınıflandırıcı

Bayes sınıflandırıcı, X girdilerine sahip bir örneğin, herbir sınıfta olabilme olasılığına bakarak sınıflandırma yapar. Bunun için X koşulu var iken herbir sınıfın olasılığı hesaplanır. En yüksek olasılığa sahip sınıf, X girdisine karşılık üretilecek Y sınıfını belirler. Naive Bayes, değişkenlerin bağımsız olduğu durumlarda kullanılan bayes kuralının özelleşmiş halidir. Naive Bayes kullanarak X örneğinin sınıfsal olasılığının hesaplanmasında aşağıdaki formül kullanılır. Herbir sınıf için şu sorunun cevabı aranır: $X=(x_1,x_2,...x_d)$ örneğinin C sınıfında olma olasılığı P(C|X) nedir?

$$P(C|X) = \frac{P(C).p(X|C)}{p(X)} \tag{3.29}$$

$$P(C|x_1, x_2, ..., x_d) = \frac{P(C).p(x_1, x_2, ..., x_d|C)}{p(x_1, x_2, ..., x_d)} \tag{3.30}$$

Denklemin parçaları aşağıdaki şekilde de yazılabilir:

$$posterior = \frac{priorxlikelihood}{evidence} \quad (3.31)$$

Burada, *Prior (önsel olasılık)* örneklerin özellik bilgisine bakmadan, sadece sınıf bilgisine bakarak bulunan olasılığı; *Likelihood (sınıf olabilirliği)* sınıf bilgisi bilindiğinde örneklerin olasılığını; *Posterior (sonsal olasılık)*sonsal olasılık değeri en büyük olan sınıfını, X örneğinin sınıfını; *Evidence* (*kanıt)* her bir sınıf için sabit olduğundan, büyüklük karşılaştırmasında etkisiz elemanı göstermektedir [4].

Naive Bayes sınıflandırma aşağıdaki örnekle açıklanmıştır.**Tablo 3.10**'da 2 sınıflı 2 boyutlu bir sınıflandırma problemi için öğrenme ve test kümeleri yer almaktadır. Amaç öğrenme kümesi yardımı ile test örneklerinin sınıf bilgisinin tahmin edilmesidir. Naive Bayes sınıflandırıcı ile problemin çözümü iki adımdan oluşur: i) modelin eğitilmesi, ii) eğitilmiş model ile test örneklerinin sınıflarının tahmin edilmesi.

Tablo 3.10: 2 Sınıflı, 2 Boyutlu Veri Kümesi

Küme	Öznitelik 1=X	Öznitelik 2=Y	Sınıf Bilgisi
Öğrenme	2	5	A
	2	7	A
	3	7	A
	4	2	B
	3	1	B
	4	1	B
	4	0	B
Test	3	2	?

Test örneğinin sınıflandırmasıaşağıdaki olasılıkların hesaplanmasını gerektirir:

$$P(A|X=3,Y=2) =? \text{ ve } P(B|X=3,Y=2) =? \quad (3.32)$$

Problemin çözümü için ilk önce model eğitilmelidir. Bu amaçla herbir öznitelik için olasılık tabloları hazırlanır(**Tablo 3.11**). Herbir sınıf için önsel olasılık hesaplanır:

$$P(C=A)=3/7,\ P(C=B)=4/7 \quad (3.32)$$

Tablo 3.11:Öznitelikler İçin Sınıf Olabilirliği

X	C=A	C=B
2	2/3	0/4
3	1/3	1/4
4	0/3	3/4

Y	C=A	C=B
0	0/3	1/4
1	0/3	2/4
2	0/3	1/4
5	1/3	0/4
7	2/3	0/4

Eğitim sürecinden sonra formüldeki değerler yerine konup, test örneğinin herbir sınıfta olma olasılığı hesaplanır ve büyük olasılığa sahip sınıf test örneğinin sınıfı olarak atanır. Aşağıdaki işlemler sonucunda test örneğinin B sınıfından olma olasılığı A sınıfından olma olasılığından daha büyük çıkmıştır. Dolayısı ile örneğin sınıfı B olarak tahmin edilmiştir.

$$P(A|X=3,Y=2) \rightarrow P(C=A).P(X=3|C=A).\ P(Y=2|C=A) \rightarrow 3/7x1/3x0=0 \qquad (3.33)$$

$$P(B|X=3,Y=2) \rightarrow P(C=B).P(X=3|C=B).\ P(Y=2|C=B) \rightarrow 4/7x1/4x1/4=1/28 \qquad (3.34)$$

3.6 SEÇİLEN ÖZNİTELİK KÜMESİNİN DEĞERLENDİRİLMESİ

Yüksek boyutlu veri kümelerinin analizinde var olan problemlerden dolayı, veri kümesinin sahip olduğu özniteliklerden seçim yada çıkarım yapılarak verinin boyutunun azaltılması gerekliliği öznitelik seçim bölümünde tartışılmıştır. Burada ele alınacak konu ise seçilmiş olan öznitelik kümesinin ne kadar başarılı olduğunun belirlenmesi işlemidir. Ancak öncelikle herhangi bir modelin başarısının ölçülmesi konusuna değinilecek daha sonra seçilmiş ve derecelendirilmiş özniteliklerin birlikte nasıl değerlendirilecekleri ele alınacaktır.

Makina öğrenme modellerinin başarısının değerlendirilmesinde tekbir geçerleme yada test kümesinin kullanılması yeterli olmayacaktır[4].Veri kümesi bölümleme konusunda tartışıldığı üzere bir veri kümesinden elde edilen bir alt küme aykırı, istisnai ve gürültü örnekler içerebileceğinden evrensel kümeyi temsil gücü düşük olacaktır. Bu sebeple tek bir sonuç doğru bir gösterge olarak kabul edilmez[4]. Veri bölümlemedeki rastsallığın önüne geçmek için bir sonuç yerine birden fazla sonuç alınır ve başarı değerlendirmesi için ortalama değer kullanılır. Bir sınıflandırma problemi için başarı ölçütü olarak doğruluk, duyarlılık, özgüllük gibi değerler kullanılabilir. Bu çalışmada,başarı ölçütü

olarak modelin doğru sınıflandırdığı örnek sayısının tüm örnek sayısına bölümü olan doğruluk(accuracy-AC),sınıflandırıcının doğru olarak negatif tahminlediği örnek sayısının negatif tahminlediği tüm örnek sayısına oranını gösteren özgüllük (specificity-Spe) ve bir sınıflandırıcının pozitif tahminlediği örnek sayısının gerçekteki pozitif örnek sayısına oranını gösteren duyarlılık (sensitivity-Sen) değerleri kullanılmıştır. Sınıflandırma başarısının hesaplanmasında karışıklık matrisisinden yararlanılır. Aşağıda karışıklık matrisi kullanılarak başarı ölçütlerinin hesaplanması formulize edilmiştir.

Tablo 3.12: Karışıklık Matrisi

	Tahminlenen Sınıf		
Gerçek Sınıf		Sınıf =1	Sınıf=0
	Sınıf=1	TP	FN
	Sınıf=0	FP	TN

$$Doğruluk = \frac{TP+TN}{TP+FN+FP+TN} \tag{3.35}$$

$$Özgüllük = \frac{TN}{TN+FP} \tag{3.36}$$

$$Duyarlılık = \frac{TP}{TP+FN} \tag{3.37}$$

Derecelendirilmiş özniteliklerin başarısının hesaplanmasında arttırımlı alt küme başarısı kullanılmıştır. Bu amaçla herbir öznitelik, sırasına göre başarısı değerlendirilecek öznitelik alt kümesine teker teker eklenmiş ve tüm özniteliklerin birlikte olması durumundaki başarısı ölçülmüştür. Bu çalışmada seçilmiş öznitelikler için nihai sıralama AHP yardımı ile oluşturulmuştur. Bu tarz bir hesaplama işleminde ilk amaç, en yüksek başarıyı en düşük sayıdaki geni içeren bir altkümeyle elde etmektir. İkinci amaç ise, AHP tabanlı derecelendirmenin başarısının ortaya konmasıdır. Arttırımlı alt küme başarısının hesaplama adımları aşağıda sözde kod olarak belirtilmiştir:

Amaç: Sıralı gen altkümesi için gen altkümesi başarısının hesaplaması
Girdi: Gen kümesi sırası
Çıktı: Sıralı gen kümesinin altkümelerinin sınıflandırma başarısı

1. *for i=1:Sıralı gen kümesindeki gen sayısı*
2. *Altkümeye i. geni ekle*
3. *Gen altkümesinin sınıflandırma başarısını hesapla*
4. *end*

3.7 ÖNERİLEN YÖNTEMLER

Öznitelik seçim yöntemlerinin veri kümesindeki çeşitlilikten ve seçim modelinden etkilendiğini birçok çalışma rapor etmiştir [1, 3, 6, 58]. Tersine, çoğu çalışmada, öznitelik seçimi için kullanılan veri kümesi hakkında detaylı bilgi verilmemektedir. Araştırmacılar yalnızca başarı değerlendirme aşaması için veri kümesi oluşturma hakkında detaylı bilgi vermektedirler [8, 9, 59]. Öznitelik seçimi için, tüm verinin kullanıldığı ya da eğitim ve test verilerine bölündüğü düşünülebilir. Ancak öznitelik seçimi, veri kümesi çeşitliliğinden etkilenmektedir ve bu nedenle araştırmacılar öznitelik seçimi için kullanılan veri kümesi hakkında kapsamlı bilgi vermelidir. Öznitelik seçiminde kullanılan veri kümesi hakkında detaylı bilginin verilmediği durumlarda, tüm verinin kullanıldığı veya verinin birçok kez bölündüğü ve en iyi eğitim kümesinin seçildiği düşünülür.

Model eğitimi için oluşturulacak tek bir öğrenme örneklemi verinin tamamını temsil edememe durumu veri kümesi bölümleme başlığı altında tartışılmıştır. Diğer taraftan, verinin farklı örneklemlerini kullanan yerel uzmanların birleştirilmesi ile evrensel uzayı temsil etmede başarılı modeller oluşturulduğu bilinmektedir [4]. Verinin yanlı örneklenmiş bir alt uzayını temsil eden öğrenme kümesi kullanan bir yerel uzmanınseçmiş olduğu özniteliklerin yerine,yerel uzmanları birleştiren bir model ile seçilmiş özniteliklerin dikkate alınması, veriyi temsil gücü daha yüksek olan bir öznitelik alt kümesi seçecektir. Böyle bir durumda 3 farklı durum izlenebilir (i) yerel uzmanlarca seçilen özniteliklerin kesişim kümesi ve ya (ii) birleşim kümesi nihai alt küme olarak alınabilir. (iii) bu çalışmada önerildiği gibi çok kriterli yaklaşımlar yardımı ile herbir öğrenme kümesini kullanan yerel öğrencileri birleştiren bir model ile seçimi yapılabilir. Çok kriterli öznitelik seçim yaklaşımında da 2 farklı öneri getirilebilir. M bölümleme ve öznitelik seçim işlemi olsun. (i) Bazı öznitelikler tüm bölümlemeler de yüksek değerler alırken, bazı öznitelikler sadece bir ya da birkaç bölümlemede yüksek değer alabilir. Condercet, MC4 algoritmaları az sayıda bölümlemede seçilmiş öznitelikleri uç olarak nitelendirilip, yanlış seçim olduğunu düşünülerek bu öznitelikleri eler. (ii) Başka bir açıdan bakıldığında, en az bir bölümleme de yüksek skora sahip öznitelikler bir uç durumun temsilcisi olabilir ve bunlar öğrenme kümelerinin temsil edemediği evrensel kümenin temsilcisi olabilir. PO yaklaşımı bu uç özniteliklerin değerli olduğunu düşünerek bunları nihai alt kümeye dâhil eder.Çok kriterli öznitelik

derecelendirme ve seçme bölümünde anlatılan PO [20-22] ve AHP [26-33, 50, 52] yöntemleri, yöneylem araştırmalarının çokça başvurduğu çok kriterli karar verme yaklaşımlarıdır. PO yaklaşımı, son yıllarda, gen alt küme seçimi problemine de uygulanmış ve pek çok gen altkümesinin sınıflandırma performansını değerlendirerek en iyi altkümelerin seçiminde başarılı olmuştur[6, 14-17, 25]. Ancak bu çalışmaların ortak özelliği, PO yaklaşımın doğası gereği seçilen genler için bir sıralama verememesidir. Oysaki sıralama, biyo-işaretçi seçiminin önemli çıktılarından biridir. AHP ise biyolojik verilere, yalnızca bir çalışmada SNP'lerin önceliklendirilmesi amacı ile uyarlanmıştır [35]. Ancak, AHP'nin yüksek hesaplama maliyetinden ötürü çok boyutlu verilerdeki tüm özniteliklerin bir anda değerlendirilmesi için uygun değildir. Dolayısı ile çalışmada AHP'nin temel yapısında bazı kısıtlamalara gidilmiştir.

Bu temellerden yola çıkarak, veri kümesindeki varyasyonun öznitelik seçim işleminde oluşturduğu problemi çözmek için,verinin farklı alt uzaylarında uzmanlaşmış yerel öğrenicilerin/uzmanlarınçok kriterli karar verme yaklaşımı ile birleştirilmesi önerilmiştir. Öğrenme kümeleri herhangi bir bölümle yöntemi ile oluşturulmuş olabilir. Önemli olan evrensel uzayı temsil edecek daha fazla sayıda öğrenme örneklemi ile karar verilmesidir. Dolayısı ile hangi bölümleme ile yapılmalıdır sorusu bu çalışmanın kapsamı dışındadır.

Bu çalışmada, AHP'nin yüksek boyutlu verilerde çalışamama ve PO yönteminin derecelendirme yapamama problemlerinin çözümü için, PO yöntemi AHP ile entegre edilerek yeni bir karma yöntem önerilmiştir. Önerilen yöntem ile PO ve AHP öznitelik derecelendirme ve seçme amacı ile biyolojik verilere ilk kez bizim tarafımızdan başarılı bir şekilde uyarlanmıştır. Bu amaçla, ilk olarak, ilgisiz tüm öznitelikler/genler PO yöntemi ile elenmiş, daha sonra geriye kalan az sayıda öznitelik/gen AHP ile sıralanmıştır. Karma yöntem kanser ve hipertansiyon (Gormez et al., 2013) veri kümelerine uygulanmıştır.

Bir diğer problem ise öznitelik seçim işleminin metot bağımlılığıdır. Ancak farklı metotlar öznitelik seçiminde karşılaşılan problemlerin üstesinden gelmek için önerilmiştir ve her biri farklı bir amaca hizmet etmektedir. Metotların zenginliğinin öznitelik seçim işlemine yansıtılarak, farklı amaçlar gözeten metotların birleştirilmesi ile öznitelik seçiminde daha başarılı modellerin ortaya çıkacağı bilinmektedir[4].

Dolayısı ile bu çalışmada farklı amaçlara hizmet eden öznitelik seçim metotlarının çok kriterli karar verme yöntemleri ile birleştirilmesi önerilmiştir. Bu amaçla birçok kriteri dikkate alarak işlem yapabilen PO yönteminin kullanılması önerilmiştir. Önerilen yöntem HGDPveri kümesine uygulanmıştır. Sonuçlar PO yöntemi ile çok amaçlı öznitelik seçiminin başarılı bir şekilde yapılabileceğini göstermiştir.

Önerilen yöntemlerin deneysel sonuçları Bölüm 4'te sunulmuştur. Sonuçlar, önerilen yöntemlerin başarılı olduğunu göstermektedir. Bu bölümde ise önerilen yöntemlerin temel prensipleri ve uyarlanış biçimleri açıklanmıştır.

3.7.1Öznitelik Seçimi için Yerel Uzmanların Pareto Optimal Yaklaşımı ile Birleştirilmesi

Gen seçimindeki klasik yaklaşım; tüm özniteliklerim öncelikle herhangi bir derecelendirme yöntemi ile derecelendirilmesi, ardından en yüksek değere sahip en iyi-*k* adet özniteliğin veya önceden belirlenen bir eşik değerinden yüksek dereceye sahip öznitelik altkümesinin seçimi şeklindedir. Bir çalışmamızda[3], öznitelik derecelendirme yöntemlerinin verikümesi çeşitliliğinden etkilendiğigösterilmiştir. Öznitelikler, *N* farklı eğitim kümesi ile derecelendirildiğinde, her seferinde dereceleri ve önem sıraları farklı olmaktadır. Bu durumda, seçim için, hangi derecelendirme değeri kullanılmalıdır sorusu gündeme gelmektedir. Ayrıca, kaç özniteliğin problemin tanımı ile (örneğin hastalık tahmini, etnik grup ayrımı) ilişkisi olduğu ve en iyi-*k* öznitelikteki k'nın kaç olacağı da diğer sorulardandır. Diğer taraftan, en uygun eşik değerinin belirlenmesi de bir başka problemdir.

Pareto Optimal yöntemi ile en iyi *k* veya eşik değerine önem vermeden uygun öznitelik altkümeleri elde edilebileceği yaklaşımı ile bu çalışmada, öznitelik seçimi için, verinin farklı alt kümelerini eğitim içinkullanan yerel uzmanlar oluşturularak veri kümesinin tüm çeşitliliğinin temsil edilmesi yoluna gidilmiştir.Verinin farklı alt uzaylarını kullanan yerel uzmanlar, çok kriterli karar verme metodu olan PO yardımıyla birleştirilmiştir. Çok kriterliöznitelik seçimi yapmak amacıyla yerel uzmanları birleştirmeyi öneren bu yöntem kanserve hipertansiyon mikrodizi veri kümeleri üzerinde test edilmiştir. Burada artık öznitelik seçimi özel anlamda gen seçimine dönüştüğünden öznitelik kavramı yerine gen kullanılmıştır.

Bu çalışmadaki çok amaçlı eniyileme problemi, *d* (öznitelik sayısı) adet parametre (karar değişkeni) ve *N* (seti bölünmesi bölümleme sayısı) amacı içeren bir *f* vektör fonksiyonunu en büyük kılmaktır. Buradaki temel amaç başka sonuçlar tarafından bastırılamayan sonuçları bulmaktır. Öncelikle,verinin farklı alt kümeleri kullanılarak oluşturulan N yerel uzman tarafından her bir genderecelendirilmiştir. Buradaki N derecelendirmedeğeri, Pareto Optimal yaklaşımında kriterlere karşılık gelmektedir. PO ile gen seçim işlemi sözde kod olarak aşağıda belirtilmiştir:

Girdi: dxN boyutlu derecelendirme matrisi (d=gen sayısı, N kriter sayısı)
Çıktı: Pareto Optimal Küme

1. *İlk geni, ParetoOptimum kümeye (POK) ekle.*
2. *for i=1:d*
3. *Eğer i. genin en az bir derecelendirme değeri, Pareto-Optimal kümedeki tüm genlerin değerlerinden daha yüksekse, i. geni, ParetoOptimal kümeye ekle, değilse, i. geni bu kümeye ekleme.*
4. *Eğer i. genin tüm derecelendirme değerleri, Pareto-Optimal kümedeki herhangi bir genin derecelendirme değerinden daha yüksek ise, i. geni Pareto Optimal kümeye ekle ve daha uygun değere sahip olmayan diğer geni kümeden çıkar.*
5. *end*

Kanser veri kümelerinde PO ile gen seçimi yapılırken, veri kümesi çeşitliliğine bağımlılığı en düşük seviyeye indirmek için, öznitelik derecelendirme işleminde verinin farklı alt kümelerini kullanan yerel uzmanlar/öğreniciler kullanılmıştır.Bu amaçla veri eğitim ve test olmak üzere iki kümeye bölünmüş ve her rastgele bölümlemede, öznitelikleri derecelendirmek amacıyla örneklerin % 60'ı yerel öğrenciler tarafından eğitim kümesi olarak kullanılmıştır. İstatistiksel olarak önemli ve doğru sonuçlar elde edebilmek amacıyla bölme ve derecelendirme işlemleri N kez tekrarlanmıştır. En sonunda, öznitelik seçimi için N kez derecelendirilmiş sonuçlar kullanılmıştır. Farklı eğitim kümelerinikullanan yerel uzmanları birleştirerek, deney sonuçlarımızın kanıtladığı gibi, veri kümesi çeşitliliğinin iyi bir temsili elde edilmiştir. Bu yaklaşım 3 kanser veri kümesi üzerinde uygulanmış ve sonuçlar bulgular bölümünde raporlanmıştır.

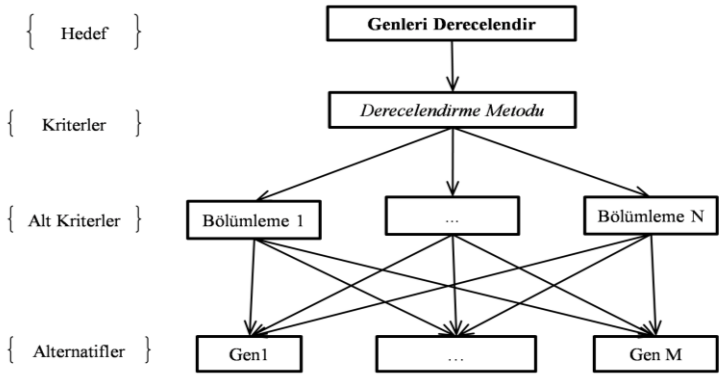
Hipertansiyon veri seti üzerindeki uygulama kanser verilerine uygulanan yaklaşıma benzerdir. Ancak burada farklı bir bölümle yöntemi uygulanmıştır. LOO çapraz geçerleme ile oluşturulmuş N=159 (örnek sayısı) altkümeyi kullanan yerel uzmanlar/öğreniciler ile öznitelikler derecelendirilmiştir. İşlem sonunda her öznitelik için N adet skor elde edilmiştir. PO yaklaşımı doğası gereği, en iyi-k ile ilgilenmeksizin birçok kriteri dikkate alarak seçim yapmaktadır. PO yönteminin çok amaçlı karar vermedeki gücü kullanılarak hastalıkla ilişkili genler seçilmiştir. Seçim sonuçları ve seçilen genlerin başarısı bulgular bölümünde gösterilmiştir.

3.7.2Öznitelik Derecelendirme için Yerel Uzmanların Analitik Hiyerarşi Proses Yaklaşımı ile Birleştirilmesi

Öznitelik seçimi, seçim yöntemi ve veri kümesi çeşitliliğine bağlı olduğu gibi aynı şey öznitelik derecelendirme için de geçerlidir. Çünkü derecelendirme yöntemleri, öznitelikleri seçmeden önce onları sıralar ve skorlarına göre bir altküme seçerler. Bunun yanında, yöntemler, her veri kümesi değişiminde farklı önem sırası oluşturur. Sıralama problemi, çok kriterli bir karar verme problemi olarak tanımlanırsa, farklı sıralama skorlarından bir ortak skor elde edilebilir. Bu ortak skorun sıralaması, özniteliklerin yeni bir önem sıralamasıdır. Bu sebeple, AHP ortak skor üretmek ve öznitelikleri önceliklendirmek amacıyla kullanılabilir.

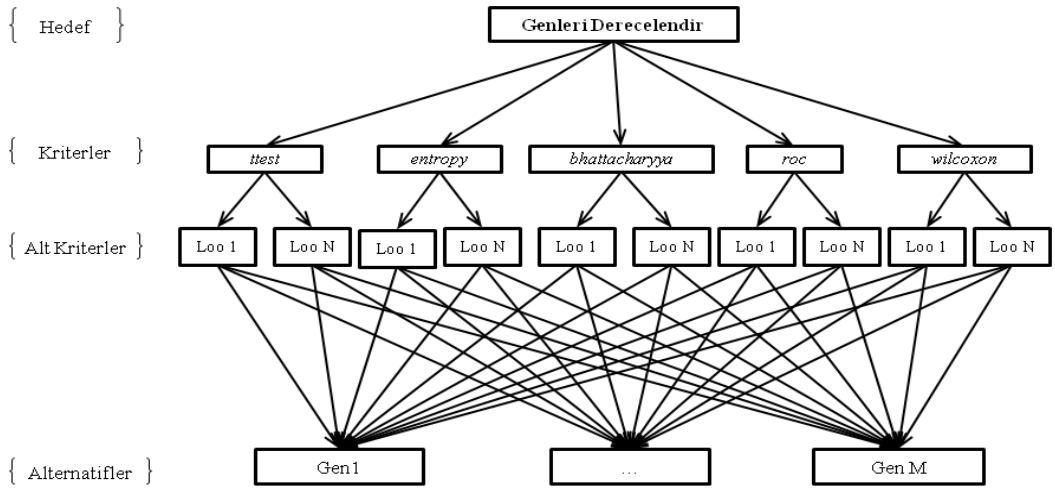
AHP tabanlı derecelendirme kanser veri kümelerine ve hipertansiyon veri kümesine farklı stratejiler ile uygulanmıştır. Takip eden 2 alt bölümde her iki kullanım için açıklamalar yer almaktadır.

Kanser veri kümelerinde, AHP tabanlı derecelendirme yönteminde kriterimiz, N adet kriteri olan derecelendirme skorudur . Skorlar, verinin N farklı alt uzayını kullanan yerel uzmanlar tarafından oluşturulmuştur. AHP hiyerarşi ağacında, alternatiflerin altkriterler bazında ikili karşılaştırılması için, çoklu eğitim kümesinden elde edilen dereceler kullanılır. Alt kriterleri karşılaştırmak için, herhangi bir uzman görüşü kullanılmaz. Tüm kriterler, eşit önceliğe sahiptir dolayısıyla her kriter için varsayılan öncelik $1/N$'dir. **Şekil 3.12**, çoklu eğitim kümelerinden elde edilen N adet alt kriteri olan AHP hiyerarşi ağacını göstermektedir. Alt kriter sayısı N=100 (100 adet küme bölümü) 'dür. Ayrıca her veri kümesinin alternatif gen sayısı (M) PO ile seçilen gen sonucu olacak şekilde**Tablo 4.6**'de gösterilmiştir.



Şekil 3.12: Kanser Verilerinde Kullanılan Ahp Hiyerarşi Ağacı

AHP tabanlı derecelendirme metodunun hipertansiyon veri kümesine uyarlamasında ise derecelendirme metotları kriterler ve LOO çapraz geçerleme ile oluşturulmuş öğrenme kümelerini kullanan yerel öğrenicilerin ürettiği derecelendirme değerleri alt kriterler olarak seçilmiştir. Her bir derecelendirme metodunun ve yerel öğrenicinin eşit öneme sahip olduğu varsayılmıştır. Dolayısı ile kriterlerin/alt kriterlerin birbirine karşı üstünlüğü tanımlanmamıştır. Hiyerarşi ağacında,kriterler bazında, genlerin LOO çapraz geçerlemeden elde edilen derecelendirme oranları ikili karşılaştırma matrislerini oluşturmak için kullanılmıştır. Derecelendirme metotlarından oluşmuş 5 kriter, LOO çapraz geçerlemeden oluşmuş 5xN alt kriter ve genlerden oluşmuş M=26(PO ile 5 yöntem için seçilen genlerin birleşimi) alternatife sahip hiyerarşi ağacı**Şekil 3.13**'de görülmektedir.



Şekil 3.13: Hipertansiyon Verisinde Kullanılan Ahp Hiyerarşi Ağacı

3.7.3Çok Amaçlı Öznitelik Seçimi için Metotların Birleştirilmesi

Öznitelik seçiminin seçim metotlarından da etkilendiğini birçok çalışma göstermiştir[3].Bu çalışmada veri seti değişikliğinden etkilenmeyen öznitelik seçim yöntemlerin yanısıra,farklı metotları birleştiren çok kriterli öznitelik seçme ve değerlendirme yöntemleri önerilmiştir. Vaka çalışması olarak HGDP ve Hipertansiyon veri kümeleri kullanılmıştır.

Hipertansiyon verisine 5 farklı seçim yöntemi uygulanmış ve herbir metot tarafından seçilmiş genlerin birleşim kümesi nihai gen kümesi olarak kabul edilmiştir. HGDP verisinde ise farklı metotlarca seçilen özniteliklerin birleşimi yerine, tüm metotların sonuçlarını dikkate alarak metotları birleştiren çok kriterlibir öznitelik seçim metodu önerilmiştir [49, 60]. Önerilen yöntemin sonuçları bölüm 4'te sunulmuştur. Sonuçlar yöntemin başarısını açıkça göstermektedir.

Yapılan 2 ayrı çalışmada, HGDP'den seçilen 12 etnik grup, çok amaçlı SNP seçimi için kullanılmıştır. Bu çalışmalardaki temel hedef, farklı amaçların eş zamanlı olarak dikkate alınmasını sağlayacak bir birleştirme işlemi yapmaktır. Amaç 1 olarak yüksek sınıflandırma doğruluğu tanımlanmıştır. İki popülasyon arasındaki genetik varyasyonların, coğrafî uzaklıkla orantılı bir şekilde artması [47] nedeni ile 2. amaç olarak, grupların genomik ve coğrafî uzaklıkları arasındaki korelasyonun maksimum olması tanımlanmıştır. Ardından her iki amaç Pareto Optimal yöntemi ile maksimize edilmiştir.

Bir SNP'nin amaç 1'e katkısını değerlendirmek için, MI ve Relieff değeri ele alınmıştır. Amaç 2 için, en iyi iki temel bileşenden gelen SNP ağırlıkları kullanılmıştır ve bu sayede genetik-coğrafî mesafelerin korelasyonu açıklanmaya çalışılmıştır [48]. Diğer bir deyişler, birinci ve ikinci temel bileşenler (PC'ler) ele alınarak her SNP için karşılık gelen yük kullanılmış Amaç 2 başarılı bir şekilde ifade edilmiştir. Amaç 2, çok kriterli (2 boyutlu -PC1 ve PC2) bir problem olarak dizayn edilmiştir. Eş zamanlı olarak Amaç 1'in sağlanması için de her SNP'nin MI ve Relieff değerleri seçim işlemine 3. ve 4. boyut olarak eklenmiştir. Böylece farklı öznitelik seçim metotları tarafından oluşturulan MI, Relieff, PC1, PC2 skorları kullanılarak çok kriterli bir seçim modeli elde edilmiştir. Çok kiriterli yaklaşımla SNP seçimi için PO yöntemi kullanılmıştır. Birden fazla kriteri

birleştiren çok kriterli bir yaklaşımınkullanılması her iki amaca da hizmet edecek şekilde SNP seçimi yapılmasını sağlamıştır.

4. BULGULAR

4.1 ÖZNİTELİK SEÇİMİNDE BAĞIMLILIKLARIN GÖSTERİLMESİ

Bilindiği üzere, öznitelik seçimi için farklı kısıtları dikkate alan birçok farklı yöntem önerilmiştir. Farklı çözüm yöntemleri iseverilen bir veri kümesi için, farklı özniteliklerin ya da öznitelik gruplarınınseçilmesine neden olmaktadır. Bu özellik, öznitelik seçim işlemininkullanılan yöntemebağımlı olduğunu gösterir. Diğer taraftan aynı öznitelik seçim algoritması, birbirinden çok çok az farklı olan iki veri kümesine uygulansa bile farklı özniteliklerin seçilmesi mümkün olabilmektedir. Bu husus ise, öznitelik seçim işleminin kullanılan veri kümesine de bağımlı olduğuna işaret eder.

Bu çalışmada, literatürde yer alan öznitelik seçim işleminin veri kümesideğişimineve seçim yönteminebağımlılığı örnek bir çalışma üzerinde gösterilmiştir. Bu amaçla dizayn ettiğimiz çalışmamızda, çapraz geçerleme ile bölümlenen veri kümesinden elde edilen farklı öğrenme kümelerinin farklı öznitelikleri seçtiği gözlemlenmiştir. Bununla birlikte, aynı eğitim kümesini kullanan farklı metotların da farklı alt kümeler seçtiğigözlemlenmiştir.

Bu bulgular, hipertansiyon veri kümesi üzerinde bir vaka çalışması yapılarak doğrulanmıştır. İncelemede, k-katlı (5 ve 10) çapraz geçerleme ve LOO yöntemleri kullanılarak, mikrodizi verisinden hem tek gen seçimi hem de bir grup gen(10 gen) seçimi için çeşitli testler yapılmıştır.

LOO çapraz geçerleme yöntemikullanılarakher seferinde bir örnek dışarıda tutulup, aynı öznitelik seçim algoritmasının en açıklayıcı geni seçmesi durumu incelendiğinde her tekrarda farklıbir genin seçildiği gözlenmiştir (**Tablo 4.2**). Benzer şekilde, K-kat çapraz geçerleme yöntemi kullanılarak her kat dışarıda bırakıldığında veya k farklı değerler aldığında da aynı öznitelik seçim algoritmasının farklı genleri seçtiği tespit edilmiştir (**Tablo 4.3** ve**Tablo 4.4**). Aynı işlem, birden fazla gen seçimi hedefi için 5 ve 10 kat çaprazlama yöntemlerii kullanılarak tekrarlanmıştır ve her kat dışarda tutulduğunda yine farklı gen grublarının seçildiği görülmüştür (**Şekil 4.1**).Bu testlerden, aynı öznitelik seçim yönteminin her tekrarda seçtiği en tahmin edici gen/gen grupları arasında anlamlı

bir örtüşme olmadığı anlaşılmıştır. Sonraki alt bölümlerde, gen seçimi için yapılan testlerin detayları verilmiştir.

Yapılan testlerde, 10 gen seçimi için hem mevcut CD, mRMR yöntemleri hem önerdiğimiz MI-d, MI-c yöntemleri kullanılırken, tekli gen seçimi için ise mRMR yöntemi hariç sözüedilen diğer yöntemler kullanılmıştır. Bilindiği gibi, tekli gen seçimi problemi açısından mRMR yöntemiMI yöntemine indirgenebilmektedir. Bu nedenle, tek gen seçimi için sadece üç yöntem (CD, MI-d, MI-c) kullanılmıştır.Tablolarda seçilen genler için, gen sembolü, gen numarası (çalışma[36] te belirtilen orjinal gen ID'leri), seçim sıklığı, CD ve MI skorları ve sınıflandırma başarısı (AC) gösterilmiştir. Burada, AC değeri, her genin k-en yakın komşu sınıflandırma yöntemi (k'nın farklı değerleri için) kullanılarak veri setini ne kadar doğru sınıflandırdığını göstermektedir. AC skoru 10-kat çapraz geçerleme kullanılarak verilmiştir

4.1.1 LOO Yöntemi ile Tek Gen Seçimi

Bu çalışmada veri kümesi bölümlemenin öznitelik seçimi üzerindeki etkisini gözlemlemek amacı ile ilk olarak LOO yöntemi kullanılmıştır. Bu yöntem ile yapılan tekli gen seçimi işleminin bulguları **Tablo 4.1**ve**Tablo 4.2**'de listelenmiştir. Veri kümesi bölümleme toplam örnek sayısı kadar (159 kez) tekrarlanmıştır. Her bölümleme sonunda 3 farklı seçim algoritması uygulanmıştır. Buna göre, 44 tane tekil gen içeren, toplamda 159*3=477 gen seçilmiştir. Seçilen genler arasındaki NULL ile sembolize edilmiş 17 genin fonksiyonu tam olarak bilinmemektedir. MI, CD, MI-c öznitelik seçim algoritmaları tarafından seçilen genlerin, k-en yakın komşu sınıflandırma algoritması kullanılarak farklı k değerleri için bulunan ortalama sınıflandırma başarıları **Tablo 4.1**'de; genlerin seçilme sıklıkları ve sınıflandırma başarıları **Tablo 4.2'** de ayrıntılı olarak gösterilmiştir.

Tablo 4.1: LOO Tek Gen Seçim Başarısı

Metot	Doğruluk(AC)				
	k=1	k=3	k=5	k=7	k=9
MI	0.35	0.33	0.28	0.34	0.28
CD	0.50	0.50	0.50	0.50	0.54
MI-c	0.38	0.40	0.37	0.34	0.28

Tablo 4.2: LOO Tek Gen Seçimi

Sembol	Gen ID	Sıklık				Doğruluk(AC)				
		Toplam	MI-d	CD	MI-c	k=1	k=3	k=5	k=7	k=9
MIST	11249	149	0	149	0	0.44	0.47	0.45	0.50	0.52
Null	10973	82	81	0	1	0.25	0.26	0.26	0.29	0.26
Null	13491	35	35	0	0	0.09	0.09	0.08	0.10	0.09
Null	13162	27	27	0	0	0.08	0.07	0.08	0.07	0.08
Null	16723	20	0	0	20	0.06	0.05	0.06	0.06	0.04
SRGAP2	13214	19	0	0	19	0.05	0.05	0.06	0.03	0.03
Null	5179	15	0	0	15	0.06	0.05	0.05	0.04	0.03
Null	14857	10	10	0	0	0.05	0.05	0.05	0.05	0.05
FGF5	6449	10	0	0	10	0.03	0.03	0.03	0.02	0.02
MRPS2	15582	9	0	0	9	0.04	0.03	0.03	0.03	0.03
Null	3122	8	0	0	8	0.03	0.02	0.03	0.03	0.03
TFAM	15231	7	0	0	7	0.03	0.02	0.01	0.03	0.03
LOC119392	14382	7	0	0	7	0.03	0.02	0.03	0.01	0.03
Null	10621	7	0	7	0	0.02	0.02	0.02	0.02	0.02
MGC20781	5109	7	0	0	7	0.04	0.04	0.03	0.03	0.04
FLJ31413	12061	6	0	0	6	0.02	0.02	0.01	0.01	0.01
Null	15764	5	0	0	5	0.01	0.03	0.03	0.03	0.03
Null	9549	5	0	0	5	0.02	0.02	0.02	0.02	0.02
MRPL9	365	5	0	0	5	0.01	0.01	0.01	0.01	0.01
NDUFV1	2024	4	0	0	4	0.01	0.00	0.01	0.01	0.00
SLC30A9	1676	4	0	0	4	0.02	0.02	0.01	0.01	0.01
Null	9609	3	0	0	3	0.00	0.00	0.00	0.00	0.00
MGC16153	9352	3	0	0	3	0.01	0.00	0.00	0.00	0.00
ERAL1	9275	3	0	0	3	0.01	0.02	0.02	0.02	0.01
SEC61B	418	3	0	0	3	0.01	0.02	0.01	0.01	0.01
Null	17676	2	0	0	2	0.00	0.01	0.01	0.01	0.01
Null	15551	2	2	0	0	0.01	0.01	0.01	0.01	0.01
RBMY1A3P	11710	2	0	0	2	0.01	0.01	0.01	0.01	0.01
ARRB1	7942	2	0	0	2	0.00	0.01	0.01	0.01	0.01
Null	2782	2	0	2	0	0.01	0.01	0.01	0.01	0.01
MGC3234	21624	1	0	0	1	0.01	0.00	0.01	0.00	0.00
COH1	18721	1	0	0	1	0.01	0.00	0.00	0.00	0.00
LOC285286	15434	1	0	0	1	0.01	0.01	0.01	0.01	0.00
FLJ37078	15401	1	0	0	1	0.00	0.00	0.00	0.00	0.00
EPS8L3	14698	1	1	0	0	0.00	0.00	0.00	0.00	0.00
NR4A1	12822	1	0	0	1	0.00	0.00	0.00	0.00	0.00
RFX2	12376	1	0	1	0	0.01	0.00	0.00	0.00	0.00
LYPLA2	3870	1	1	0	0	0.01	0.01	0.01	0.01	0.01
Null	3506	1	0	0	1	0.00	0.01	0.01	0.00	0.01
MGC35295	3283	1	0	0	1	0.01	0.01	0.01	0.01	0.01
Null	3038	1	0	0	1	0.00	0.00	0.00	0.00	0.00
Null	1902	1	0	0	1	0.00	0.00	0.00	0.00	0.00
CKMT1	1000	1	1	0	0	0.00	0.00	0.00	0.00	0.00
KRTAP20-1	686	1	1	0	0	0.00	0.00	0.00	0.00	0.00

4.1.2 K-kat Çapraz Geçerleme ile Tek Gen Seçimi

Çalışmada kullanılan diğer Veri Kümesi Bölümleme yöntemi olarak 5-katçapraz geçerleme ile 3farklı öznitelik seçim yöntemi birlikte kullanılarak oluşturulan 15 farklı analiz sonucunda, her bir analizin farklı genleri seçtiği **Tablo 4.3**'degösterilmiştir.**Tablo 4.4**'de ise, 10-kat çapraz geçerleme ile elde edilen tek gen seçim sonuçları da her bir analizin farklı genleri seçtiği ve ortak seçilen bir gen olmadığı görülmektedir. Benzer şekilde, 10-kat çapraz geçerleme ile 3 farklı öznitelik seçim yöntemi birlikte kullanılarak yapılan toplam 30 analiz sonucunda, 26'sı yalnızca 1 kez seçilen, 28 tekil gen bulunmuştur.

Tablo 4.3: 5-Kat Çapraz Geçerleme İle Tek Gen Seçimi

Sembol	Gen ID	Sıklık
Null	10973	1
Null	13162	1
Null	14857	1
LOC285692	15235	1
FLJ35848	18162	1
SEC61A2	1711	1
Null	2782	1
KIAA1729	6582	1
RAB9A	10297	1
MIST	11249	1
RFX2	12376	1
Null	14199	1
LOC151475	16881	1
Null	17936	1
K-ALPHA-1	21453	1

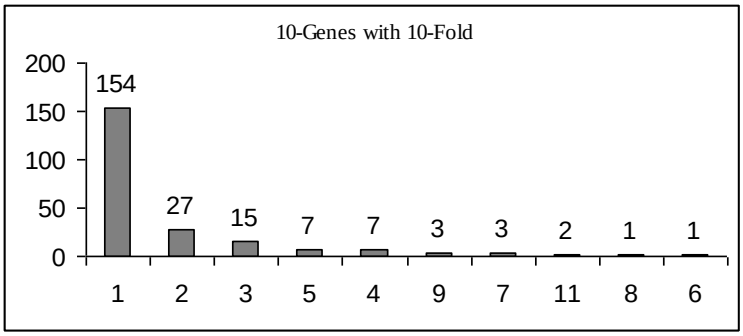
Sonuç olarak, yapılan inceleme, gen seçimlerinin veri kümesine ve modele bağımlı yapısını net olarak ortaya koymuştur. Tüm bulgular tek bir veri kümesi bölümlemeyöntemi ve tek bir öznitelik seçme yöntemi uygulanarak doğru sonuçlar elde edilemeyeceğini, belirli bir yöntemi aynı öğrenme kümesi ile kullanmanında kararsız sonuçlar doğuracağını göstermiştir.

Tablo 4.4: 10-Kat Çapraz Geçerleme İle Tek Gen Seçimi

Sembol	Gen ID	Sıklık
MIST	11249	2
ARPP-21	17392	2
MRPL9	365	1
SEC61B	418	1
Null	2782	1
MPP4	3734	1
Null	4678	1
Null	5667	1
FGF5	6449	1
Null	8103	1
MUC2	9437	1
Null	10621	1
Null	10973	1
DPYSL4	11551	1
FLJ31413	12061	1
KALI	12776	1
Null	14199	1
EPS8L3	14698	1
FLJ36601	14800	1
Null	14857	1
FLJ37078	15401	1
LOC285286	15434	1
Null	15551	1
MGC2941	15558	1
FLJ33084	16524	1
Null	16723	1
CPSF1	20187	1
DSCR3	22026	1

4.1.3 10-kat Çapraz Geçerleme ile 10 Gen Seçimi

Önceki bölümlerde (Bölüm 4.1.1 - 4.1.2) tek gen seçiminde veri kümesi ve seçim yöntemine bağımlılık dolayısı ile kararsız sonuçlar elde edilmiştir. Aynı bağımlılığın bir grup gen seçiminde de (10 gen) ortaya çıkıp çıkmayacağını görmek amacıyla MI, CD, MI-c, mRMR analiz yöntemi ve 10-kat çapraz geçerleme yöntemi kullanılarak gen seçimleri gerçekleştirilmiştir. Bu amaçla, her biri 10 gen içeren 40 öznitelik altkümesi elde edilmiştir.Yapılan analizlerde yalnızca 2 genin 11 kere seçildiği, 154 genin ise yalnızca 1 kez seçildiği görülmüştür. **Şekil 4.1**'de verilen grafikte, x-ekseni 400 seçilmiş gen üzerinden, bir genin seçilme sıklığını, y-ekseni ise kaç tane genin bu sıklıkla seçildiğini göstermektedir.



Şekil 4.1: 10-Kat Çapraz Geçerleme İle, 10-Gen Seçim Histogramı

Yaptığımız deneyler tüm veri kümesinin bir örneğinin dışarıda bırakılıp, diğerinin dahil edilmesi ile oluşturulan neredeyse özdeş altkümelerinde bile gen seçim yöntemlerinin farklı seçimler üretebileceğini göstermiştir. Tek gen seçim işleminde, en az 2 analiz tarafından seçilmiş genler **Tablo 4.5** listelenmiştir. Farklı veri kümeleri ve farklı metotların farklı genler seçmesi tek bir yöntem yada tek bir eğitim kümesi ile elde edilen sonuçların güvenilmez olduğunu ortaya koymuştur. Bu problemin en basit çözümü tüm analizlerin birleşim kümesini almaktır. Sonuç olarak, en çok ortaklaşa seçilen genleri ıslak laboratuvar doğrulaması için aday olarak tanımlamak amacıyla çeşitli analiz yöntemlerini kullanmanın daha uygun olacağı düşülmüştür.

Tablo 4.5: Ortaklaşa Seçilen Genler

Sembol	GenID	Sıklık	CD	MI-d	MI-c	Doğruluk (%) ± Standart Sapma				
						k=1	k=3	k=5	k=7	K=9
Null	14857	3	0,01	0,12	0,05	53 ± 0,08	54 ± 0,08	58 ± 0,09	64 ± 0,10	63 ± 0,08
MIST	11249	3	0,08	0,09	0,09	55 ± 0,09	52 ± 0,11	53 ± 0,11	53 ± 0,13	56 ± 0,11
Null	10973	3	0,03	0,12	0,10	57 ± 0,08	60 ± 0,08	58 ± 0,06	61 ± 0,08	59 ± 0,05
Null	2782	3	0,07	0,10	0,08	52 ± 0,09	52 ± 0,08	55 ± 0,11	55 ± 0,09	53 ± 0,09
Null	16723	2	0,01	0,11	0,06	55 ± 0,09	58 ± 0,18	59 ± 0,16	58 ± 0,15	65 ± 0,09
Null	15551	2	0,00	0,11	0,07	61 ± 0,15	60 ± 0,16	68 ± 0,09	72 ± 0,12	67 ± 0,14
LOC285286	15434	2	0,03	0,04	0,11	58 ± 0,14	63 ± 0,13	63 ± 0,08	59 ± 0,11	56 ± 0,13
FLJ37078	15401	2	0,01	0,03	0,06	55 ± 0,08	64 ± 0,06	61 ± 0,07	61 ± 0,11	61 ± 0,15
EPS8L3	14698	2	0,00	0,12	0,05	58 ± 0,09	59 ± 0,11	59 ± 0,11	60 ± 0,14	62 ± 0,16
Null	14199	2	0,07	0,08	0,08	56 ± 0,11	55 ± 0,07	58 ± 0,09	61 ± 0,06	54 ± 0,10
Null	13162	2	0,04	0,12	0,08	54 ± 0,10	57 ± 0,10	55 ± 0,08	57 ± 0,06	60 ± 0,10
RFX2	12376	2	0,07	0,06	0,06	55 ± 0,11	51 ± 0,12	50 ± 0,12	57 ± 0,11	52 ± 0,08
FLJ31413	12061	2	0,00	0,04	0,13	54 ± 0,08	62 ± 0,13	58 ± 0,13	62 ± 0,13	62 ± 0,14
Null	10621	2	0,08	0,06	0,07	48 ± 0,12	53 ± 0,17	55 ± 0,12	59 ± 0,14	54 ± 0,12
FGF5	6449	2	0,02	0,07	0,10	63 ± 0,11	64 ± 0,08	59 ± 0,10	64 ± 0,12	63 ± 0,07
SEC61B	418	2	0,05	0,12	0,12	49 ± 0,12	59 ± 0,11	57 ± 0,07	63 ± 0,09	60 ± 0,11
MRPL9	365	2	0,00	0,09	0,09	48 ± 0,12	51 ± 0,12	54 ± 0,10	54 ± 0,11	61 ± 0,11

4.2 KANSER SONUÇLARI

Bölüm 4.1'de gösterilen öznitelik seçim yönteminin veri kümesinin değişiminden etkilenmesini minimize etmek amacı ile bu çalışmada PO&AHP yöntemlerine dayanan yeni bir karma yöntemönerilmiştir. Önerilen karma yöntem, hastalık-ilişkili genlerin seçim ve derecelendirme başarısını artırmak için, en çok tercih edilen beş öznitelik derecelendirme metodu (*ttest, entropy, bhattacharyya, roc, wilcoxon*) ile birlikte kullanılmıştır. Yeni yöntem öznitelik seçimi için çoklu öğrenme kümesi kullanmaktadır. Bu nedenle veri kümesi %60 öğrenme, %40 test kümesi olacak şekilde 100 kez bölümlenmiştir. Tüm öznitelikler 100 öğrenme kümesi kullanılarak derecelendirilmiş ve 100×*f*(*f*: öznitelik sayısı) boyutlarında bir derecelendirme matrisi oluşturulmuştur. Her rastgele bölümlemede, tüm derecelendirme yöntemleri tamamen aynı bölünmüş veri kümesini kullanmıştır.

Önerilen karma yöntemde, gen altkümeleri seçme amacı ile öznitelikleri tüm kriterler uzayında birbiri ile yarıştıran PO yaklaşımı uygulanmıştır.PO yöntemi, tüm değerlendirme kriterleri açısından en az bir optimum değere sahip olan yani diğer genler tarafından bastırılamayan tüm genleri seçilmiş gen kümesine dahil etmektedir. **Tablo 4.6**'de, çoklu eğitim kümeleri ile gerçekleştirilen, beş derecelendirme yönteminin herbiri ile birlikte PO yaklaşımı kullanılarak elde edilen gen altküme seçimlerinin sonuçları gösterilmektedir.

Tablo 4.6: Pareto Optimal Kümedeki Bastırılamayan Genlerin Sayısı

Yöntem/Veri kümesi	# Gen		
	Kolon	Duke	DLBCL
T-test	27	124	10
Entropy	85	80	100
Bhattacharyya	28	161	6
Roc	37	102	35
Wilcoxon	57	102	30

PO ile seçilen gen alt kümesindeki tüm genler eşit öneme sahip olduğundan bu genlerin herhangi bir önceliksıralaması yoktur. Bu problemin çözümü için, AHP yöntemi PO yaklaşımı ile birleştirilerek karma bir yöntem elde edilmiştir. Bu amaçla, PO tarafından seçilen genler AHP ile derecelendirme (önceliklendirme) işlemine tabi tutulmuştur. AHP ile sıralamanın sonucunda eğer genlerin öncelikleri varsayılan öncelikten daha büyükse, bu genler birinci derecede önemli genler; genlerin öncelikleri varsayılan öncelikten daha küçükse, bunlar da ikinci derecede önemli genler olarak kabul edilirler.

Tablo 4.7'de, her veri kümesi ve her derecelendirme yöntemi için; varsayılan öncelik değeri (VÖ=1/alternatif sayısı), varsayılan öncelikten daha düşük (#K) ve daha yüksek olan (#B) genlerin sayıları verilmiştir.

Tablo 4.7: AHP Önceliklendirmesine Göre Seçilen Genlerin Sayısı

Veri Kümesi/ Yöntem	Kolon			Duke			DLBCL		
	VÖ	#K	#B	VÖ	#K	#B	VÖ	#K	#B
T-test	0,04	14	**13**	0,01	64	**60**	0,1	4	**6**
Entropy	0,01	46	**39**	0,01	60	**20**	0,01	78	**22**
Bhattacharyya	0,04	14	**14**	0,01	94	**67**	0,17	1	**5**
Roc	0,03	16	**21**	0,01	50	**52**	0,03	11	**24**
Wilcoxon	0,02	28	**29**	0,01	50	**52**	0,03	13	**17**

varsayılanöncelik (VÖ)'den daha küçük (K) vedaha büyük (B) değere sahip genlerin sayısı

Bu çalışmada önerilen karma yaklaşımı mevcut bazı yöntemlerle karşılaştırmak için, **Tablo 4.9**'dailgili çalışmaların[6, 10, 37] kolon veri kümesiüzerindeki deneysel sonuçları ve bu çalışmada önerilen derecelendirme+PO&AHP karma yönteminsonuçları özetlenmiştir.

Yaygın kullanılan mRMR [37] öznitelik seçim yöntemi, MID ve MIQ olmak üzere iki alt yönteme sahiptir. Çalışmada, seçilen genlerin sınıflandırma başarısının değerlendirmesi için Naive Bayes sınıflandırıcı LOO çapraz doğrulama tekniği kullanılmıştır. MID ve MIQ yöntemlerinin 10 gen için en yüksek doğruluk oranları sırasıyla % 88.7 ve %93.5 olarak rapor edilmiştir. Bu çalışmada önerilen karma yöntem ile elde edilen doğruluk değerleri ise dört derecelendirme yöntemi için sırasıyla 2,4,6 gen ile %91.9'dur.

Tablo 4.8'de mRMR ve derecelendirme+PO&AHP karma yöntemi beş derecelendirme yöntemi için detaylı bir şekilde karşılaştırılmıştır. Burada görülen doğruluk değerleri, mRMR [37]çalışmasında verildiği gibi LOO çapraz doğrulama ve NaiveBayes sınıflandırıcılar kullanılarak ile elde edilmiştir. Sonuçlar incelendiğinde, PO&AHP ile kullanılan basit derecelendirme yöntemlerinin başarısının, bir wrapper yöntem olan mRMR'den daha iyi olduğu görülmüştür. Öte yandan, PO&AHP'nin mRMR ile birleştirilmesi halinde, mRMR'nin de performansının artacağı düşünülmektedir.

Tablo 4.8: Kolon Veri Kümesi için Deneysel Sonuçların Mrmr ile Karşılaştırılması

Metot/ Gen Sayısı	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
mRMR_MID	0.839	0.871	0.871	0.871	0.855	0.839	0.855	0.871
mRMR_MIQ	0.839	0.871	0.806	0.871	0.871	0.903	0.903	**0.919**
T-test +PO&AHP	**0.855**	**0.919**	**0.903**	0.903	**0.903**	0.903	**0.919**	**0.919**
Entropy +PO&AHP	**0.855**	0.871	**0.903**	**0.919**	**0.903**	0.903	0.903	**0.919**
Bhattacharyya+PO&AHP	**0.855**	0.855	**0.903**	0.903	**0.903**	**0.919**	**0.919**	**0.919**
Roc+PO&AHP	0.823	0.855	**0.903**	0.903	**0.903**	**0.919**	**0.919**	**0.919**
Wilcoxon +PO&AHP	0.774	0.806	0.839	0.871	0.871	0.855	0.871	0.871
Metot/ Gen Sayısı	**10**	**12**	**15**	**20**	**30**	**40**	**50**	
mRMR_MID	0.887	0.887	0.887	0.871	0.887	0.887	0.887	
mRMR_MIQ	0.935	0.919	0.887	0.887	0.871	0.871	0.887	
T-test +PO&AHP	0.903	0.903	0.903	0.887	-	-	-	
Entropy +PO&AHP	0.919	0.919	0.919	0.903	0.887	0.903	0.919	
Bhattacharyya+PO&AHP	0.919	0.919	0.903	0.903	-	-	-	
Roc+PO&AHP	0.903	0.887	0.887	0.919	0.903	-	-	
Wilcoxon +PO&AHP	0.871	0.887	0.887	0.887	0.887	0.887	0.887	

Önerilen karma yöntemi literatürdeki çalışmalarla karşılaştırmak amacı ile kullandığımız diğer bir çalışma [10], kolon veri kümesi için SNR, FFS1, FFS2, SVM-RFE olmak üzere dört ileri doğru öznitelik seçim yöntemi kullanmıştır. Seçilen öznitelik kümesinin başarısının değerlendirilmesi bölümünde tartışıldığı üzere bir sınıflandırma modelinin başarısı, veri kümesi bölümlemesine bağımlıdır. Bu bağımlılıktan kaçınmak için, Luo ve diğ. [10] tarafından önerildiği gibi veri kümesi 100 kez (%60 eğitim, %40 test) bölünmüş ve seçilen gen kümelerinin sınıflandırma başarısını hesaplamak için rastgele 100 bölümün başarısının ortalaması kullanılmıştır. Luo ve diğ.SVM sınıflandırıcı kullanılmıştır. Bu çalışmada bildirilen doğruluk ise K-NN (k=3) sınıflandırıcısı ile elde edilmiştir

Luo ve diğ.çalışmasında, 20, 2, >80 ve >80 gen kullanılarak elde edilen en yüksek doğruluk değerleri sırasıyla dört yöntem için ve %78.5, %81.4, %80.9 ve %81.4 olarak rapor edilmiştir[10]. Bu çalışmada elde edilen en düşük doğruluk değeri çalışma [10]'nun en yüksek doğruluk değerini geçmiştir (*wilcoxon*+PO&AHP için 30 genle %86.6). Bu çalışmada elde edilen en yüksek doğruluk (*entropy*+PO&AHP için 30 genle %91) Luo ve diğ.'nin çalışmasında elde edilen en yüksek doğruluk değerinden (%81.4) %10 oranında daha yüksektir. Bu çalışmanın sonuçları ile Duke ve DLBCL veri kümelerinin karşılaştırılması **Tablo 4.10**'te verilmiştir.

Karşılaştırma amacı ile kullanılan bir diğer çalışma [6], ttest+SVM-RFE ile seçtiği genlerin, sınıflandırma başarımını test etmek amacı ile 10 kat çapraz geçerleme

kullanmış ve bunu 250 kez çalıştırarak ortalama doğruluğu yayınlamıştır. Çalışma[6]'da; 8,16 ve 32 genin seçilmesi halinde ortalama doğruluk deperleri sırasıyla %84.6, %85.3 ve %87.6 olarak sunulmuştur. Bu çalışmada, ttest yönteminin önerilen karma yöntem ile birleştirilmesi (ttest+PO&AHP) sonucunda seçilen11 gen ile elde edilen doğruluk %90.6'dır. Önerilen yöntemin başarısı, çok amaçlı öznitelik seçim yaklaşımı kullanarak, örnek alt kümelerindeki çeşitliliğin seçim işlemine etkisini azaltmaya dayanmaktadır.

Tablo 4.10, Duke ve DLBCL veri kümeleri üzerinde Luo ve diğ.[10]çalışması ile bu çalışmada elde edilen sonuçların karşılaştırılmasını göstermektedir. Duke veri kümesi için, Luo ve diğ.(2011) SNR, FFS1, FFS2, SVM yöntemleriile 100, 50, 100 ve 100 gen kullanarakrapor ettikleri doğruluk oranları sırasıyla % 84.2, %72, %84 ve %86.8'dır. Bu çalışmada elde edilen en yüksek doğruluk ise %95 (ttest+PO&AHP)olmuştur. Bu sonuç Luo ve diğ.(2011) elde ettiği en yüksek doğruluk miktarı olan %86.8'den %9 oranında kadar daha yüksektir.Benzer sonuçlar, önerilen yöntemin DLBCL veri kümesine uygulandığında da gözlemlenmiştir. Buna göre; DLBCL veri kümesinde SNR, FFS1, FFS2, SVM yöntemleri uygulandığında 20, 100, 100 ve 100 gen ile doğruluk oranları sırasıyla %85.2, %79.8, %90.7 ve %95.5 olarak rapor edilmiştir. Bu çalışmadan elde edilen deneysel sonuçlarda ise, beş farklı derecelendirme metodunun önerilen karma yöntemle kullanılması (derecelendirme+PO&AHP) sonucunda elde edilen doğruluk değerleri %87.6 ile %94.5 arasındadır.Buna ilaveten, Roc ve wilcoxon derecelendirme yöntemleri ile elde edilen yüksek doğruluk oranına(%94.5) sadece 3 gen seçerek ulaşılabilirken, Luo ve diğ.(2011) SVM-RFE yöntemi ile elde ettiği en yüksek doğruluk değerine (%95.5) ancak 100 gen seçerek ulaşılabilmiştir. Bu sonuçlardan, veri kümesindeki değişintiyi hesaba katan karma yöntemin, evrensel kümeyi başarı ile temsil edebilecek genleri seçtiği söylenebilir.

Tablo 4.9: Deneysel Sonuçların Kolon Veri Kümesi için Karşılaştırılması

(NaiveBayes- LOO CV)			
Yöntem		En yüksek AC – #E	#G
mRMR	mRMR_MID	0.887 – 7	10
	mRMR_MIQ	0.935 – 4	10
Derecelendirme+ PO&AHP	T-test	0.919 – 5	2
	Entropy	0.919 – 5	4
	Bhattacharyya	0.919 – 5	6
	Roc	0.919 – 5	6
	Wilcoxon	0.887 – 7	9
(100 kez %60-%40 bölme)			
Yöntem		En yüksek AC	#G
Luove diğ. (2011)	SNR	0.785	20
	FFS1	0.814	2
	FFS2	0.809	>80
	SVM-RFE	0.814	>80
Derecelendirme+ PO&AHP	T-test	0.906	11
	Entropy	0.910	17
	Bhattacharyya	0.908	13
	Roc	0.903	8
	Wilcoxon	0.866	30
Chenve diğ. (2007)	T-test+SVM-RFE (250 kez 10 kat çapraz)	0.876	32

#E yanlış sınıflandırılan örnek sayısını, #G seçilen gen sayısını gösterir.

Tablo 4.10: Deneysel Sonuçların Duke Ve Dlbcl Veri Kümeleri için Karşılaştırılması

Öznitelik Seçim Yöntemi		Duke		DLBCL	
		#G	AC	#G	AC
Derecelendirme+ PO&AHP	T-test	33	0.957	7	0.878
	Entropy	8	0.858	72	0.876
	Bhattacharyya	111	0.948	2	0.896
	Roc	9	0.943	3	0.945
	Wilcoxon	9	0.943	3	0.945
Luo ve diğ. (2011)	SNR	100	0.842	20	0.852
	FFS1	50	0.720	100	0.798
	FFS2	100	0.840	100	0.907
	SVM-RFE	100	0.868	100	0.955

#G seçilen gen sayısını göstermektedir.

Öznitelik seçim işleminin geliştirilmesi amacıyla önerdiğimiz PO&AHP yaklaşımı temelde, farklı veri alt kümelerin kullanan birçok seçim metodunu çok kriterli bir düzlemde birleştirmeye dayanmaktadır. Bu işlem, her biri genel öğrenme kümesinden alınmış yerel örnek kümesini kullanan ve ona göre işlem yapan yerel uzmanların birleştirilmesidir. Böylece oluşturulan öğrenme modellerinin birbirini tamamlaması sağlanarak toplamda model başarısı arttırılır [4].

Bu bölümde, yerel uzmanların PO&AHP karma yaklaşımı ile birleştirilmesi ile uzmanların ortalaması alınarak birleştirilmesinin karşılaştırması sunulmuştur. Ayrıca, yerel uzmanların birleştirilmesi yanında tüm veriyi tek bir kez kullanan genel uzmanın başarısı da karşılaştırmaya dâhil edilmiştir. Sonuç olarak karşılaştırma sonuçlarında üç farklı uzman yer almaktadır:

i. **Topluluk+PO&AHP:** her bir yerel uzmanın sonuçlarının PO ile birleştirilerek gen seçimin yapılması ve daha sonra AHP ile seçilen genlerin sıralanmasını içeren modeli temsil eder.
ii. **Topluluk**: her bir yerel uzmanın sonuçlarının ortalaması alınarak gen seçiminin yapılmasını içeren modeli temsil eder.
iii. **Tüm veri,tek seçim:**Veri kümesinin tamamını kullanarak gen seçiminin yapılmasını içeren genelmodeli temsil eder.

Bu tarz bir karşılaştırma için, her seferinde veriden 2 örnek çıkarılarak veri kümesinin farklı varyasyonları elde edilmiştir. Buradaki hedef, farklı örnek kümelerinde yöntemlerin durumunu ortaya koymaktır. Çünkü her uygulamada başarılı olacak bir model yoktur. Ancak yapılacak birçok deneme içerisinde, modellerin başarılı olma oranı karşılaştırma amacı ile kullanılabilir [4].

Modeller3 açıdan karşılaştırılmıştır:

i. **Tek gen**: her bir modelin tek gen seçmesi durumu.
ii. **PO tarafından seçilen gen kadar:** PO yöntemi doğası gereği, dikkate aldığı kriterlerin değişimi durumunda farklı sayıda öznitelik seçebilmektedir. Dolayısıile, PO yöntemi kullanarak yerel uzmanların birleştirilmesi için önerdiğimiz model,gerekirci (deterministic) bir yöntem değildir. Uzmanların oluşumu rassallığa dayandığından önerilen model de sezgisel bir yöntemdir.
iii. **AHP tarafından seçilen gen kadar:** Karşılaştırma işleminde alternatiflerin varsayılan önceliğinden büyük önceliğe sahip genlerin sayısını ifade etmektedir. AHP yönteminde her bir alternatifin eşit öneme sahip olduğu düşünüldüğünde alternatifler için varsayılan öncelik 1/alternatif sayısı kadardır. PO yöntemi ile yerel uzmanları birleştiren model tarafından seçilen genlerden, AHP sıralaması

sonrası varsayılan öncelikten büyük değere sahip olanlarının daha önemli olduğu söylenebilir.

Uzmanların birleştirilmesi yöntemlerini karşılaştırmak için, veri kümesinden her seferinde 2 örnek çıkarılarak 100 farklı alt uzay elde edilmiştir. Her bir veri kümesinden 20 farklı alt küme alınarak 20 tane yerel uzman oluşturulmuştur. Yerel uzmanlar 2 farklı şekilde birleştirilmiştir. Birincisi bu çalışmada önerildiği gibi, tüm yerel uzmanları dikkate alan çokkriterli PO&AHP karma yöntemi (topluluk+PO&AHP) diğeri ise yerel uzmanların ortalamasını alan (topluluk) birleştirici yöntemdir. Ek olarak tüm veriyi gören ve tek bir öznitelik seçme işlemi yapan bir genel uzman(tum veri tek secim) da oluşturulmuştur. Modellerin seçmiş olduğu genlerin sınıflandırma doğruluğu (accuracy-AC) 10 kat çapraz geçerleme yöntemi kullanılarak hesaplanmıştır.

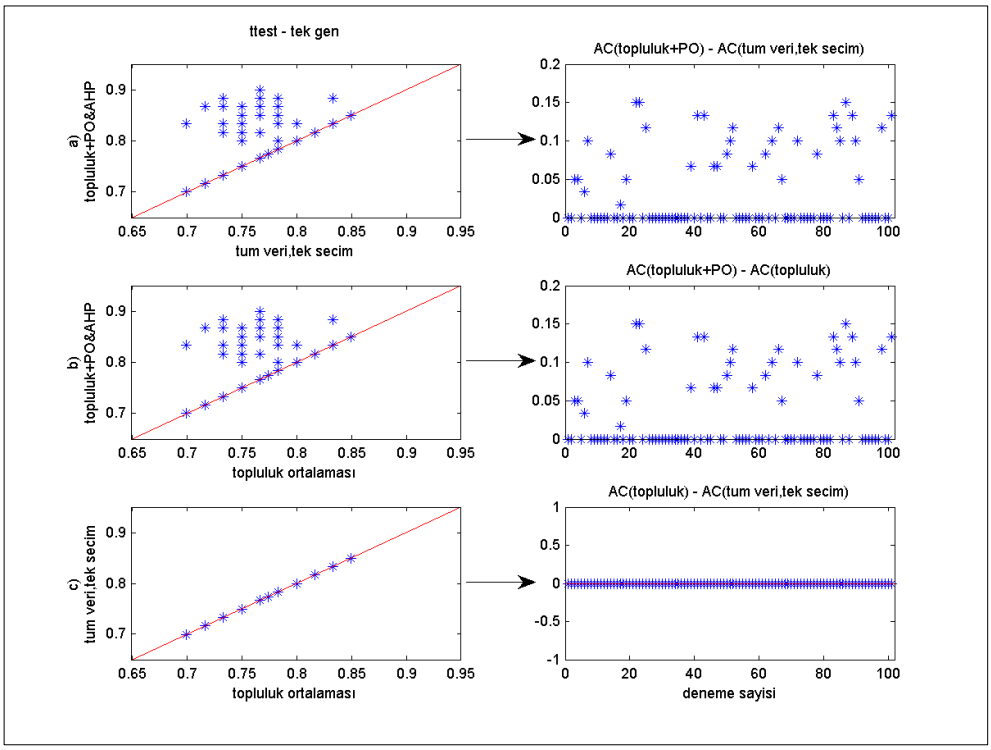
3 modelin(topluluk+PO&AHP, topluluk, tum veri tek secim) 5 farklı derecelendirme yönteminin (*ttest,entropy,batt,roc,wilcoxon*) 3 farklı veri kümesindeki (Kolon, Duke, DLBCL) karılaştırma sonuçları **Şekil 4.2-Şekil 4.43**'de sunulmuştur. Şekillerde, ilk kolonda yer alan grafikler bir modele karşılık diğer modelinsınıflandırma doğruluğunu göstermektedir. Bu grafiklerde, iki modelin başarısının eşit olma durumu köşegen üzerinde yer alan çizgi ile temsil edilmektedir. Dolayısı ile köşegen üzerindeki noktalar, iki modelin birbirini yenemediği durumları göstermektedir. Köşegen-x ekseni arasında noktalar,x eksenindeki modelin daha başarılı olduğu durumları, köşegen-y ekseni arasındaki noktalar, y eksenindeki modelindaha başarılı olduğu durumları temsil etmektedir. Şekillerde, ikinci kolondaki grafikler ise karşılaştırılan modellerin doğruluk oranlarının farkını göstermektedir (x ekseni deneme sayısını gösterir). y=0 sıfır doğrusu üzerindeki noktalar, iki modelin eşit başarı durumunu, doğrunun altında kalan noktalar birinci kolondaki grafiklerde x ekseninde yer alan modelin başarılı olduğu durumları ve doğrunun üstünde kalan noktalar y eksenindeki modelin başarılı olduğu durumları göstermektedir.

Grafikler incelendiğinde, yerel uzmanları çok kriterli yaklaşım olan PO&AHP karma yöntemi ile birleştiren modelin birçok durumda daha başarılı olduğu görülmektedir. Bu sonuçlar, tüm veriyi tek başına kullanmak yerine, alt kümelerde uzmanlaşmış yerel öğrencilerin çok kriterli yaklaşımla birleştirmenin daha başarılı olduğunu açıkça göstermektedir. Ayrıca, çok amaçlı optimizasyonun, ortama, maksimum, minimum gibi

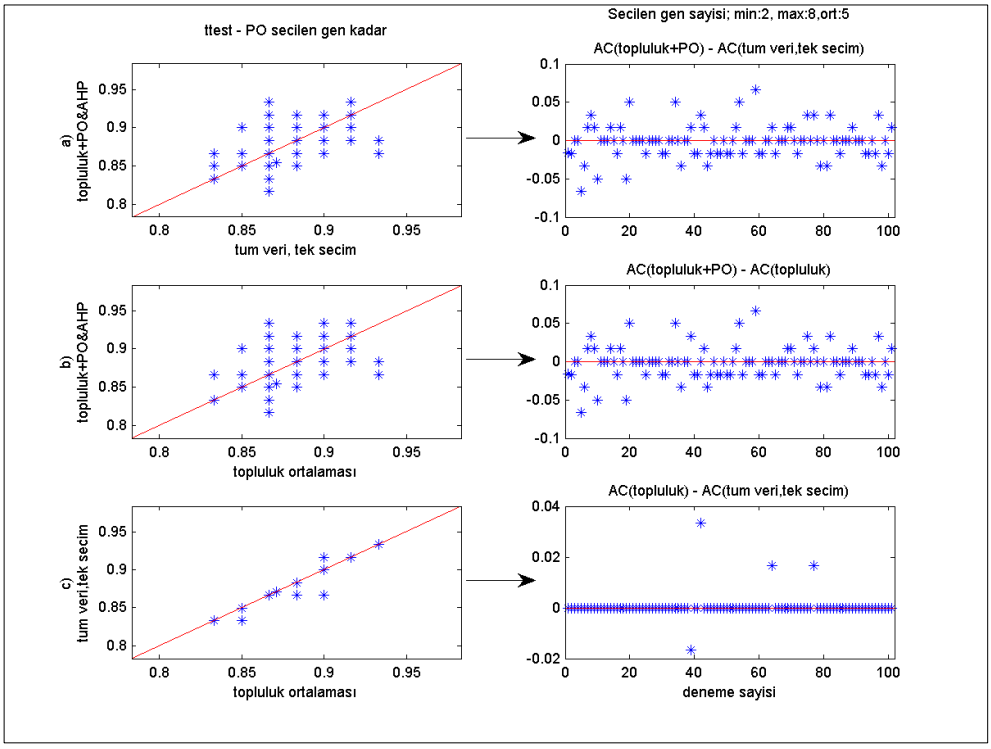
yöntemlerle tekbir amaca indirgenerek kullanılmasından daha başarılı olduğunu da göstermektedir. Buna karşın, Duke veri kümesinde tek gen seçimi işleminde *roc* ve *wilcoxon* yöntemlerini kullanan yerel uzmanların, önerilen karma modelle birleştirilmesi diğer iki modele göre başarısız olmuştur. Diğer bir sonuç ise bazı durumlar için 3 modelinde aynı başarıyı göstermesidir.

Yerel uzmanların ortalamasının alındığı model ile genel uzman karşılaştırıldığında ise, birçok durum için bu iki modelin eşit olduğu gözlemlenmiştir. Buradan hareketle tum veriyi kullanan genel öğrenicinin, yerel uzmanların ortalaması iye aynı olduğu ortaya çıkmaktadır. Eşitlik durumunun gözlendiği birçok durumda PO&AHP karma modelinin bu iki modeli geçtiği de açıkça görülmektedir. Buradan hareketle, eğer yerel öğrenicilerin gücü kullanılarak daha başarılı model oluşturulmak isteniyorsa, ortalama almanın bir etkisi olmadı, buna karşılık her bir yerel öğreniciyi dikkate alan çok kriterli yöntemlerin kullanılmasının iyi sonuçlar verdiği görülmektedir.

Tüm sonuçlar, yöneylem araştırmalarında çokça kullanılan PO, AHP gibi çok kriterli karar verme yöntemlerinin yerel uzmanları birleştirmede kullanılabileceğini göstermiştir. Verinin farklı alt kümelerini başarılı bir şekilde birleştiren çok kriterli karar verme yöntemlerinin, öznitelik seçim işleminde seçim modeline bağımlılığı azaltmak amacı ile farklı amaçlara hizmet eden seçim modellerinin de başarılı bir şekilde birleştirebileceğini düşündürmüştür. İleriki çalışmalarda hem yerel uzmanların hem yerel modellerin birlikte birleştirilmesi planlanmaktadır.

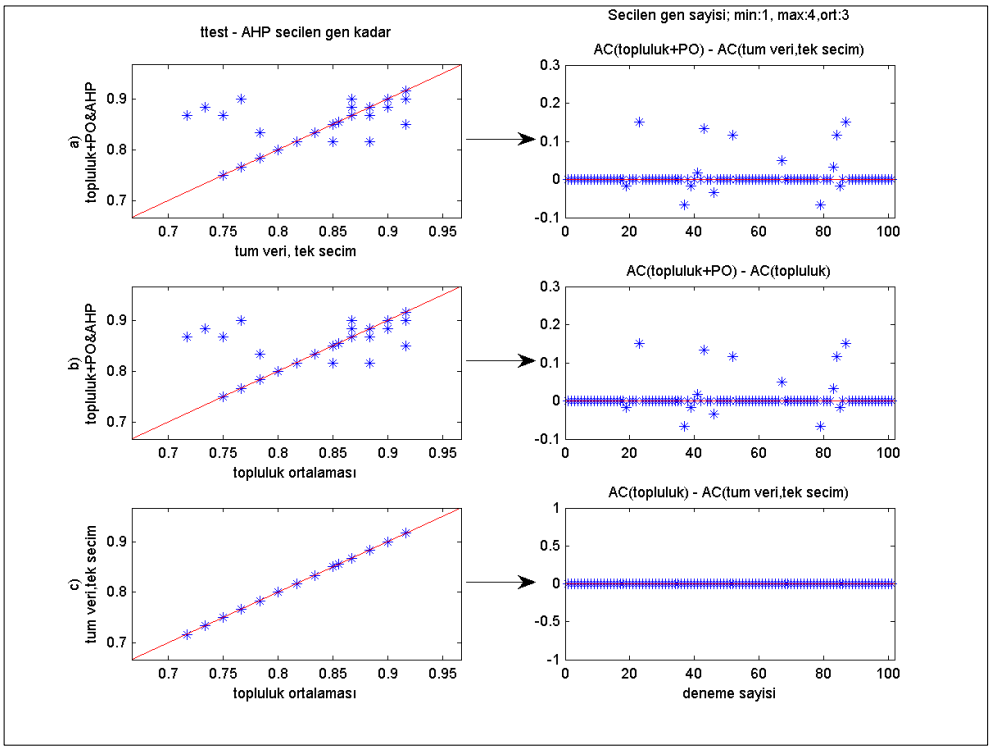


Şekil 4.2: Tek Geninkarşılaştırması-Kolon Verisi, *ttest* Yöntemi

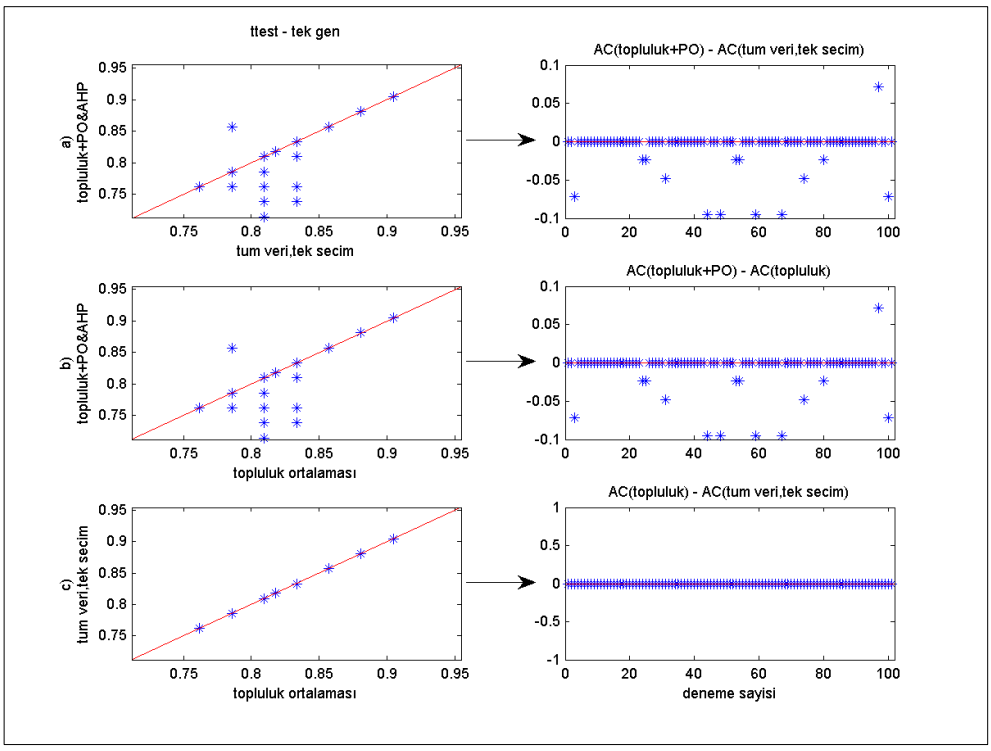


Şekil 4.3:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

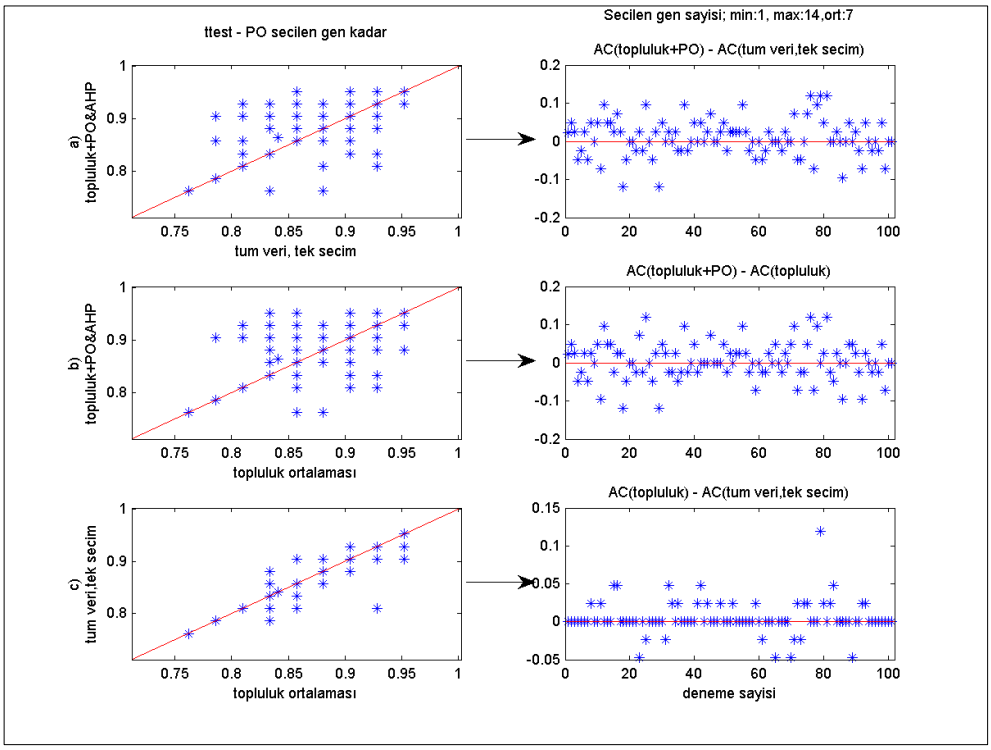
Kolon Verisi, *ttest* Yöntemi



Şekil 4.4:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-Kolon Verisi, *ttest* Yöntemi

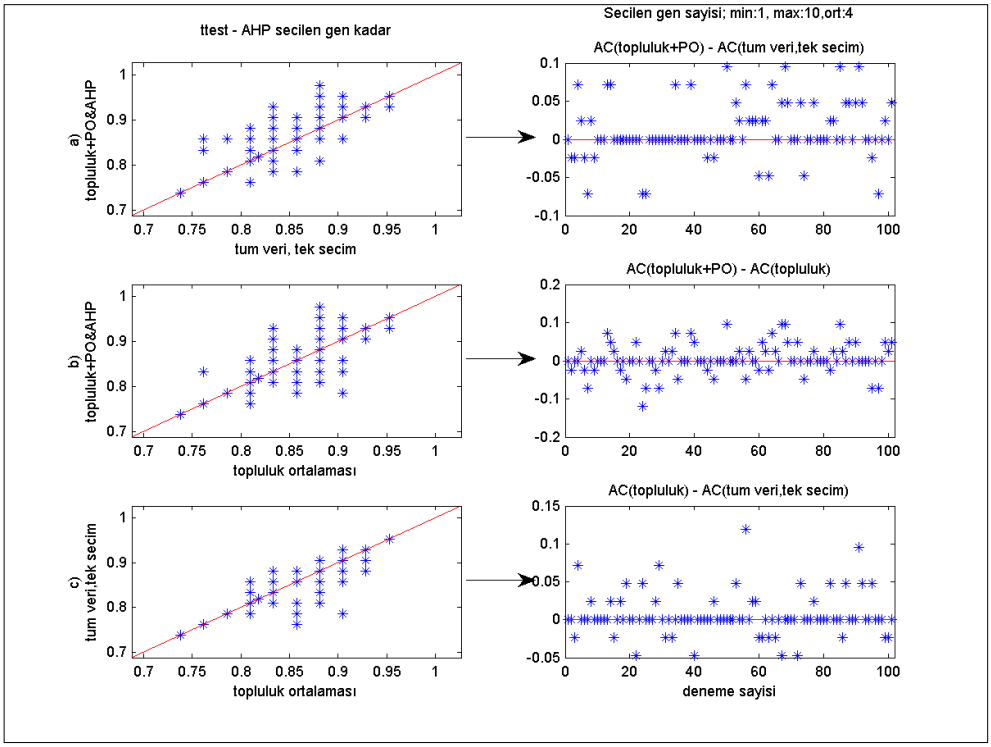


Şekil 4.5: Tek Geninkarşılaştırması-Duke Verisi, *ttest* Yöntemi



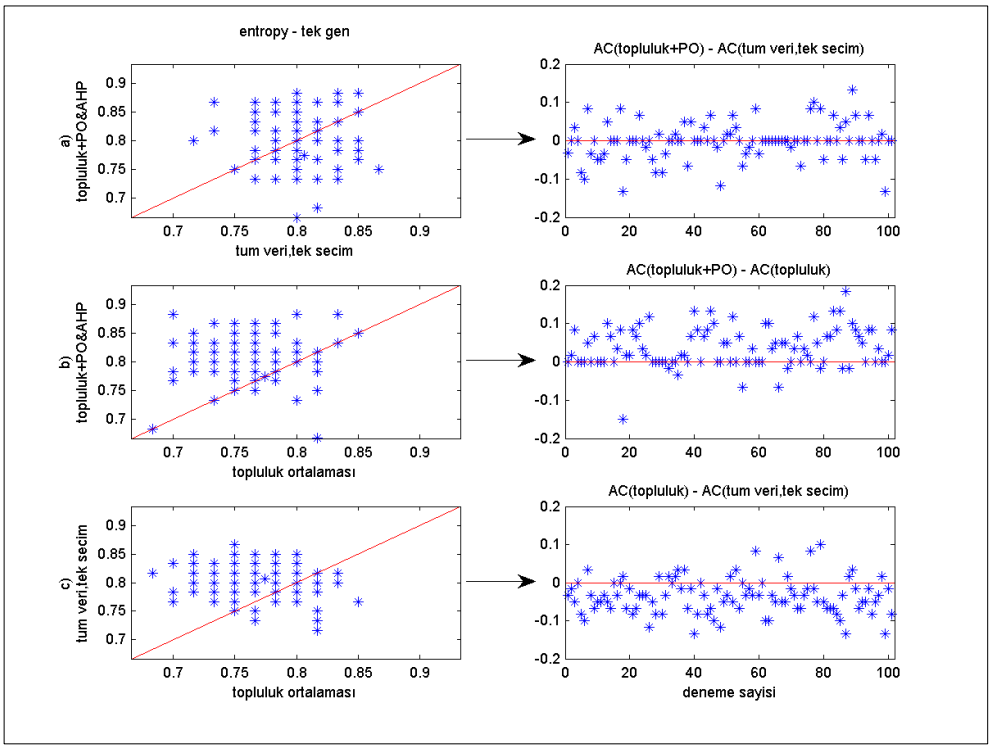
Şekil 4.6:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

Duke Verisi, *ttest* Yöntemi

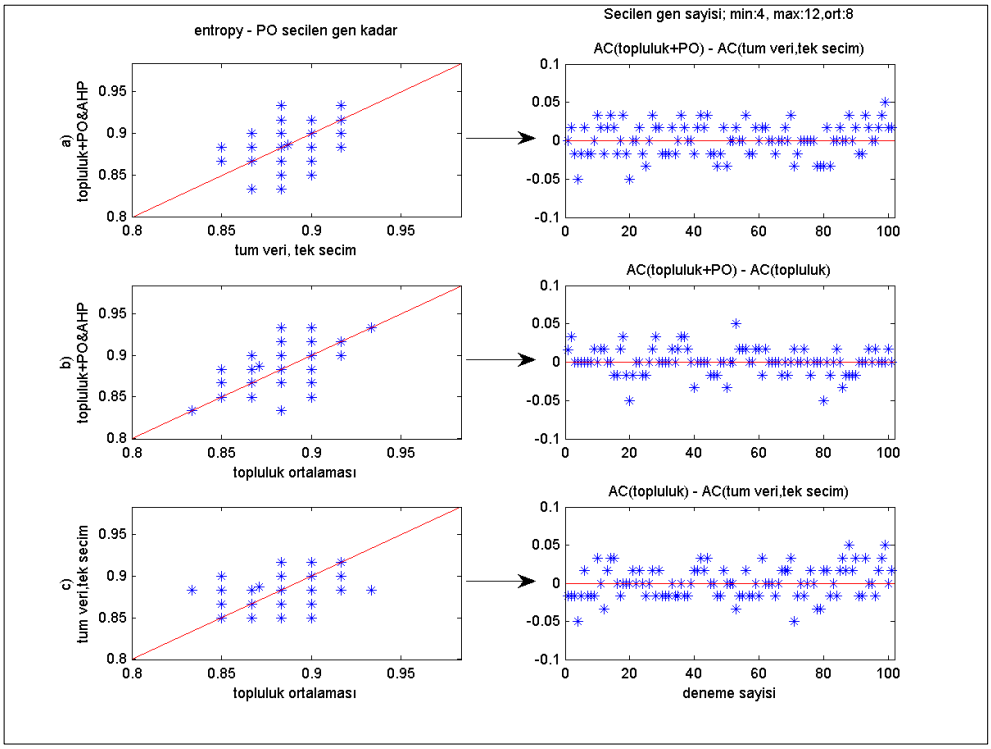


Şekil 4.7:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-

Duke Verisi, *ttest* Yöntemi

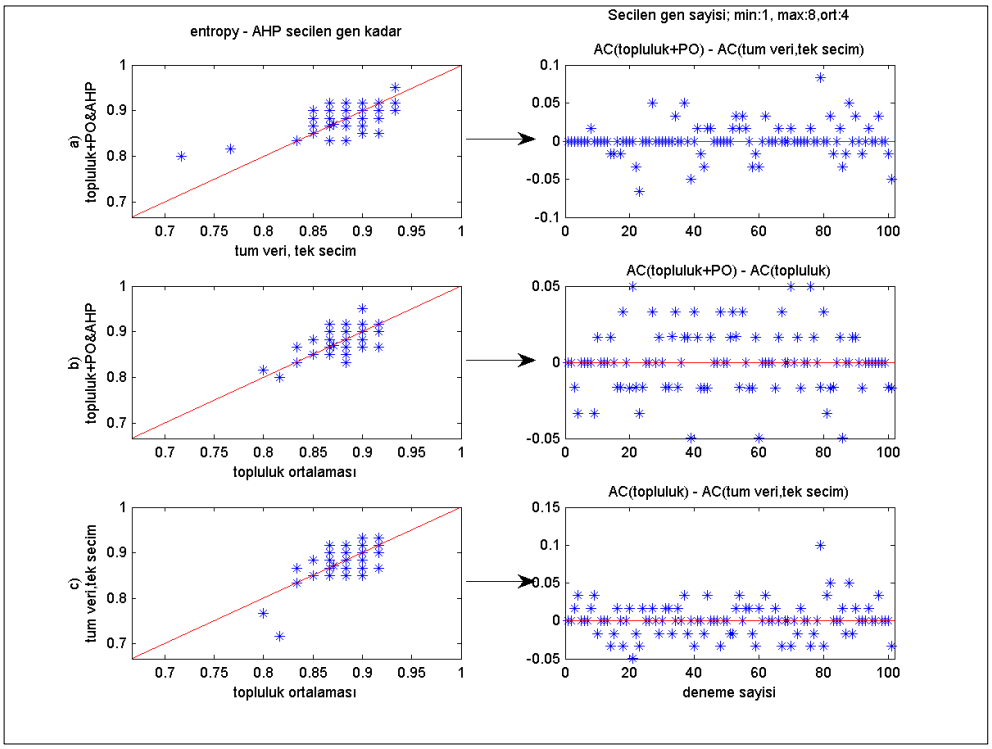


Şekil 4.8: Tek Genin Karşılaştırması-Kolon Verisi, *entropy* Yöntemi

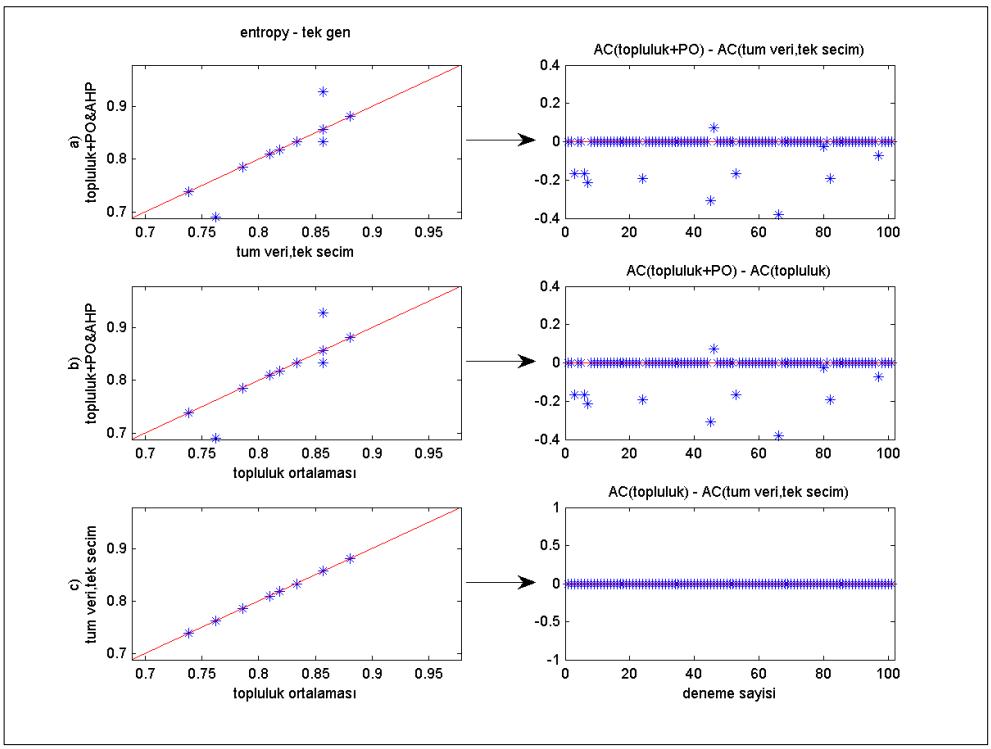


Şekil 4.9:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

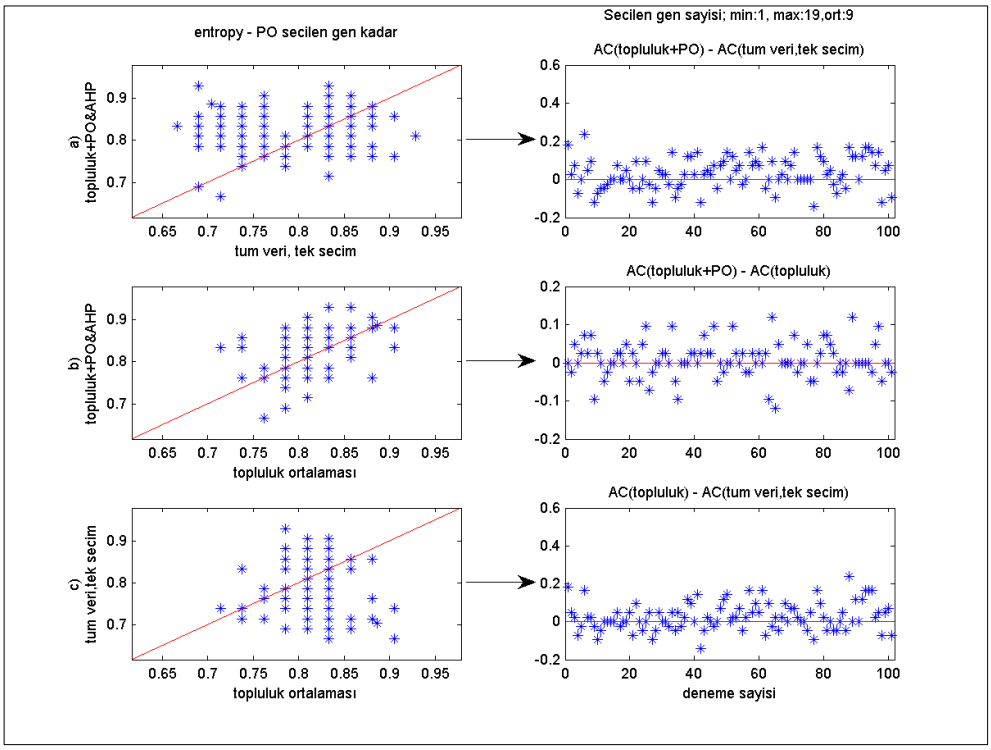
Kolon Verisi, *entropy* Yöntemi



Şekil 4.10:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-Kolon Verisi, *entropy* Yöntemi

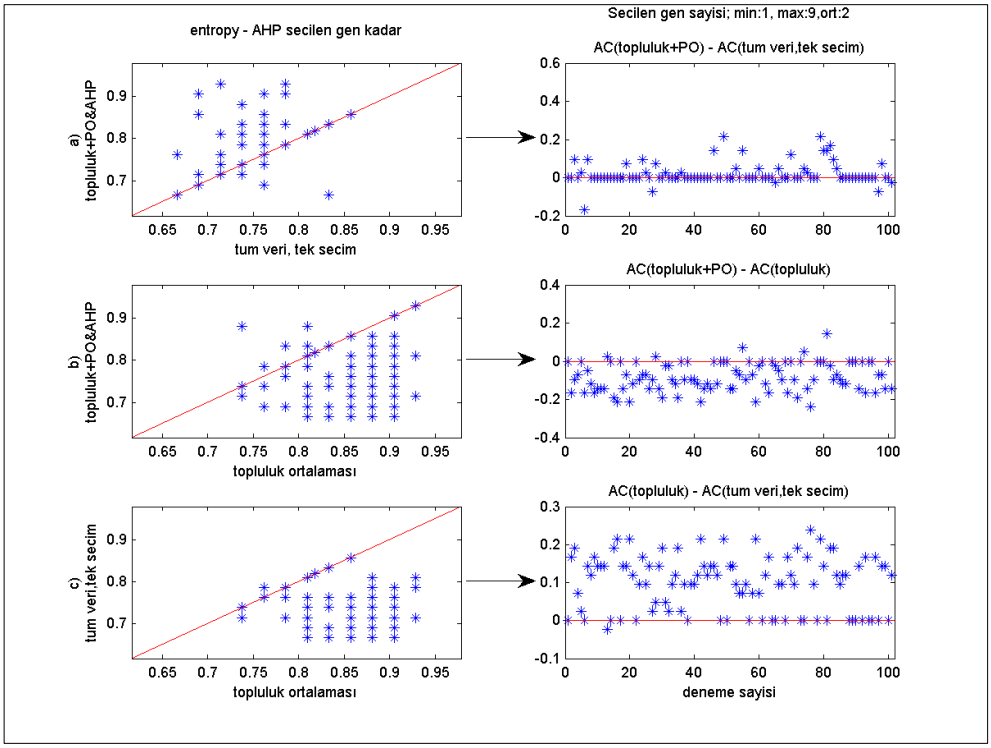


Şekil 4.11: Tek Geninkarşılaştırması-Duke Verisi, *entropy* Yöntemi



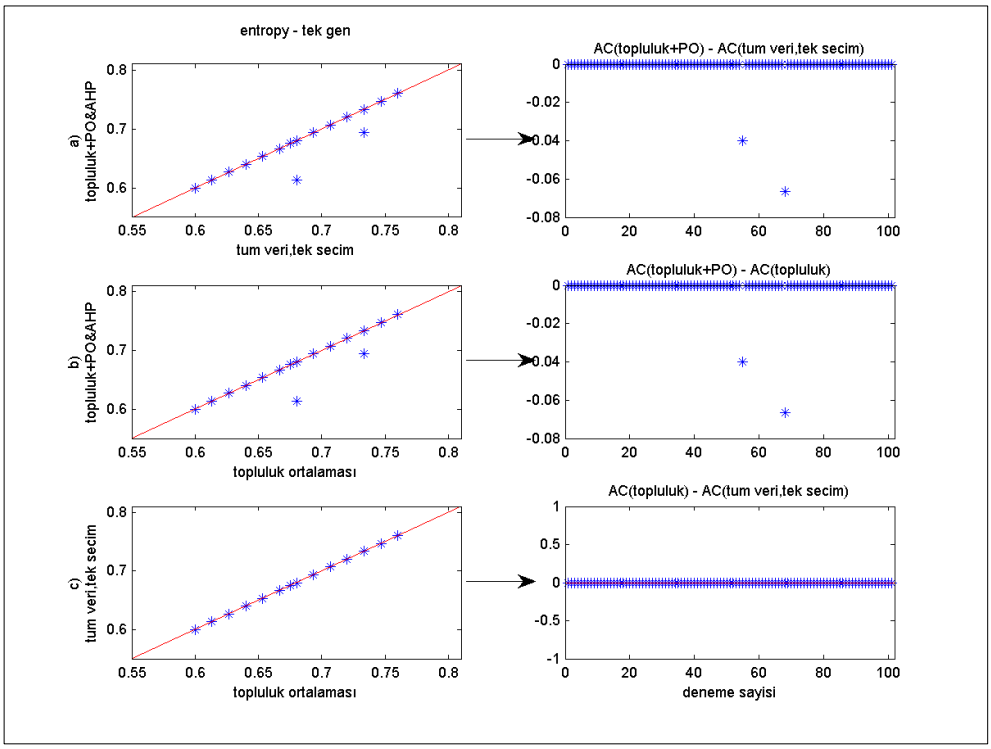
Şekil 4.12:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

Duke Verisi, *entropy* Yöntemi

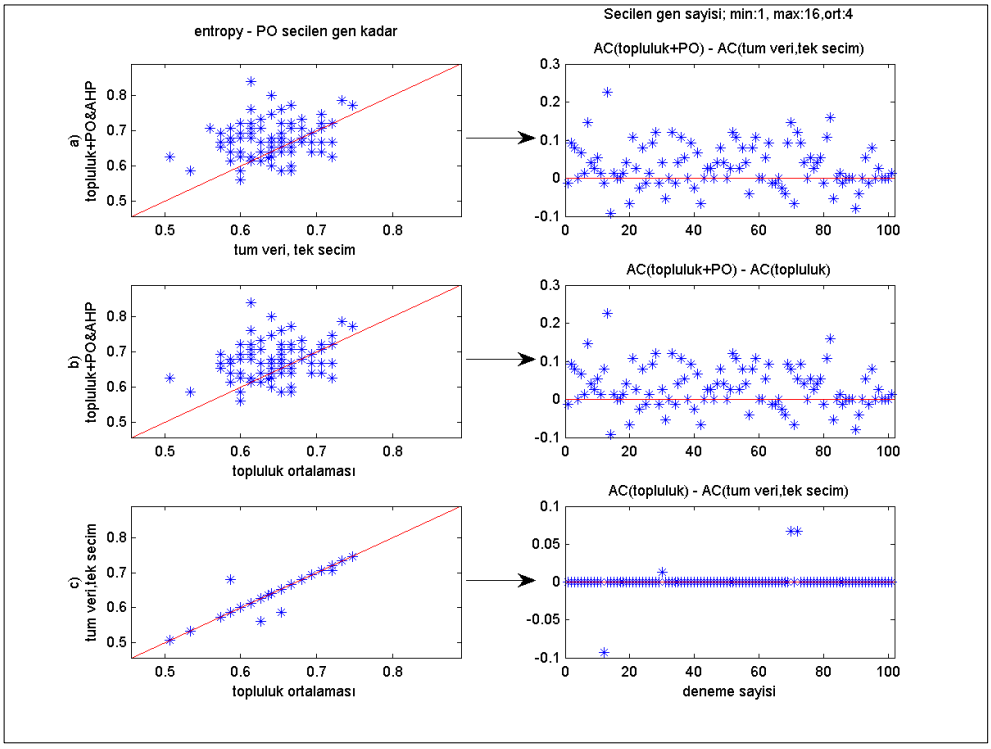


Şekil 4.13:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-

Duke Verisi, *entropy* Yöntemi

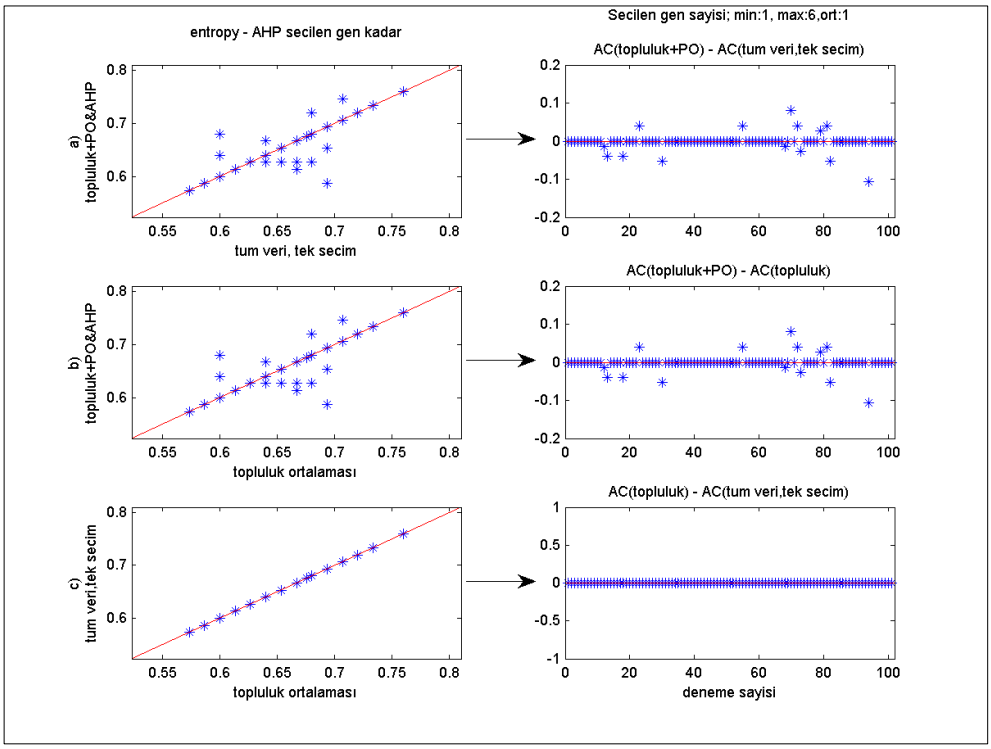


Şekil 4.14: Tek Geninkarşılaştırması-DLBCL Verisi, *entropy* Yöntemi

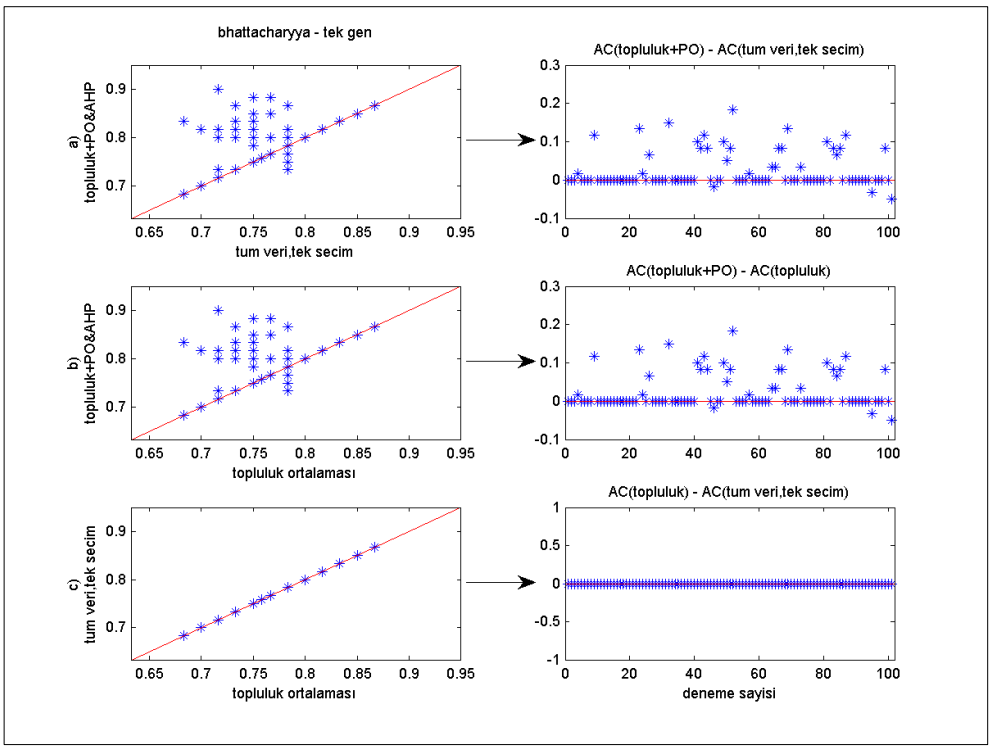


Şekil 4.15:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

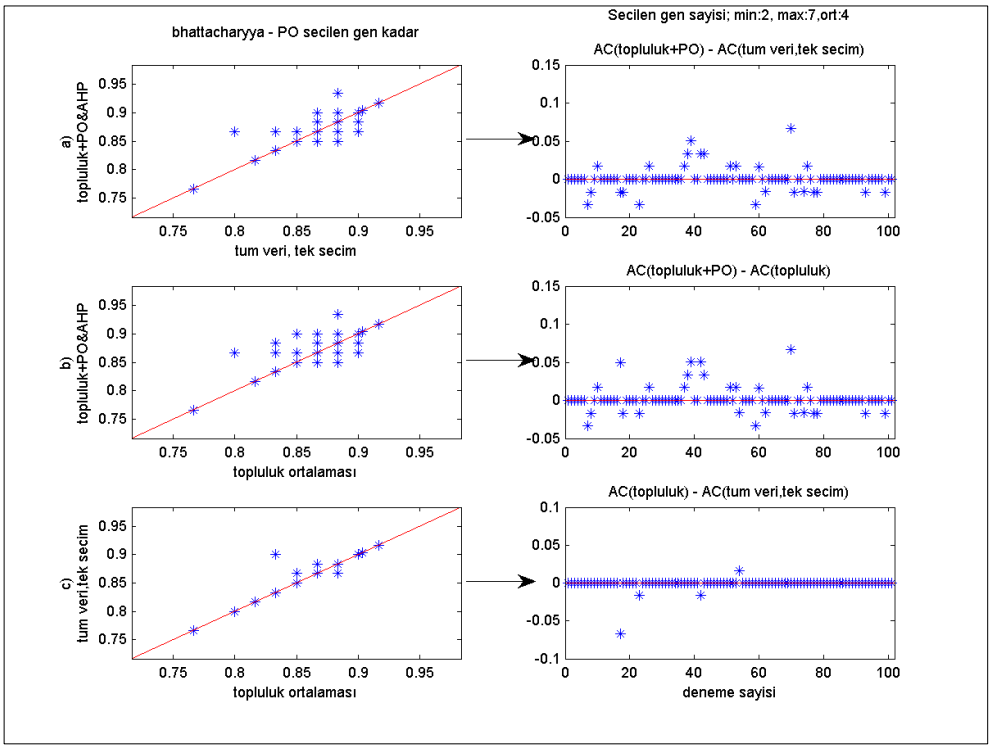
DLBCL Verisi, *entropy* Yöntemi



Şekil 4.16:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-DLBCL Verisi, *entropy* Yöntemi

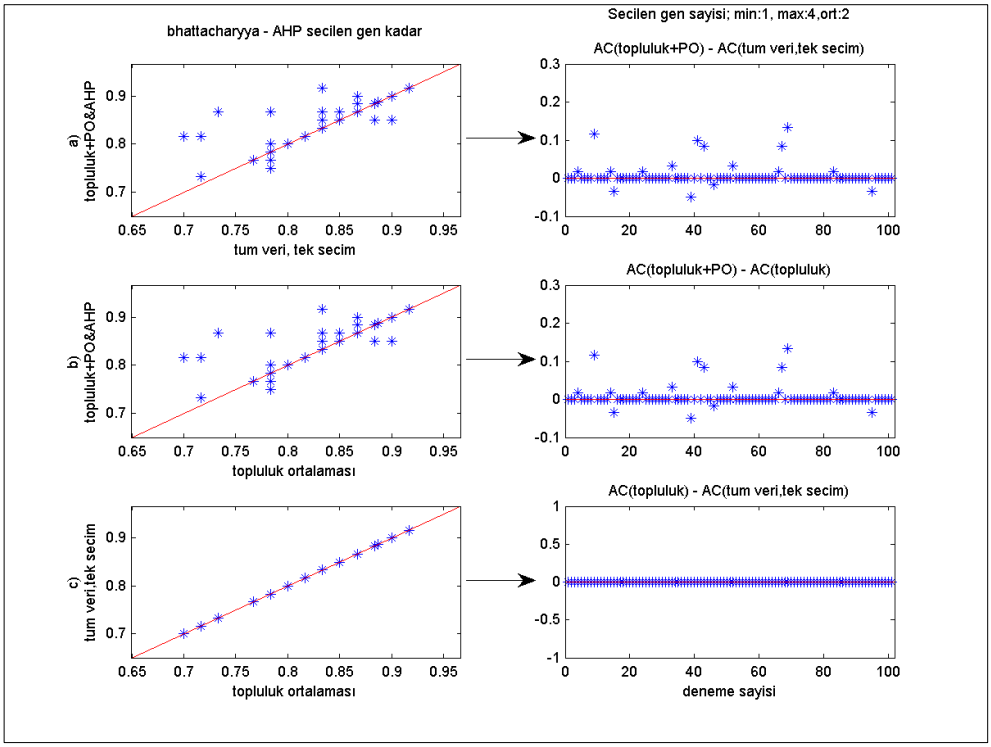


Şekil 4.17: Tek Geninkarşılaştırması-Kolon Verisi, *bhattacharyya Yöntemi*



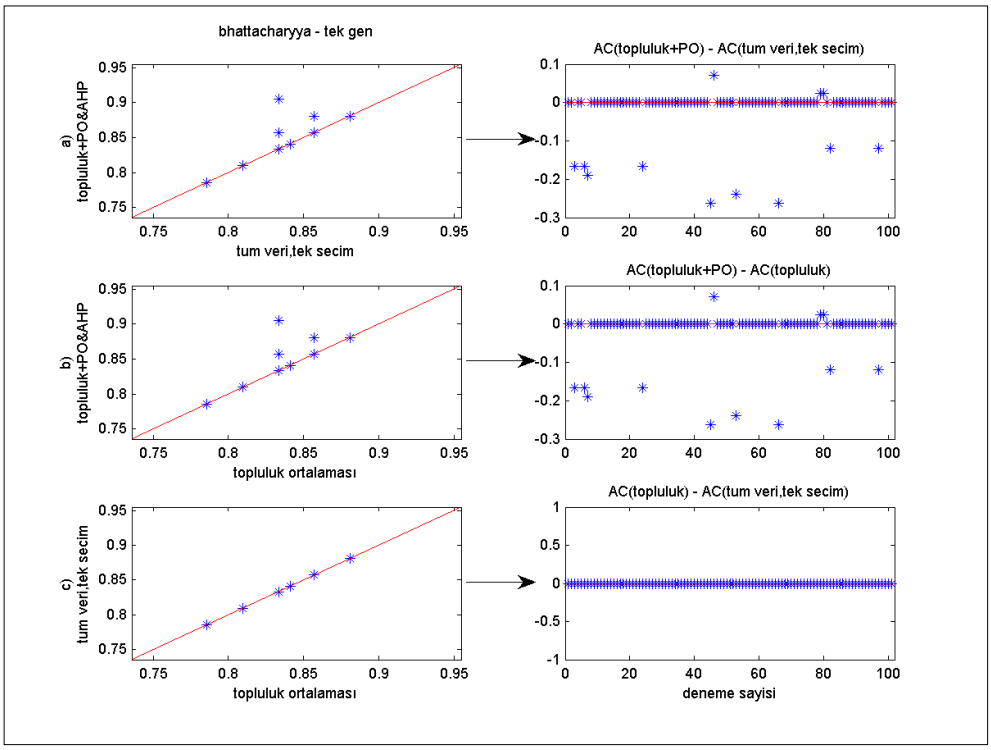
Şekil 4.18:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

Kolon Verisi, *bhattacharyya Yöntemi*

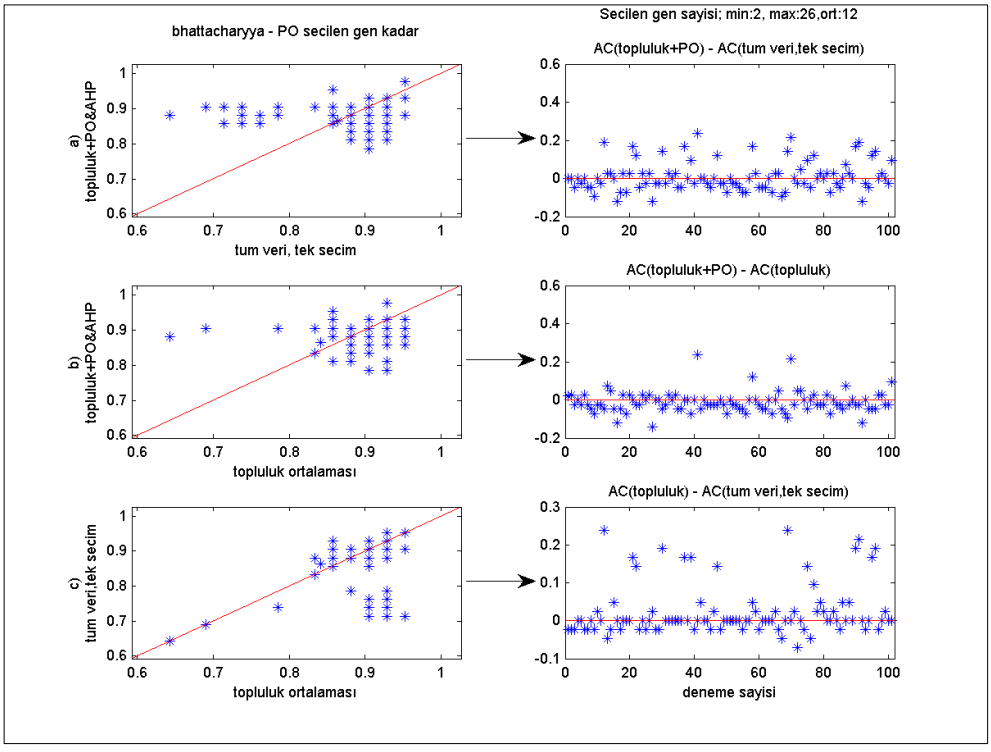


Şekil 4.19:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-

Kolon Verisi, *bhattacharyya Yöntemi*

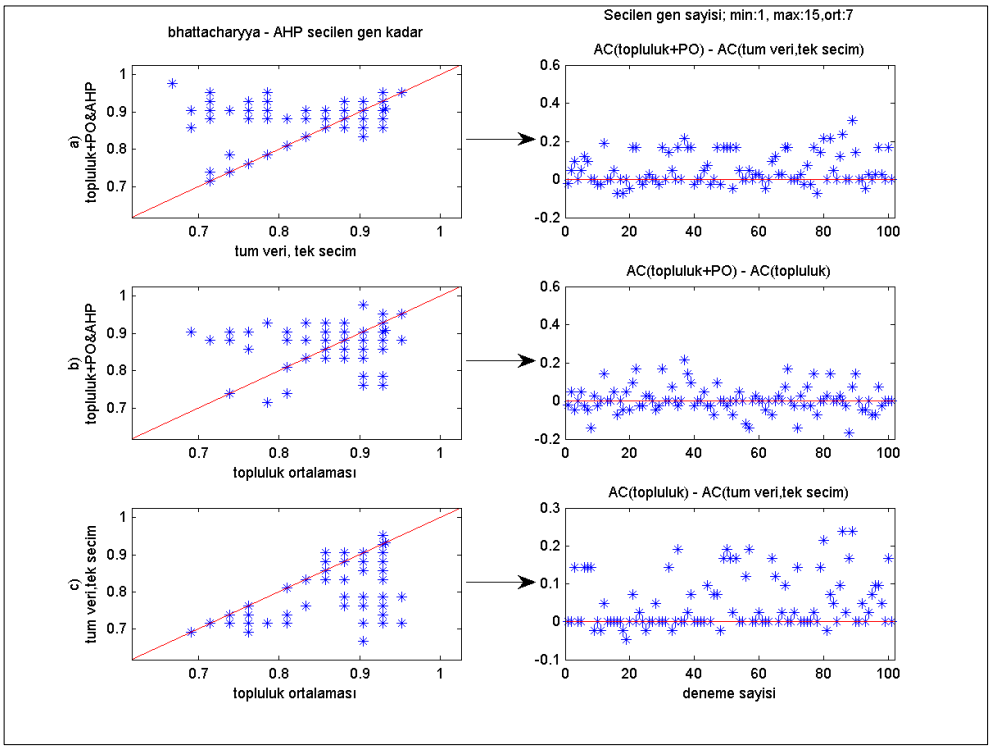


Şekil 4.20: Tek Geninkarşılaştırması-Duke Verisi, *bhattacharyya Yöntemi*

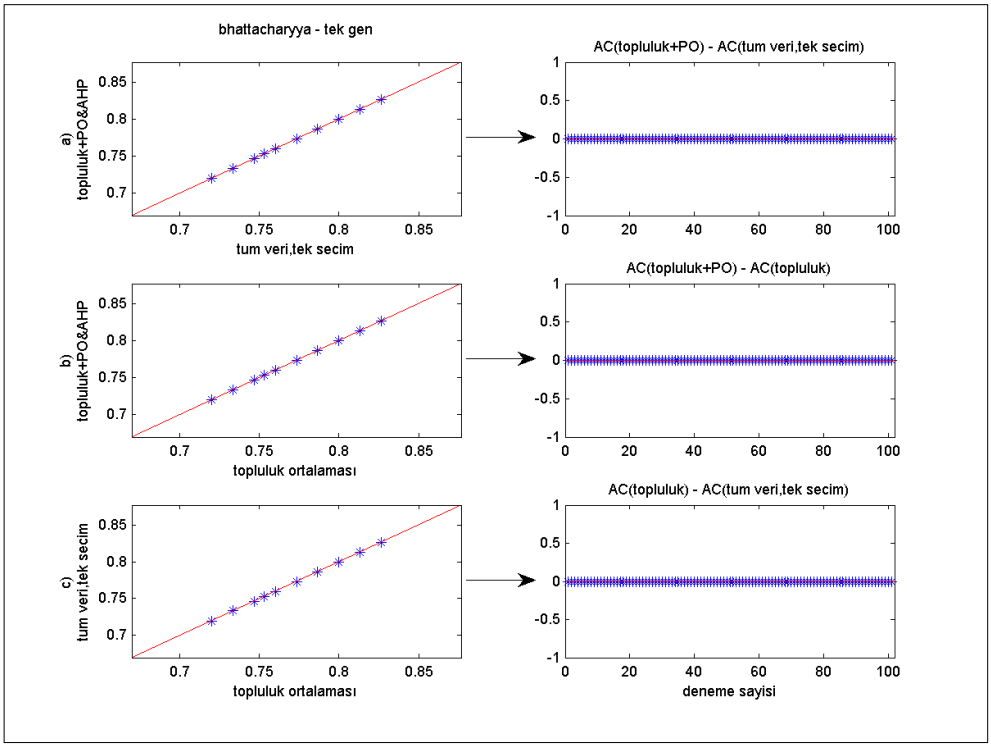


Şekil 4.21:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

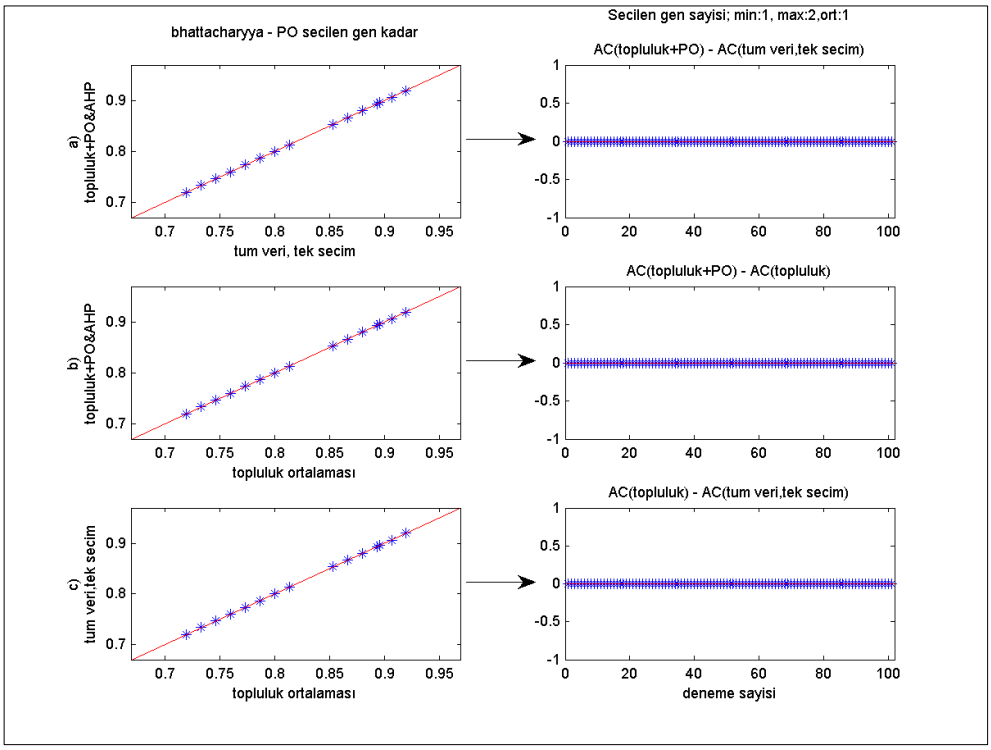
Duke Verisi, *bhattacharyya Yöntemi*



Şekil 4.22:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-Duke Verisi, *bhattacharyya Yöntemi*

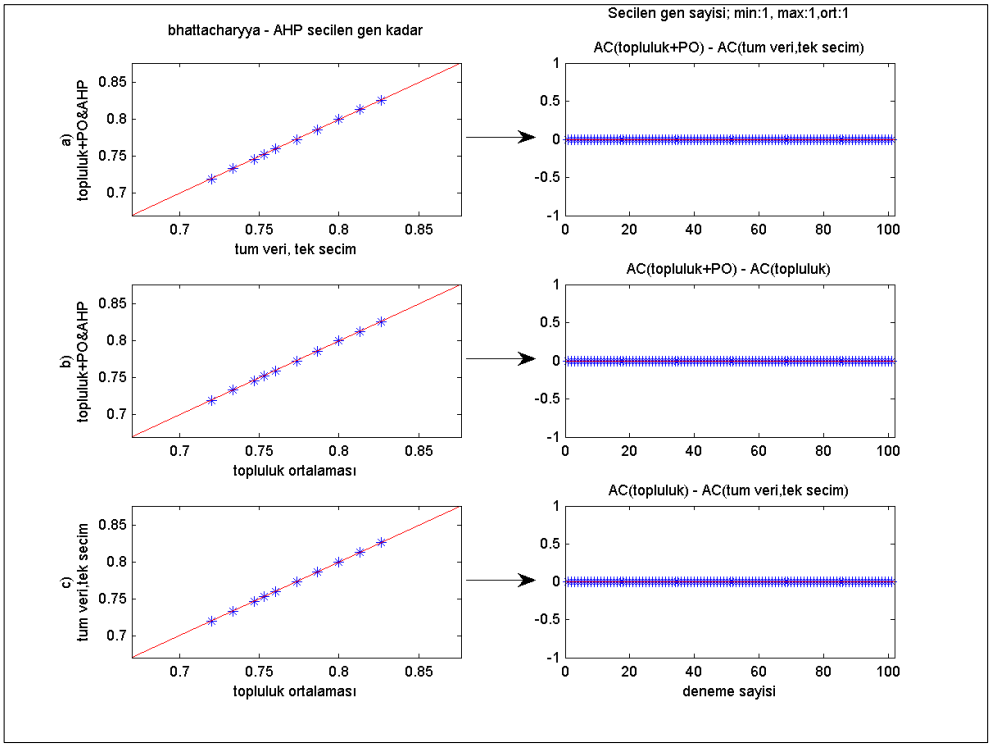


Şekil 4.23: Tek Geninkarşılaştırması-DLBCL Verisi, *bhattacharyya Yöntemi*



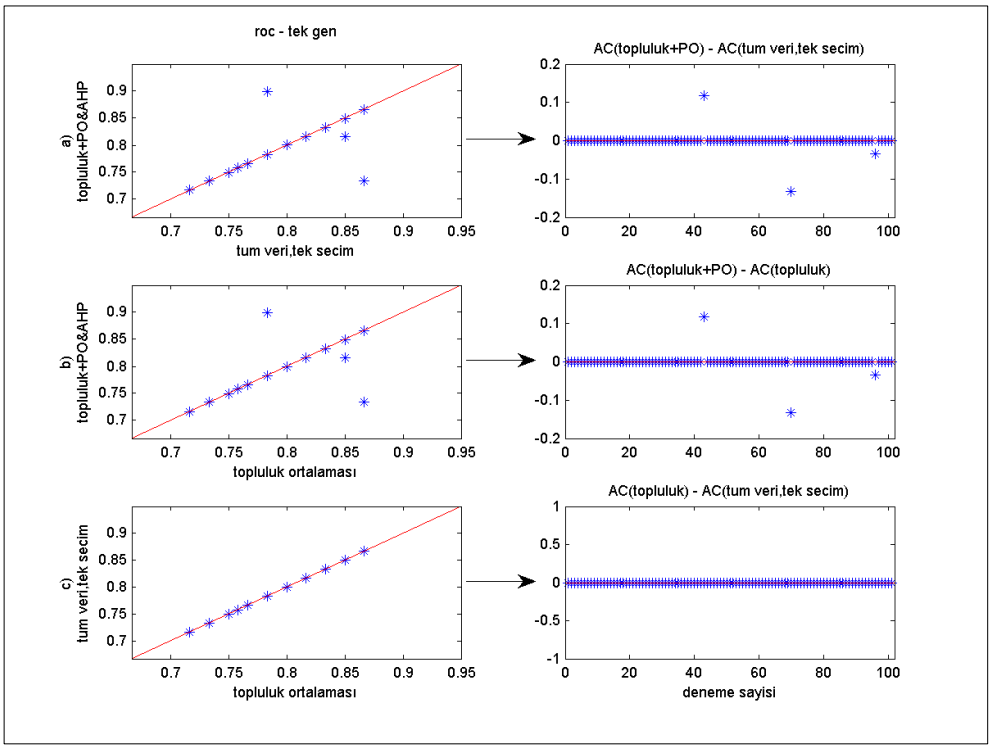
Şekil 4.24:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

DLBCL Verisi, *bhattacharyya Yöntemi*

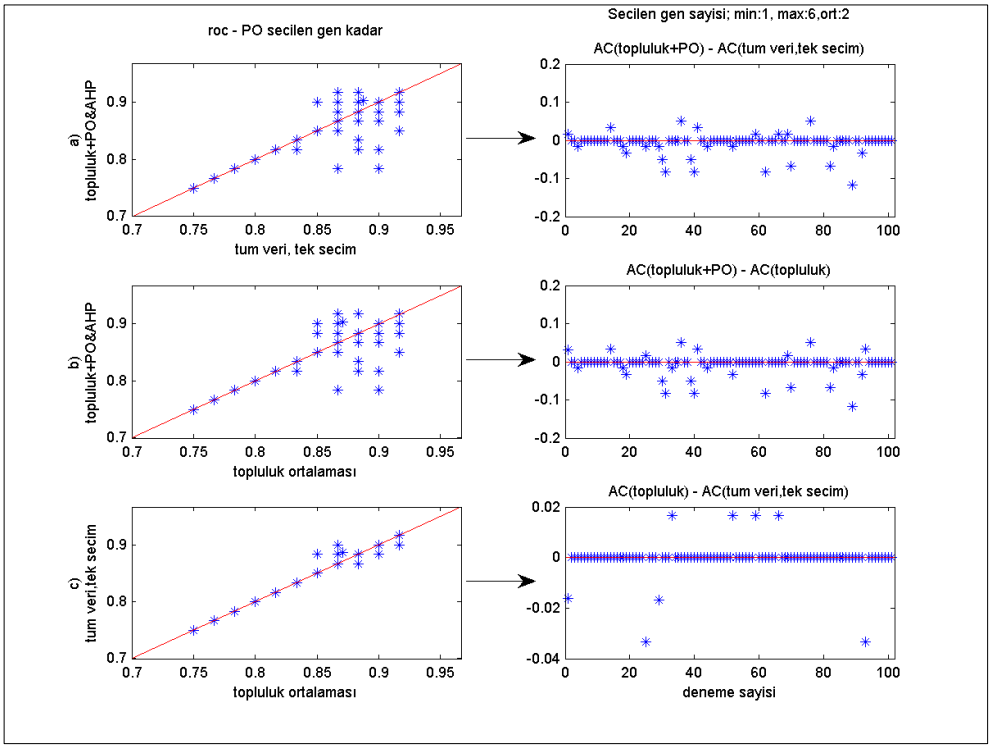


Şekil 4.25:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-

DLBCL Verisi, *bhattacharyya Yöntemi*

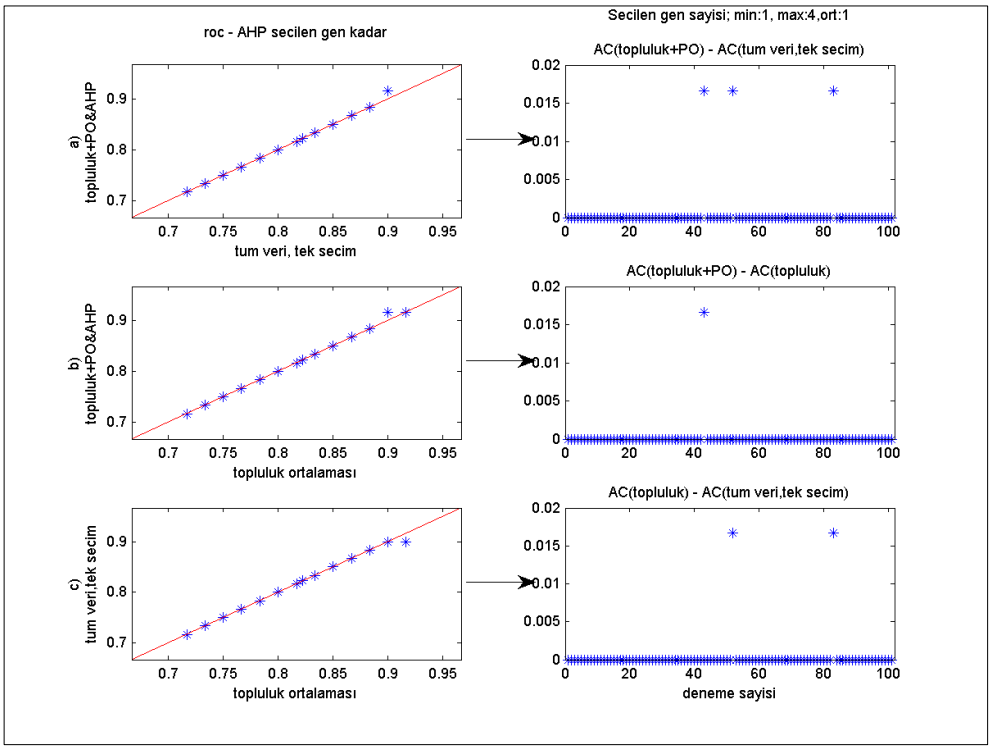


Şekil 4.26: Tek Genin Geninkarşılaştırması-Kolon Verisi, *roc* Yöntemi

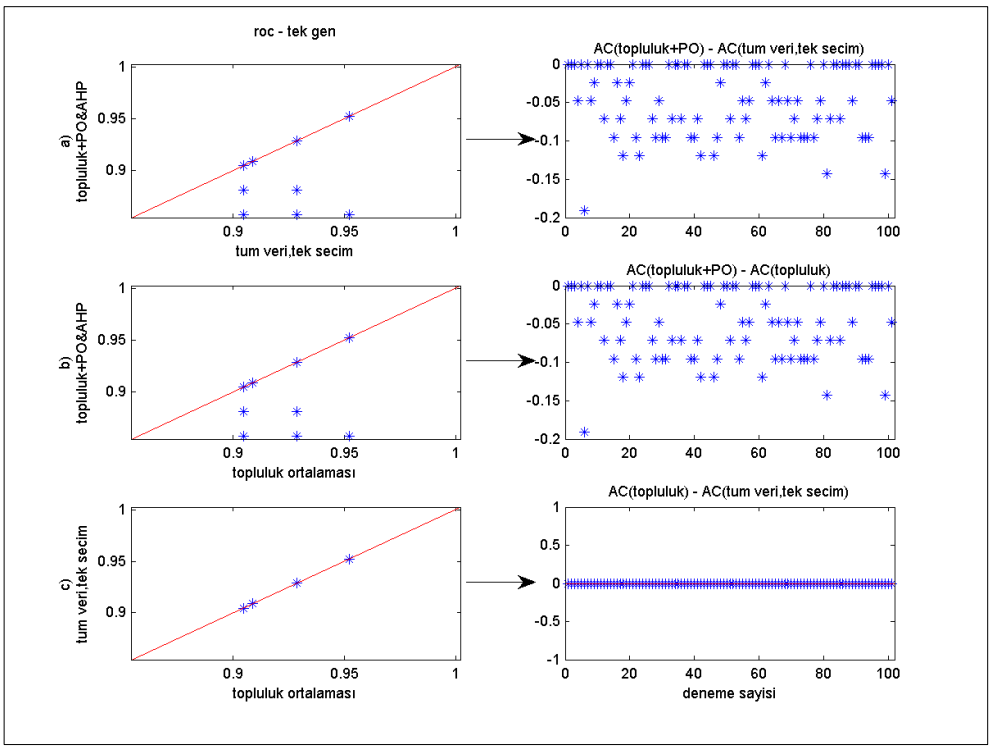


Şekil 4.27:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

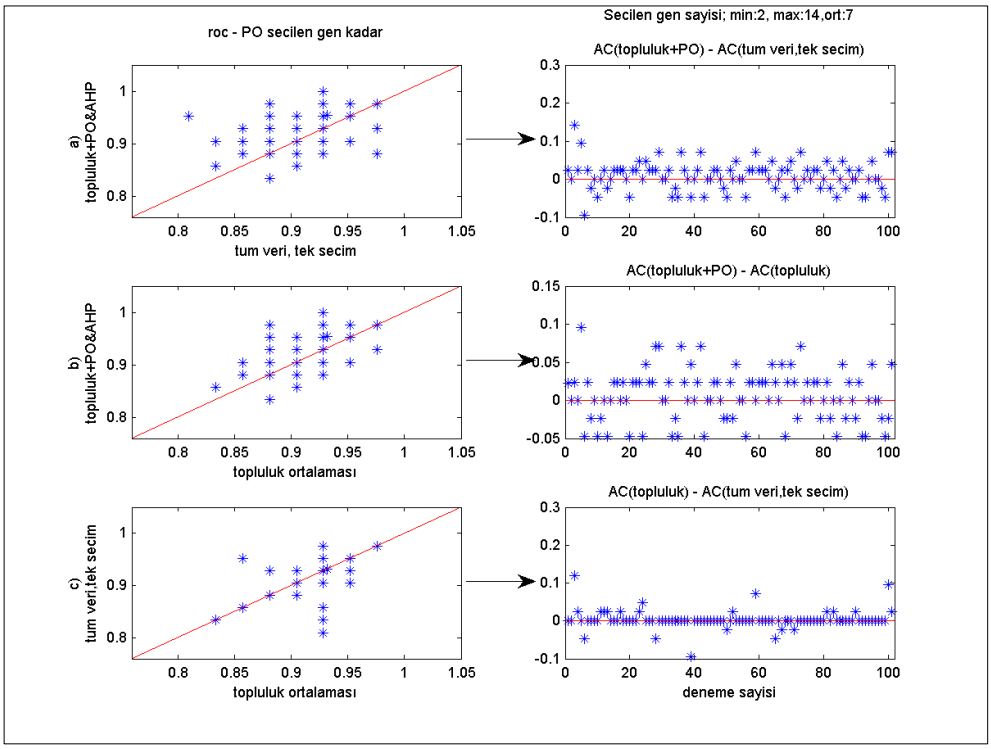
Kolon Verisi, *roc* Yöntemi



Şekil 4.28:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-Kolon Verisi, *roc* Yöntemi

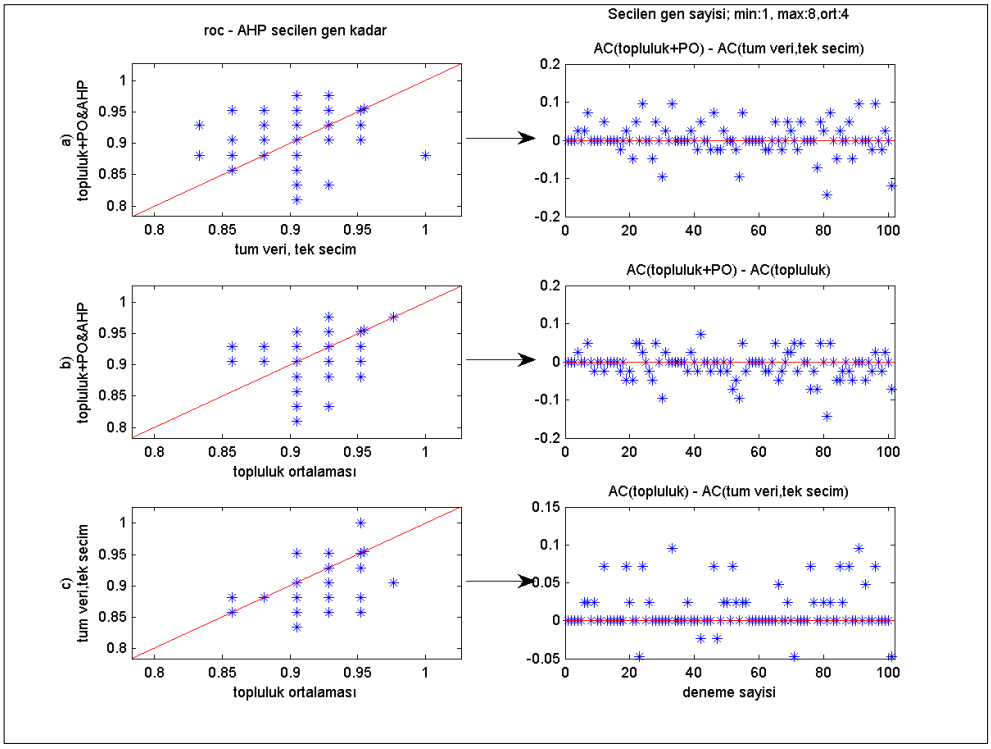


Şekil 4.29: Tek Geninkarşılaştırması-Duke Verisi, *roc* Yöntemi



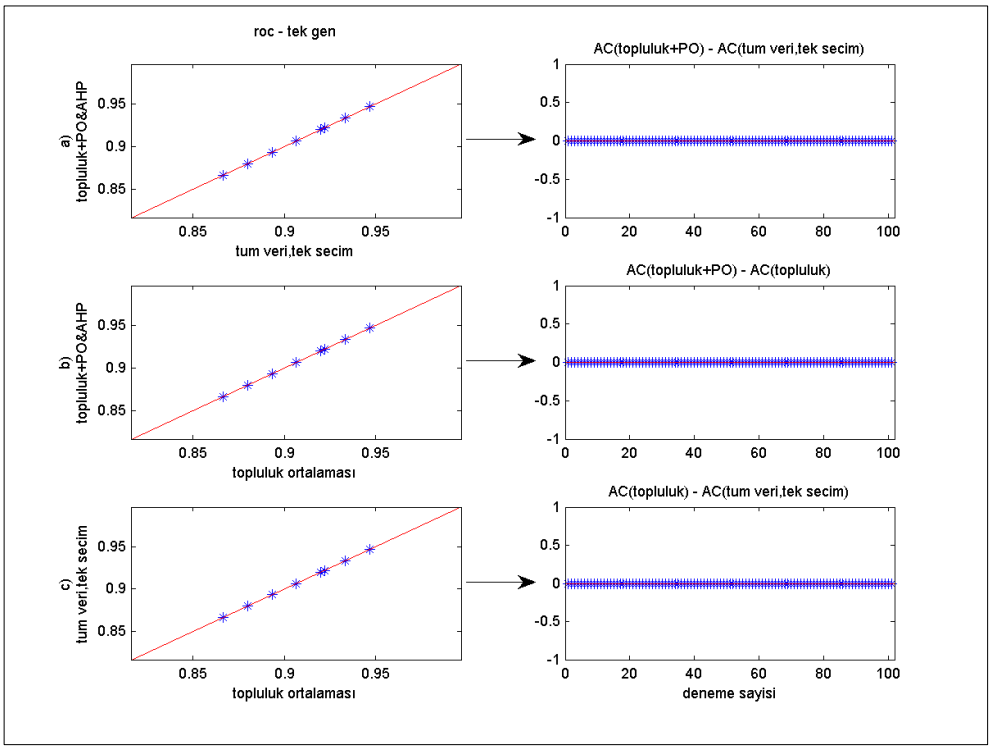
Şekil 4.30:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

Duke Verisi, *roc* Yöntemi

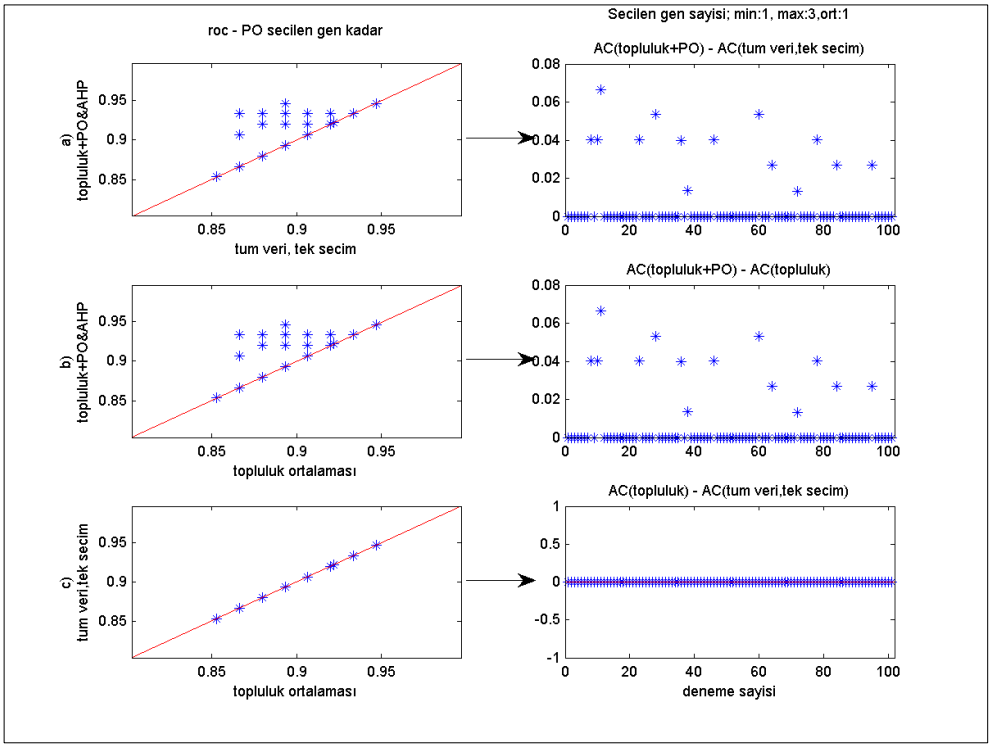


Şekil 4.31:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-

Duke Verisi, *roc* Yöntemi

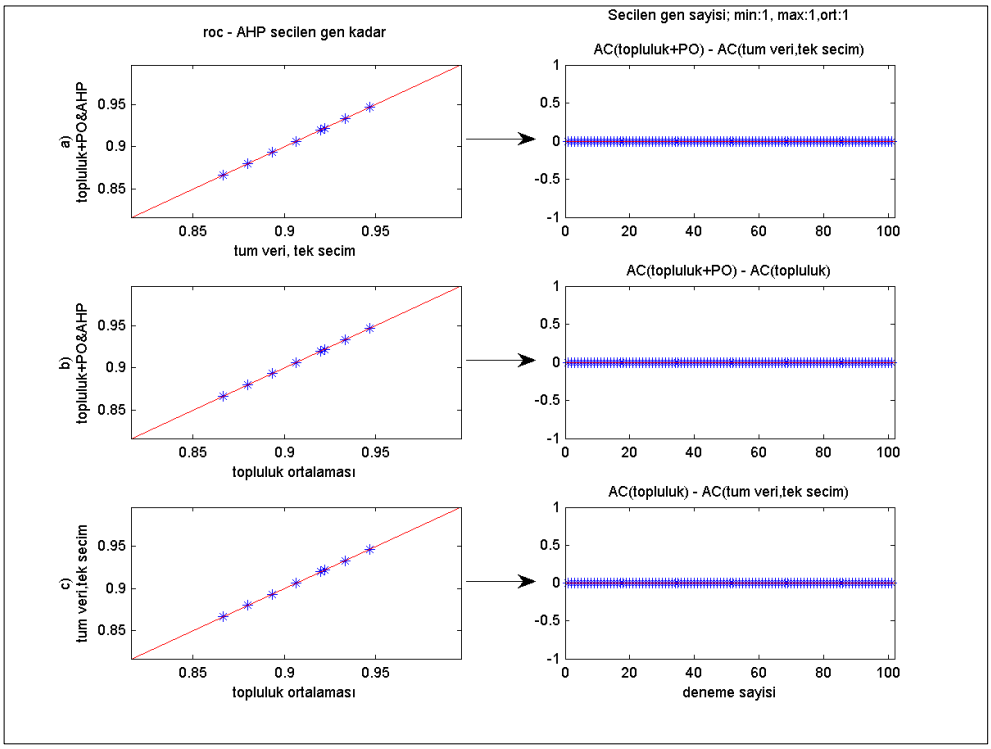


Şekil 4.32: Tek Geninkarşılaştırması-DLBCL Verisi, *roc* Yöntemi

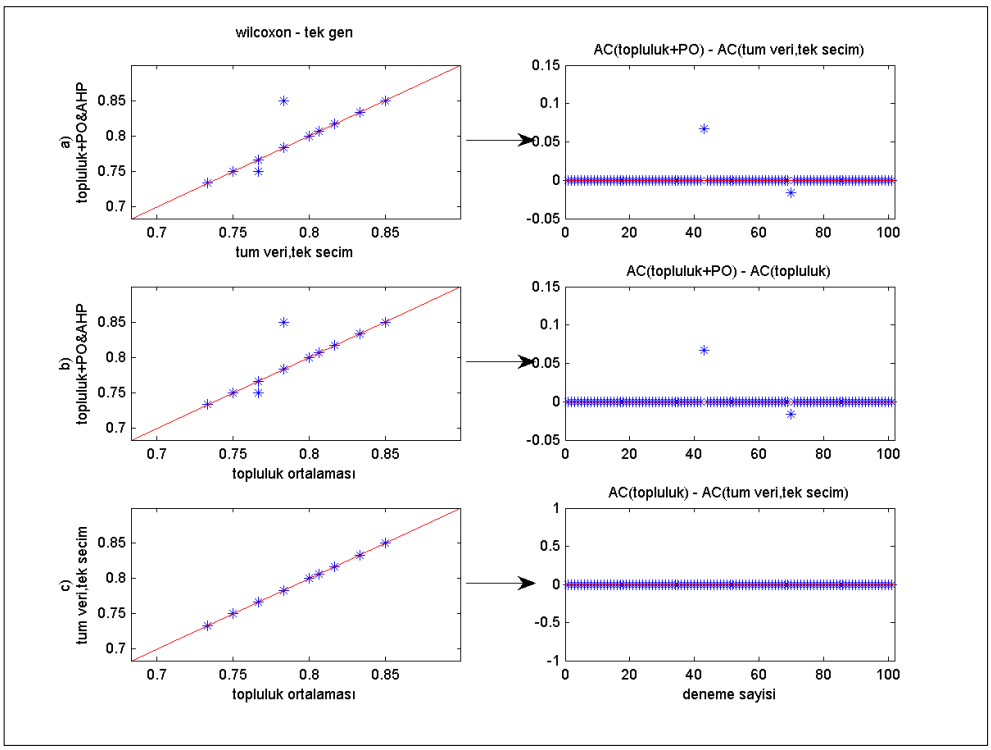


Şekil 4.33:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

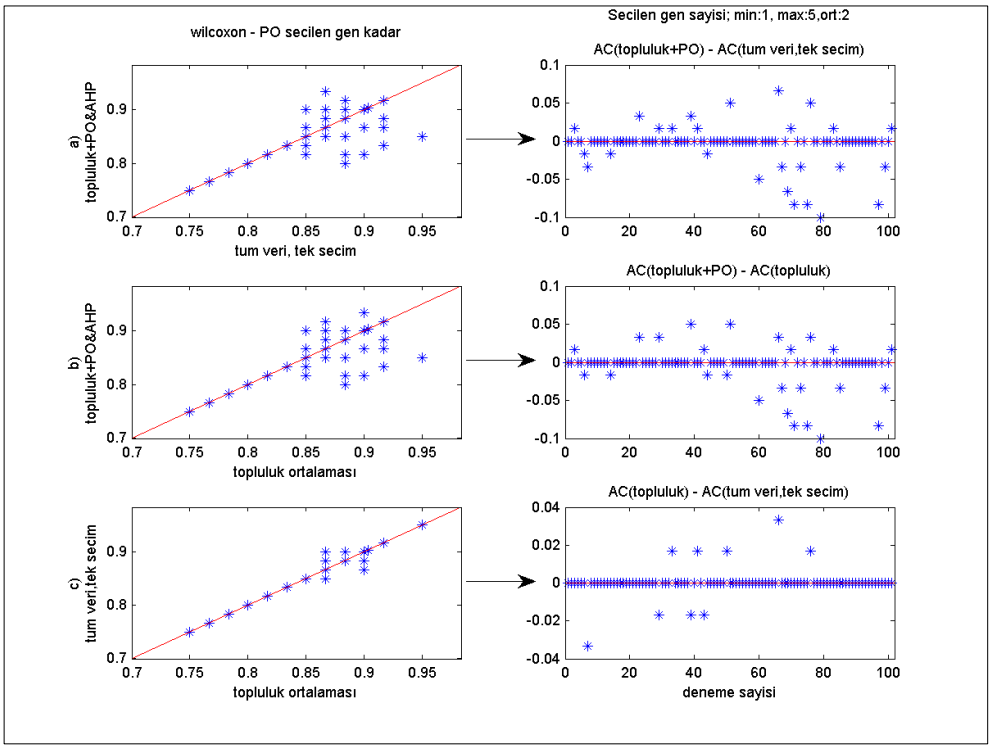
DLBCL Verisi, *roc* Yöntemi



Şekil 4.34:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-DLBCL Verisi, *roc* Yöntemi

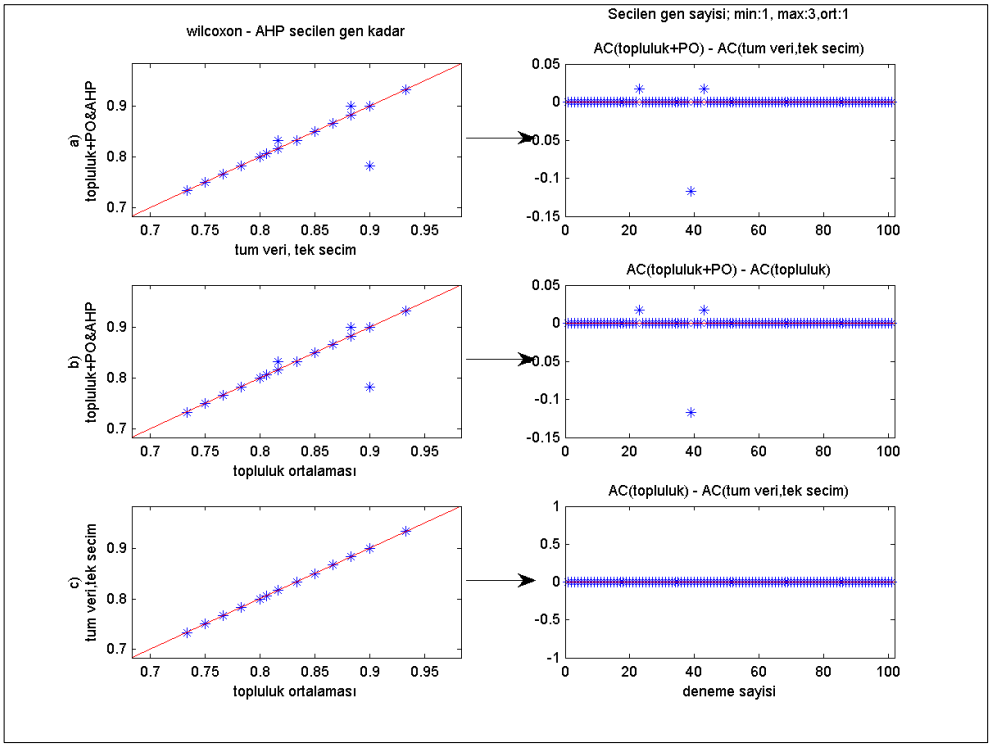


Şekil 4.35: Tek Genin Karşılaştırması-Kolon Verisi, *wilcoxon* Yöntemi



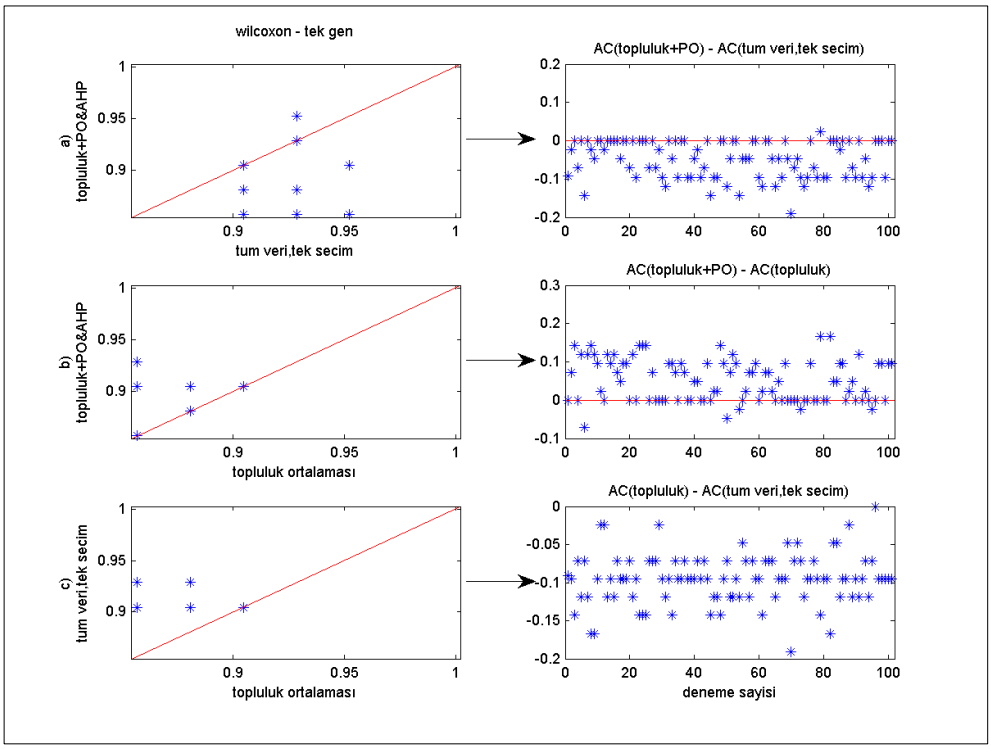
Şekil 4.36:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

Kolon Verisi, *wilcoxon* Yöntemi

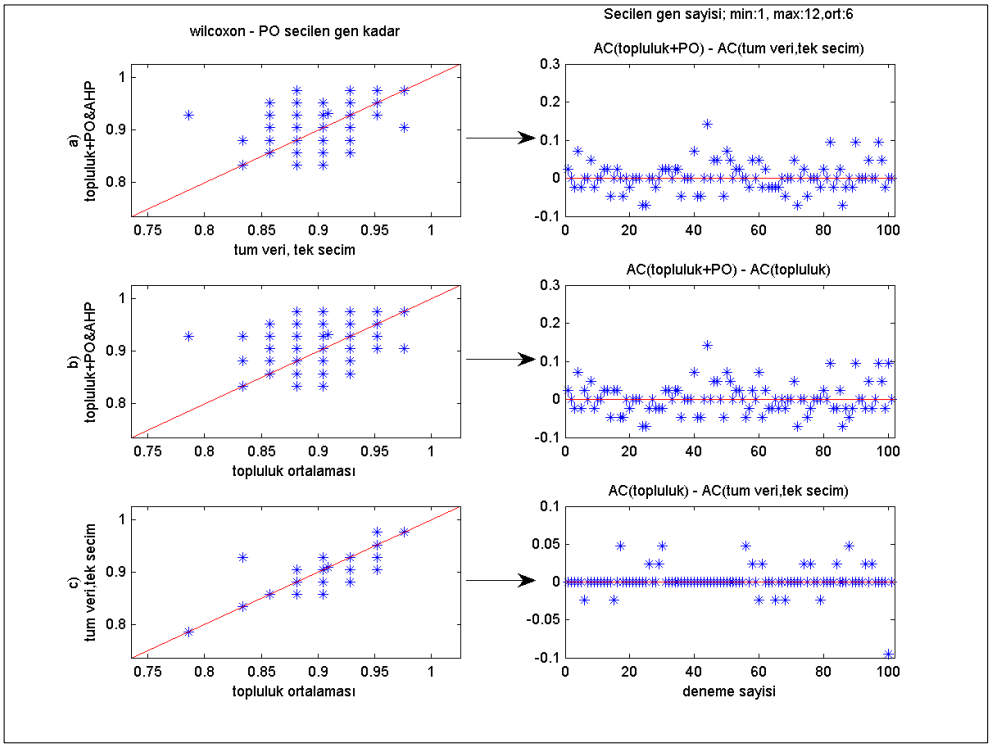


Şekil 4.37:AHP Tarafından Seçilen Kadar Geninkarşılaştırılması-

Kolon Verisi, *wilcoxon* Yöntemi

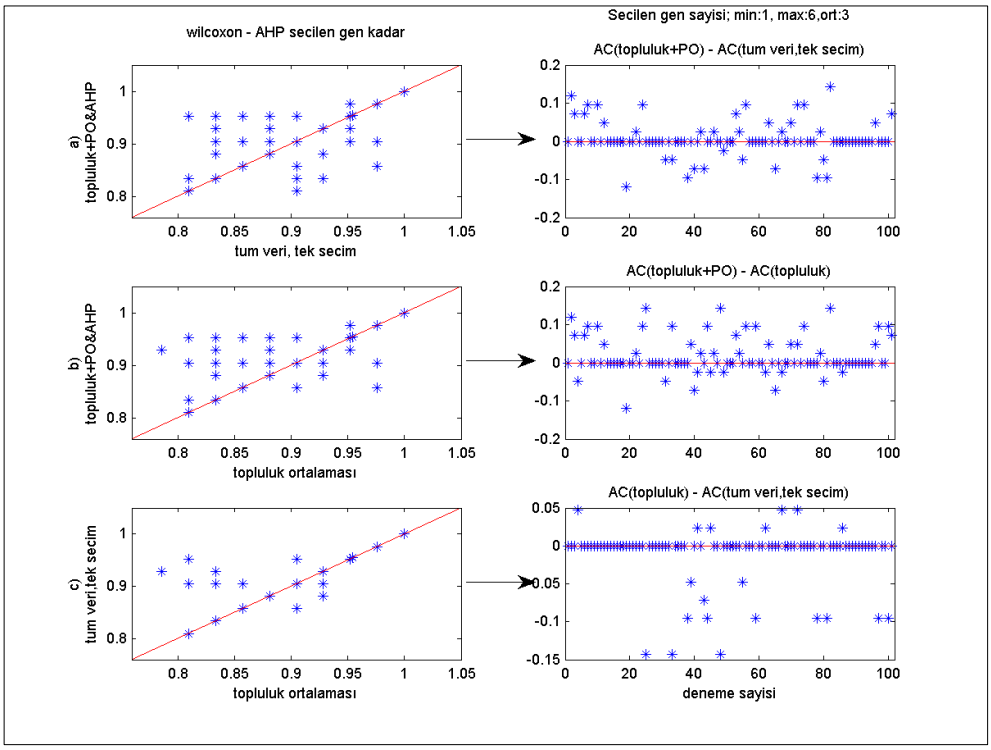


Şekil 4.38: Tek Geninkarşılaştırması-Duke Verisi, *wilcoxon* Yöntemi

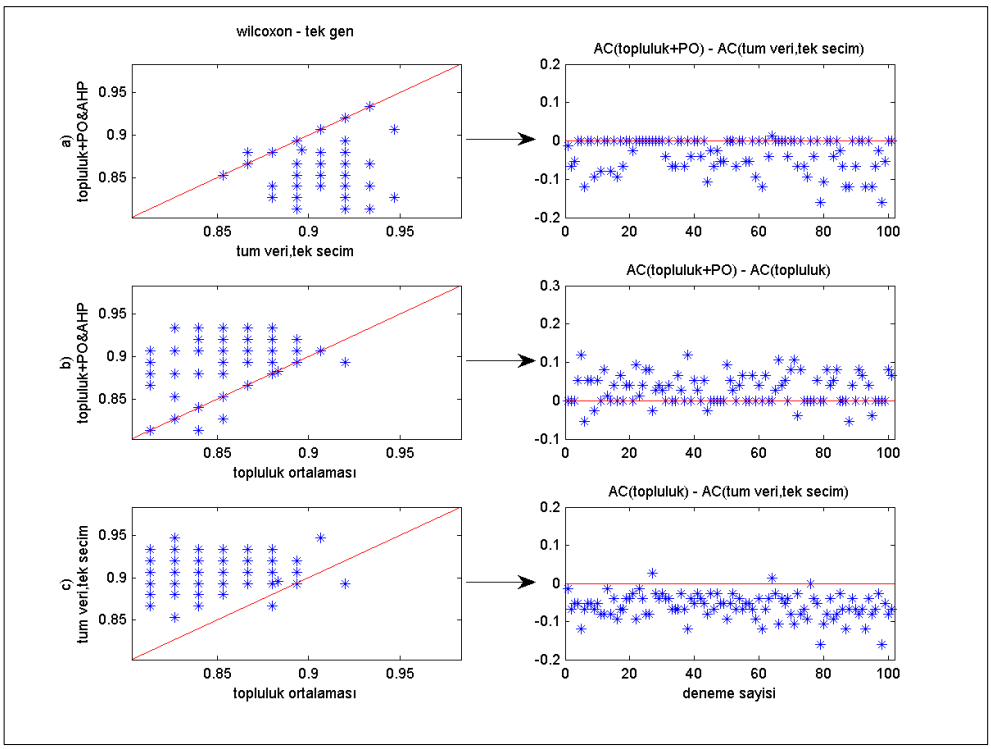


Şekil 4.39:PO Tarafından Seçilen Kadar Genin Karşılaştırılması-

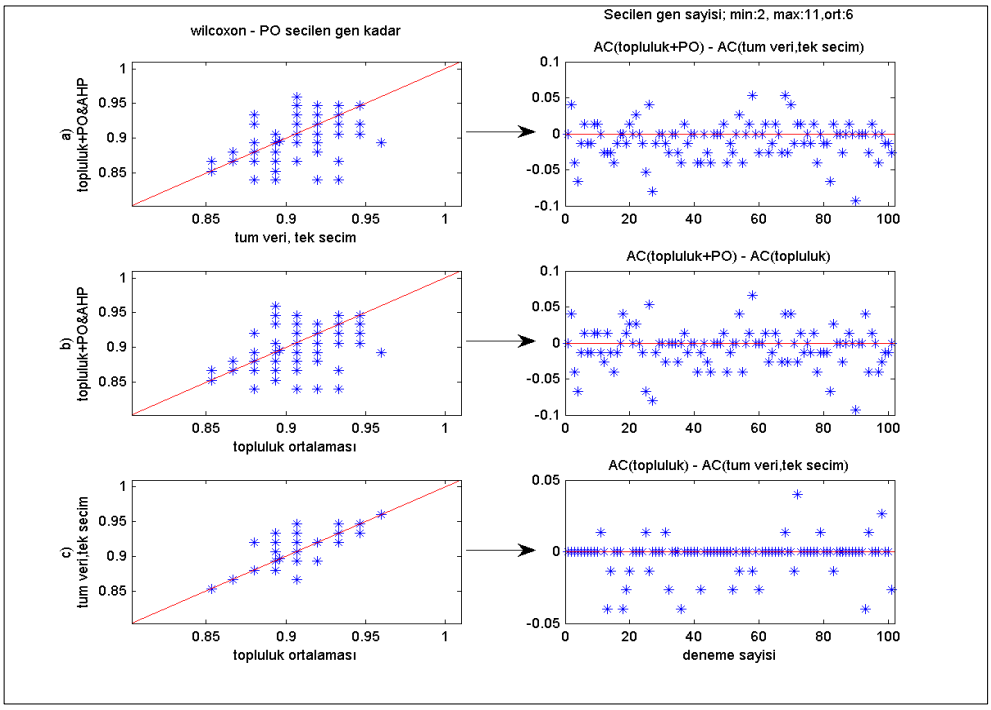
Duke Verisi, *wilcoxon* Yöntemi



Şekil 4.40:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-Duke Verisi, *wilcoxon* Yöntemi

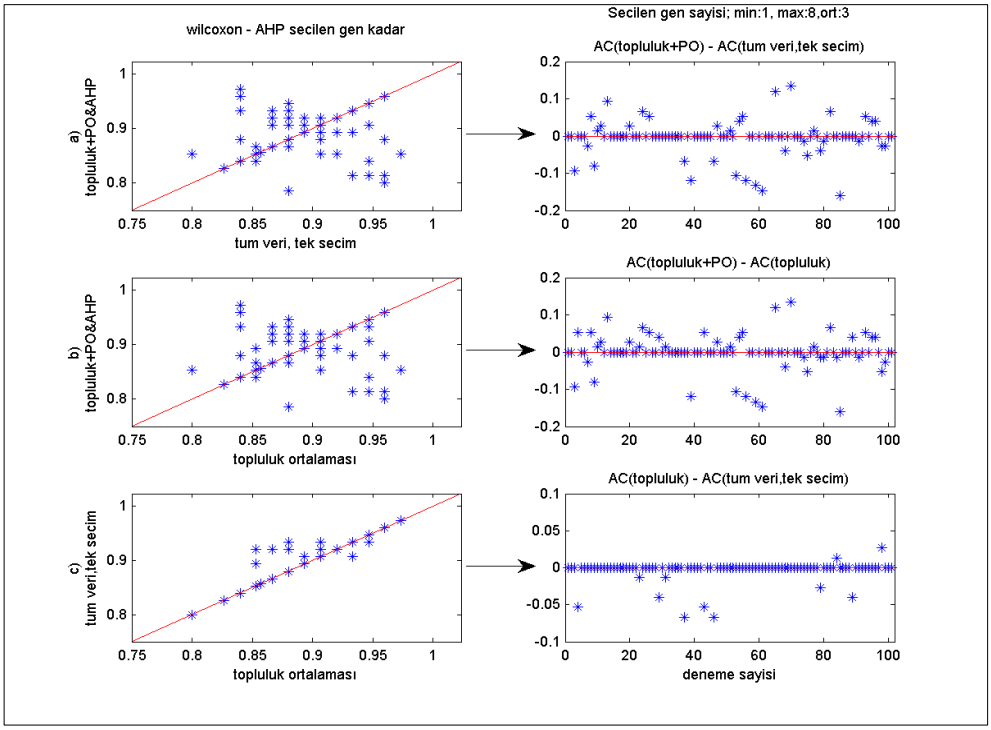


Şekil 4.41: Tek Geninkarşılaştırması-DLBCL Verisi, *wilcoxon* Yöntemi



Şekil 4.42:PO Tarafından Seçilen Kadar Genin Karşılaştırılması

DLBCL Verisi, *wilcoxon* Yöntemi



Şekil 4.43:AHP Tarafından Seçilen Kadar Genin Karşılaştırılması-

DLBCL Verisi, *wilcoxon* Yöntemi

4.3HİPERTANSİYON SONUÇLARI

Bu bölümde, PO&AHP karma yönteminin, öznitelik seçim metodu ve veri kümesi çeşitliliğinden en az etkilenen bir yöntem olduğunu göstermek amacıyla, hipertansyon veri kümesi üzerinde gerçeklenen farklı bir uygulaması incelenmiştir.

Bilindiği gibi, öznitelik derecelendirme yöntemleri öznitelikleri sınıf ayırabilirlik düzeylerine göre derecelendirir. Bunun için özniteliklerin sınıf bilgisi ile olan alakaları farklı yöntemlerle değerlendirilir. Derecelendirme yöntemlerine bağımlılığı önlemek için, sıkça kullanılan beş derecelendirme metotu (ttest, entropy, bhattacharyya, roc, wilcoxon) kullanılmıştır. Metotlar MATLAB Bioenformatik Toolbox kullanılarak uygulanmıştır.

Ayrıca çalışmada, örnek bağımlılığını en aza indirgemek amacıyla öznitelik derecelendirme işlemin uygulanırken birini dışarıda bırak (LOO) çapraz geçerleme metodu kullanılmıştır. LOO işleminde her tekrarda bir örnek dışarıda bırakılır, sonra aynı öznitelik derecelendirme algoritması N-1 örneği kullanarak genleri derecelendirir. Bu işlem her örnek dışarıda bırakılana kadar tekrar edilir, böylece işlem N defa çalıştırılır

Tüm öznitelikler (22184 adet gen) seçilen beş derecelendirme metodu ile N defa derecelendirilmiş ve her bir derecelendirme metodu için Nxf (N=örnek sayısı, f=öznitelik sayısı) boyutlarında bir derecelendirme matrisi üretilmiştir. Genlerin seçilmesinde ve derecelendirilmesinde bu beş matris kullanılmıştır.

Bu çalışmada, PO ve AHP yöntemlerinin birleştirildiği edildiği karma bir yöntem kullanılmıştır. PO yöntemi yardımıyla özniteliklerin kendi arasında yarıştırılması sağlanarak, yüksek boyutlu biyoenformatik veriler daha düşük boyutlara indirgenebilmiştir.Bununla birlikte,seçilen genlerin kendi arasında bir önem derecesi yoktur. Bu problemin çözümü için, seçilen genlerAHP yöntemi ile önceliklendirilmiştir.

Pareto Optimallik yönteminin amacı, her bir metot için hastalıkla ilişkili genleri seçmektir. Hipertansiyon ile ilişkili genler için Pareto Optimallik seçim sonuçları **Tablo 4.11**'de gösterilmiştir. Bu işlem sonucunda, 5 PO seçim kümesi için (herbir

derecelendirme metodu ile) toplam 26 farklı gen seçilmiştir ve daha sonra bu genler AHP ile derecelendirilmişlerdir (**Tablo 4.12**).

Tablo 4.11: Pareto-Optimal Kümelerdeki Genler

Metot	#G	Gen ID
Ttest	4	2782, 10621, 11249, 12376
Entropy	18	418, 2161, 3303, 3911, 4031, 4560, 4621, 5011, 9053, 9632, 9916, 11740, 11945, 11983, 15277, 16695, 17213, 19830
Bhattacharyya	18	418, 2161, 3303, 3911, 4031, 4560, 4621, 5011, 9053, 9632, 9916, 11740, 11945, 11983, 15277, 16695, 17213, 19830
Roc	5	1711, 5179, 5838, 10621, 17210
Wilcoxon	3	5179, 5838, 17210

#G seçilen genlerin sayısını göstermektedir.

Tablo 4.12: Genlerin AHP Skorları

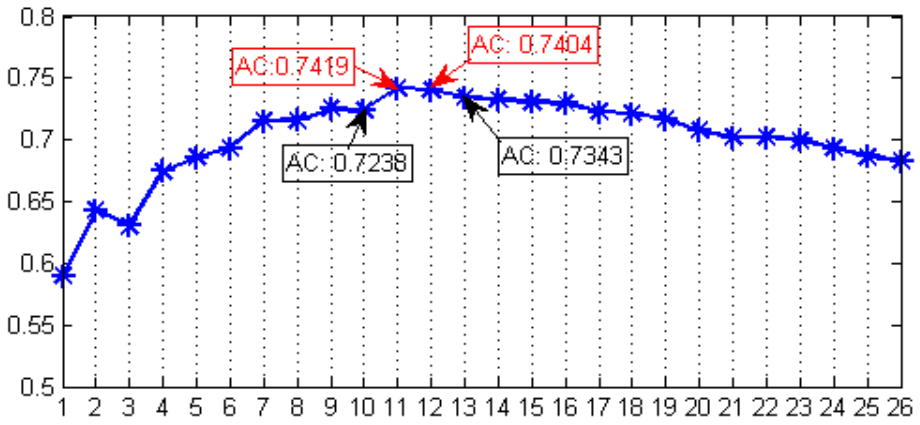
AHP skor	ID	Sembol	AC	Sen	Spe
0.088	17210	RPL7	0.590	0.318	0.844
0.086	5179	NULL	0.631	0.623	0.638
0.086	5838	FLJ35728	0.608	0.545	0.666
0.068	1711	SEC61A2	0.634	0.601	0.665
0.064	10621	NULL	0.608	0.673	0.546
0.062	12376	RFX2	0.611	0.595	0.626
0.059	11249	MIST	0.603	0.599	0.606
0.057	17213	ITGB1	0.598	0.362	0.820
0.052	2782	NULL	0.574	0.599	0.550
0.052	19830	CIB1	0.578	0.390	0.754
0.050	418	SEC61B	0.559	0.388	0.718
0.046	4621	RPS17	0.558	0.515	0.598
0.038	11983	S100A2	0.604	0.659	0.553
0.023	11740	STK25	0.490	0.476	0.503
0.019	9916	PRKCB1	0.487	0.398	0.572
0.018	11945	PSMB1	0.480	0.567	0.398
0.014	16695	NXF3	0.520	0.669	0.379
0.014	4560	SPF45	0.477	0.374	0.573
0.011	15277	C6orf153	0.491	0.550	0.435
0.008	4031	MBD6	0.473	0.341	0.597
0.007	2161	ETS2	0.489	0.344	0.624
0.006	5011	KIS	0.503	0.345	0.651
0.004	3911	HIP-55	0.548	0.585	0.514
0.004	9632	NULL	0.508	0.452	0.561
0.003	9053	RER1	0.517	0.551	0.485
0.003	3303	SYNCOILIN	0.438	0.314	0.554

AHP yönteminde tüm alternatifler eşit öneme öneme sahipseler, alternatiflerin AHP skorları 1/alternatif sayısı (1/26=0.38) ile belirlenen varsayılan önceliğe eşit olacaktır. **Tablo 4.12**'den görüleceği üzere, seçilen genlerden 12'si varsayılan öncelikten daha

büyük skora sahip olduklarından, bu genler diğerlerinden daha önemli kabul edilmişlerdir.

Burada, AHP ile seçilen gen kümelerinin başarımını değerlendirmek için 10-kat çapraz geçerleme ve doğrusal ayrıştırma analiz sınıflandırıcısı kullanılmıştır. Bir modelin sınıflandırma başarısı veri kümesi varyasyonuna bağımlı olduğundan, bu bağımlılıktan kaçınmak amacı ile 2000x10 kat çapraz geçerleme seti oluşturulmuştur. Seçilen genlerin rapor edilen sınıflandırma başarısı 2000 çapraz geçerleme setinin ortalamasıdır.

Sıralanmış gen kümesinin başarımını hesaplamak için arttırımlı altküme değerlendirme süreci uygulanmıştır. Bu süreçte, her adımda, başarısı değerlendirilecek gen kümesine seçilmiş, sıralanmış genlerden bir tane eklenmiştir. Böylece her bir tekrarda alt kümenin büyüklüğü arttırılmıştır. Gen kümesinin sırası **Tablo 4.12** 'de görüldüğü gibi AHP tabanlı derecelendirmeden elde edilmiştir. Bu tarz bir değerlendirme sürecindeki ilk amacımız, en iyi başarıya ve en az sayıda gene sahip gen alt kümesini seçmektir. İkinci amacımız ise, AHP tabanlı sıralamanın iyi bir başarıya sahip olduğunu göstermektir. **Şekil 4.44** AHP tabanlı sıralanmış gen kümesinin arttırımlı alt kümelerinin başarımlarını (AC) göstermektedir.

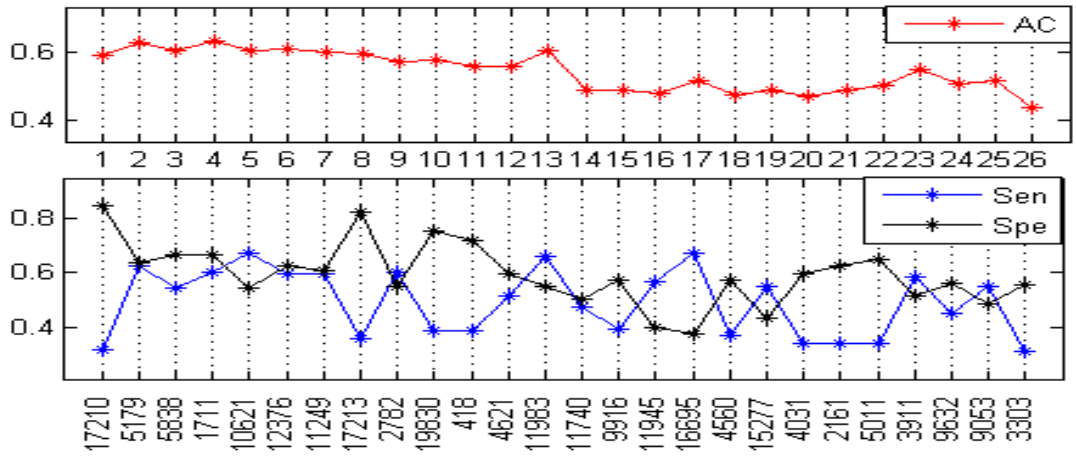


Şekil 4.44: AHP Tabanlı Sıralama Arttırımlı Alt Küme Doğrulukları

Arttırımlı alt küme başarımları, AHP skorları varsayılan öncelikten daha büyük olan ilk 12 geni içeren alt kümenin en iyi olduğunu göstermektedir. Önemli genler **Tablo 4.12**'de gri zemin ile belirtilmiştir. Tüm genlerin bir arada kullanılması durumunda başarı ~%68 iken, daha önemli genlerin(ilk 12 gen) başarısı %74'ü geçmiştir. Grafikte

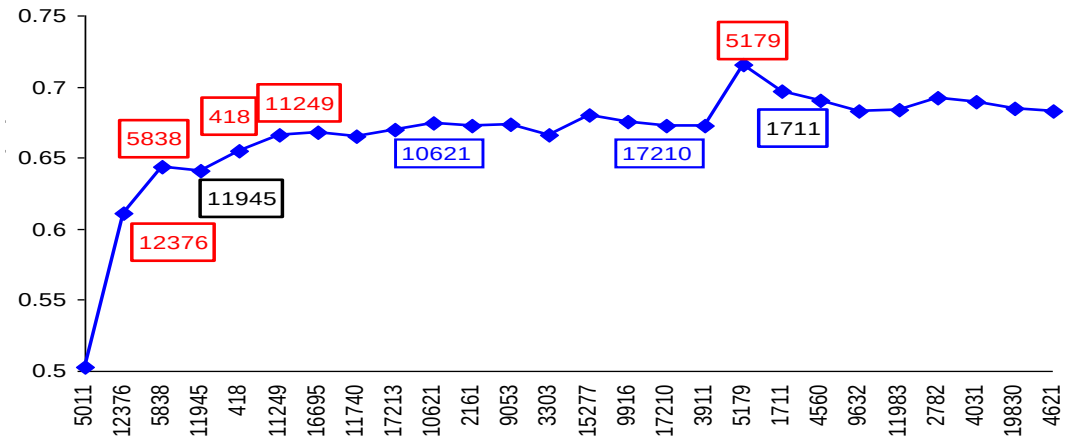
daha önemli genlerin eklenmesi ile başarının adım adım arttığı, daha az önemli genlerin (13-26 arası) eklenmesi ile de yavaş yavaş düştüğü görülmektedir.

Gen kümemizin birlikteliğinin anlamlılığını ölçmek amacı ile genlerin tek başlarına olması durumundaki başarıları ölçülmüştür (**Şekil 4.45**, **Tablo 4.12**). Genlerin birlikte kullanılması ile ulaşılan doğruluğun hiç bir genin tek başına elde edemeyeceği açıkça görülmektedir. Ayrıca önemli ve daha az önemli genlerin doğruluk düzeylerindeki farklılık AHP tabanlı sıralamanın bir doğrulamasıdır.



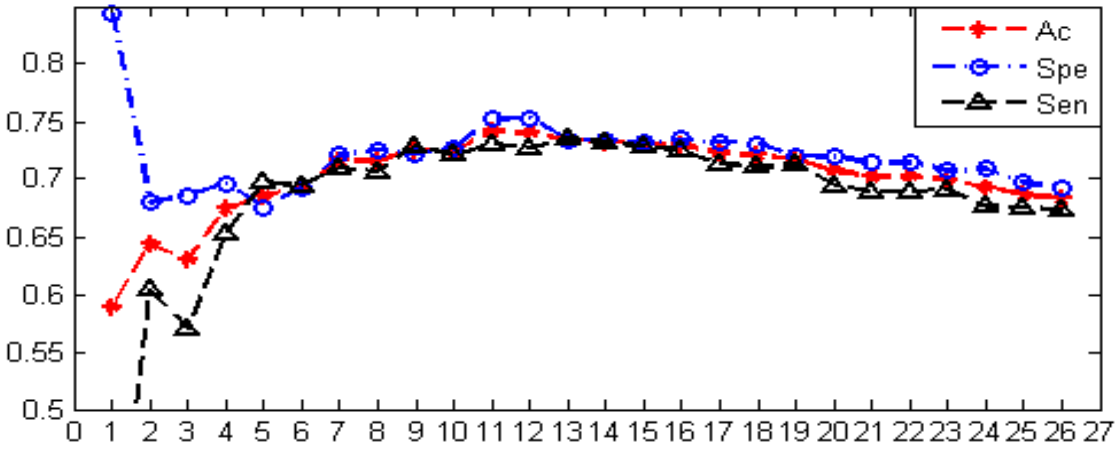
Şekil 4.45: Seçilen Genlerin Tekli Başarımları

Ayrıca AHP tabanlı sıralamayı doğrulamak için rastgele sıralanmış gen kümesinin arttırımlı altküme doğrulukları (AC) hesaplanmıştır. **Şekil 4.46**'de, genellikle önemli genler eklendikçe başarının arttığını görülmektedir. Aksine AHP skoru varsayılan önceliğin altında olan daha az önemli genler eklendiğinde AC azalmaktadır. 5011 geni ile elde edilen %50 başarı, daha önemli görülen 12376 genin eklenmesi ile bir anda %60 oranını geçmiş, 5838 genin eklenmesi ile de %65 yükselmiştir. Bu üçlüye daha az önemli 11945'in eklenmesi başarıyı biraz azaltmış, tekrar önemli genlerin eklenmesi ile başarı yine yükselmeye başlamıştır. **Şekil 4.46**'de, başarıyı arttıran önemli genler kırmızı çerçeve ile beklenen seviyede artış sağlamayan bazı önemli genler ise mavi çerçeve ile gösterilmiştir. Diğer taraftan bazı önemli genler (1711) başarıyı negatif etkilemiştir. Bizce önemsiz genlerin altkümeye önemli genlerden önce eklenmesi buna sebep olmuştur. Sonuç olarak, bu sonuçlar AHP tabanlı sıralamayı doğrulamıştır.



Şekil 4.46: Rastgele Sıralama Arttırımlı Alt Küme Doğrulukları

Önerilen karmametot ile seçilen ve sıralanan genlerin doğruluk (Accuracy-AC), özgüllük (specificity-Spe) ve duyarlılık (sensitivity-Sen) değerleri **Şekil 4.47**'te gösterilmiştir. Daha önemli olduğunu düşündüğümüz ilk 12 gen ile 2000x10 kat çapraz geçerleme setlerinin ortama doğruluğu %74 oranına ulaşmıştır.

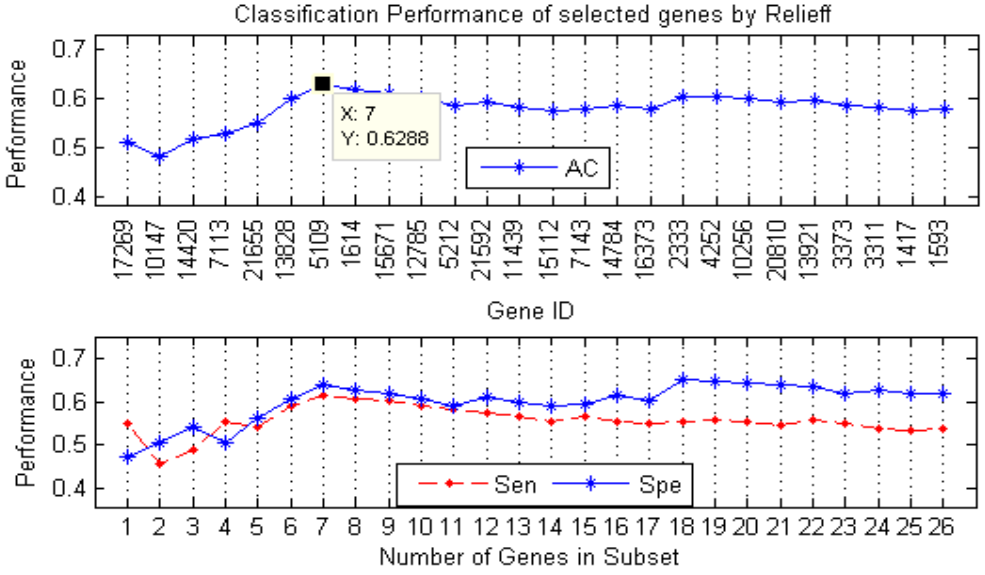


Şekil 4.47: AHP Tabanlı Sıralama Arttırımlı Alt Küme Başarımları

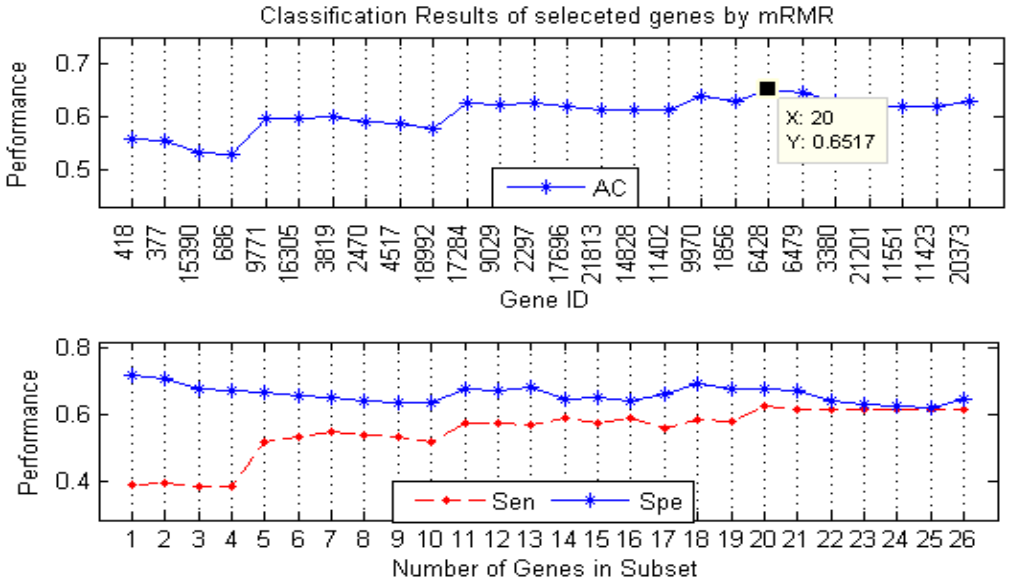
Önerilen karmametot ile seçilen ve sıralanan genlerin doğruluk (Accuracy-AC), özgüllük (specificity-Spe) ve duyarlılık (sensitivity-Sen) değerleri Şekil 4.20'te gösterilmiştir.

Son olarak, önerilen karma yöntemin başarısını literaturde çokça kullanılan öznitelik seçme yöntemleri ile karşılaştırmak amacı ile çok bilinen mRMR ve Relieff yöntemlerini kullanılmıştır. Tüm veri kümesi bu iki metoda girdi olarak verilmiş ve 26 tane gen seçimi yapılmıştır. Daha sonra bu metotlarca seçilen gen kümelerinin de

arttırımlı alt küme başarılarıhesaplanmıştır. Metotların başarımları **Şekil 4.48** ve **Şekil 4.49**'de gösterilmiştir. Relieff ve mRMR yöntemlerinin en yüksek doğruluk(AC) oranları sırasıyla %62 ve %65 olarak ölçülmüştür. Önerilen karma yöntem ile seçilen genlerin başarısı ise %74'tür. Bu sonuçlar önerilen yöntemin son zamanlarda sıkça kullanılan bu iki metottan daha iyi sonuçlar ürettiğini göstermiştir. Burada önerilen yöntemin veri kümesindeki çeşitliliği temsil etme gücünden dolayı daha başarılı olduğunu söyleyebiliriz.



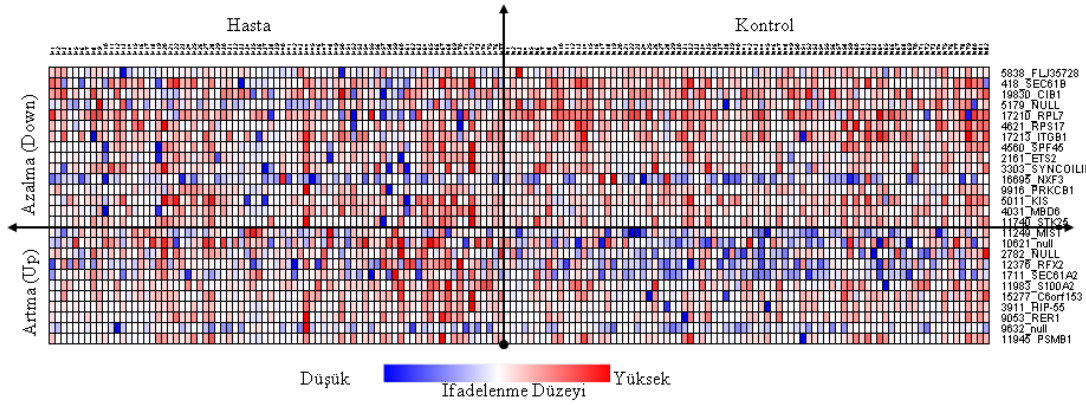
Şekil 4.48: Relieff ile Seçilen Genlerin Arttırımlı Alt Küme Başarımları



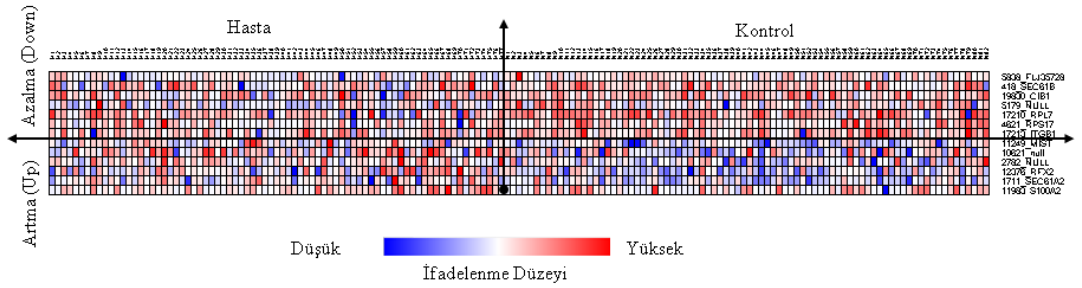
Şekil 4.49:mRMR ile Seçilen Genlerin Arttırımlı Alt Küme Başarımları

Seçtiğimiz genlerin hipertansiyon hastalığı ile ilişkisinin biyolojik düzeyde incelemesi yapılmıştır. Klasik mikrodizi analizlerinde hasta-kontrol gruplarında farklı ifadelenmiş genlerin bulunması için bir p-değeri hesaplanmaktadır. Bu noktada ilk çalışma tarafında hesaplanmış p-değerleri ve Genepattern[61] programı tarafından hesaplanmış p-değerleri **Tablo 4.13**'de sunulmuştur. Ayrıca genlerin genepattern programı tarafında hesaplanmış olan 2 grup arasındaki değişim oranını gösteren kat değişimi (fold change) ve regülasyondaki etki bilgileride sunulmuştur.

Klasik mikrodizi analizlerinde genlerin regülasyondaki etkileri ve hasta-kontrol grubu dikkate alınarak ifadelenme düzeylerindeki yoğunluğun görselleştirilmesi amacı ile yoğunluk haritası (heat map) kullanılır. Haritada satırlar genleri, sütunlar bireyleri ve her hücre genin ifadelenme düzeyini temsil etmektedir. Kırmızı hücreler yüksek ifadelenme düzeyini mavi hücreler düşük ifadelenme düzeyini gösterir. Önerilen karma metot tarafından seçilen 26 genin ve ilk AHP sıralamasında göre daha önemli olan 13 genin yoğunluk haritası Genepattern [61] kullanılarak oluşturulmuşturve sonuçlar sırasıyla **Şekil 4.50** ve **Şekil 4.51**'de gösterilmiştir. Genler hasta grubuna göre regülasyondaki artma (up) ve azalma (down) etkilerine göre gruplanmış, ilk makalenin p-değerlerine göre de sıralanmıştır.Grafiklerde azalma bölümüne bakıldığında sağlıklı grup için genlerin ifadelenme seviyesinin yüksek olduğu buna karşılık hasta grupta azalma olduğu açıkça görülmektedir (hasta grubunda kırmızılıklar azalmış, beyaz ve mavilikler artmıştır). Diğer taraftan artma bölümü incelendiğinde; sağlıklı grupta genlerin ifadelenme düzeylerinin düşük olduğu görülmektedir.Seçilen genlerin klasik mikrodizi açısıyla bakıldığında başarılı bir ayrım yapabildiği görülmektedir.**Şekil 4.51**'de, ilk 13 gen (AHP skoru varsayılan önceliğe eşitS100A2 geni de dâhil edilmiştir) ile oluşturulan yoğunluk haritasında azalma-artma ve sağlıklı-kontrol farkları daha net görülmektedir.



Şekil 4.50: Karma Metot Tarafından Seçilen 26 Genin Yoğunluk Haritası



Şekil 4.51: Karma Metot Tarafından Seçilen 13 Genin Yoğunluk Haritası

Tablo 4.13incelendiğinde p-değerleri arasında benzerlik görünsede ayrıntılı bakıldığında, seçilecek *p* eşik değeri için bazı genlerilk çalışmaya göre en iyi listeye girerken genepattern'a göre liste dışında kalacaktır yada tam tersi durum olabilir. Örneğin$p$ < 0.05 için RPS17 geni,p< 0.001 için FLJ35728 geniilk çalışmada değerlendirmeye alınmazken, genepattern'a göre alınacaktır. Bu örnekler sadece 26 genlik bir listedeki farklılıklara işaret eder, oysa 22184 gen için oluşan farklılıklar çok daha fazla olabilir. Ayrıca kat değişimi de klasik mikrodizi analizlerinde dikkate alınan bir diğer parametredir. Ancak veri kümesini kullanan orijinal çalışma kat değişimini hiç kullanmamıştır. Bu sonuçlardan hareketle, klasik mikrodizi analizlerindeuygulanan farklı metotların beklendiği gibi farklı sıralamalar oluşturduğu görülmektedir.

Tablo 4.13: Seçilen Genlerin Klasik Mikrodizi Analiz Bilgileri

ID	Sembol	Regülasyon	p-değeri Lynn ve diğ.	p-değeri genepattern	kat değişimi
17210	RPL7	Azalma(Down)	0.014422	0.015997	1.115524
5179	**NULL**	Azalma(Down)	0.008113	0.006199	0.912796
5838	FLJ35728	Azalma(Down)	0.0027465	0.00100	0.872342
1711	SEC61A2	Artma(Up)	0.0011572	0.00120	1.136973
10621	**NULL**	Artma(Up)	0.0003896	0.00020	1.381725
12376	RFX2	Artma(Up)	0.0007185	0.00140	1.084153
11249	MIST	Artma(Up)	0.0002108	0.00020	0.784146
17213	ITGB1	Azalma(Down)	0.094046	0.10338	1.052475
2782	**NULL**	Artma(Up)	0.0005289	0.00060	0.889242
19830	CIB1	Azalma(Down)	0.00752	0.00460	1.11719
418	SEC61B	Azalma(Down)	0.0051878	0.00320	1.069486
4621	RPS17	Azalma(Down)	0.05436	0.04179	1.052613
11983	S100A2	Artma(Up)	0.048968	0.01740	1.157612
11740	STK25	Azalma(Down)	0.38201	0.52350	1.02287
9916	PRKCB1	Azalma(Down)	0.26478	0.30514	1.019204
11945	PSMB1	Artma(Up)	0.56767	0.77664	1.010845
16695	NXF3	Azalma(Down)	0.24839	0.25335	0.864656
4560	SPF45	Azalma(Down)	0.13511	0.13057	0.85339
15277	C6orf153	Artma(Up)	0.0741	0.05699	1.028677
4031	MBD6	Azalma(Down)	0.34029	0.35653	1.029759
2161	ETS2	Azalma(Down)	0.13629	0.14117	1.044442
5011	KIS	Azalma(Down)	0.26572	0.27734	1.027795
3911	HIP-55	Artma(Up)	0.12593	0.12637	1.036947
9632	**NULL**	Artma(Up)	0.36401	0.47391	0.925556
9053	RER1	Artma(Up)	0.28064	0.32973	1.022914
3303	SYNCOILIN	Azalma(Down)	0.22046	0.23175	1.042901

Lynn ve arkadaşlarının[36] ve genepattern sonuçlarına göre, AHP skoruna göre daha önemli olduğunu belirttiğimiz ilk 12 genden 9'unun hipertansiyon hastaları ve sağlıklı kontrol gurubu arasındaki ifadelenme düzeyindeki farklılık istatistiksel olarak anlamlıdır ($p < 0.01$). Diğer yandan $p < 0.05$ eşik seçildiğinde ise ilk 13 genden (AHP öncelik >=varsayılan AHP önceliği) sadece birinin (ITGB1) ifadelenme düzeyindeki fark anlamlı değildir. Ayrıca, SEC61A2, RFX2, MIST, CIB1, SEC61B isimli seçtiğimiz genler, ilk çalışmanın da [36] hipertansiyon ile ilişkilendirdiği genler arasında yer almaktadır.

İlk çalışmanın[36]nın p-değerlerine bakıldığında 103 gen, [61]'deki çalışmaya göre ise 111 gen seçilmekteyken, önerilen karma metot kullanıldığında sadece 26 gen seçilmiştir. Karma yöntemdeki AHP önceliklendirme sıralaması dikkate alınırsa toplam 12 gen seçilmesi yeterli olmaktadır. Genlerin hastalık ilişkilerinin incelenmesinde az sayıda gen kullanımının önemli bir zaman kazancı sağlayacağı aşikârdır. Bu çalışmada önerilen yöntem ile toplam 12 gen seçilerek daha yüksek doğruluk sağlanabileceği ve hızlı mikrodizi analizi yapılabileceği gösterilmiştir.

Farklı bir gözlem olarak, beş derecelendirme metodu kullanılarak seçilmiş genlerin AHP sıralaması incelendiğinde*; ttest, roc ve wilcoxon* tarafından seçilmiş olan tüm genlerin (1711, 2782, 5179, 5838, 10621, 11249, 12376,17210) ilk 12 gen arasında olduğu ve 17210 hariç diğerlerinin p-değerlerinin 0.01'den küçük olduğu görülmektedir. Öte yandan, diğer derecelendirme yöntemleri*Bhattacharyya ve entropy*metotlarından seçilen genlerden ise sadece 4 tanesinin (418, 4621, 17213, 19830) ilk 12 gen içinde yer aldığı ve sadece 2 tanesinin (418, 19830) p-değerlerinin 0.01'den küçük olduğu tespit edilmiştir. Bu sonuçlardan *ttest, roc ve wilcoxon* ile daha başarılı gen seçimi yapıldığı, *bhattacharyya ve entropy* ile hastalıkla ilişkisi daha düşük genlerin seçildiği ve bu metotların seçime gürültü karıştırdığı söylenebilir.

Diğer araştırma konusu, seçilen genlerin hastalık ilişkili olup olmadığının tespitidir.Hastalıklara neden olan doğru aday genleri belirlemek, gen-gen etkileşimlerini çözebilmek ve hastalığa neden olan yolakların aydınlatılabilmesi için çeşitli alternatif analiz araçlarına ihtiyaç duyulmaktadır[62]. Bu amaçla klasik microarray analiz yöntemlerinin yanında makina öğrenme yöntemleri de son zamnlarda sıkça kullanılmaktadır [63-66]. Bu bölümde, önermiş olduğumuz makine öğrenme yöntemleri temelli karma yöntemin seçmiş olduğu genlerin hipertansiyon ile ilişkileri ele alınmıştır.

Hipertansiyon, dünya genelinde kardiyovaskuler hastalıklar arasında en fazla morbidite ve mortaliteye sebep olan hastalıklardan birisidir. Dünya Sağlık Örgütü kriterlerine göre kan basıncının 140/90 mmHg dan daha fazla ölçüldüğü kronik patofizyolojik bir durumdur [67]. Etiyolojik açıdan primer (esansiyel) ve sekonder hipertansiyon olmak üzere iki grupta sınıflandırılmaktadır. Sekonder hipertansiyon %5-10'luk kısmını oluştururken günümüzde vakaların %90-95'ini primer hipertansiyon

oluşturmaktadır ve etiyolojisi bilinmemektedir. Temel olarak genetik ve çevresel faktörlere bağımlı olduğu düşünülmektedir [68].

Kompleks genetik bir hastalık olan hipertansiyonun bildirilmiş ailesel bir geni yoktur. Ancak aile ve ikizlerde yapılan çalışmalarda kan basıncının değişkenliğinde kalıtımın yaklaşık %30-50, çevresel faktörlerin yaklaşık %50 etkisinin olduğu saptanmıştır [62]. Artmış kan basıncına neden olan genlerin sayısı tam olarak bilinmemektedir [69]. Çok genli, çok fonksiyonlu hastalıkları çözmek için yapılan GWAS çalışmalarında 20'den fazla kromozomal bölgede hipertansiyon ile ilişkili gen tanımlanmıştır [62].Bu genler direk veya dolaylıyollarla hastalığa neden olurlar.

Önerilen metot tarafından seçilen genlerin hipertansiyon ile ilişkisi intenette ve Genemania[70] programı ile literatürdekiçalışmalarda araştırılmıştır. Elde edilen sonuçlar **Tablo 4.14**'de sunulmuştur.Tabloda yer alan bazı kavramlar ise aşağıda açıklanmıştır. Kavramlar Genemaina[70] programının yardım menüsünden alınmıştır.

- ✓ **Gen-genetkileşimi:** genler, gen ürünleri (protein, mRna) ve yolaklar (pathway) arasındaki fonksiyonel ilişkilerdir[71]. Biyolojik süreçlerde etkileşim halinde olan genler birbirlerinin ifadelenme düzeylerini arttırır yada azaltırlar [72]. Süreç boyunca devam eden bu etkileşim sürecin devamlılığı için çok önemlidir.
- ✓ **Genetik Etkileşim:**Eğer bir genin sentezlenmesinde meydana genlen bozulmanın etkileri diğer genin sentezinde meydana gelen bozulma ile düzelilebiliyorsa bu iki gen fonksiyonel olarak ilişkilidir
- ✓ **Fiziksel Etkileşim:** Protein-protein etkileşimini ifade eder. Yani 2 genin ürünleri olan proteinlerin etkileşimi gözleniyorsa bu iki gen genetik etkileşim halindedir.
- ✓ **Gen kolokalizasyonu:** 2 genin aynı dokuda yada hücrenin aynı bölgesinde birlikte ifadelenmesidir.
- ✓ **Birlikte İfadelenme:** 2 geninin ifadelenme düzeylerinin benzerliğini (eş zamanlı azalma-artma) gösterir. Benzer ifadelenme düzeyi genlerin genetik etkileşim halinde olduklarını gösterir.
- ✓ **Bir protein domainini paylaşma:** Eğer 2 gen aynı protein domainine sahipse, bu genlerin ürünleri bağlantılıdır.
- ✓ **Aynı yolakta yer alma:** Eğer 2 gen bir yolaktaki aynı reaksiyona katılıyorlarsa bu iki genin ürünleri bağlantılıdır.

Tablo 4.14: Seçilen Genlerin Hipertansiyon İle İlişkisi

SEC61A2 ve SEC61B	SEC61 kompleksi, Endoplazmik Retikulum (ER) membranında protein translokasyonunda görev yapar. SEC61A2 ve SEC61B genleri SEC61 kompleks protenin a ve b domainidir. Bu kompleks ApoB geni ile etkileşim[73] ve fiziksel etkileşim[74] halindedir. APOB geni de birincil hipertansiyon için biyo-işaretçi olarak tanımlanmıştır[75]. Diğer bir çalışmada ise APOB geninin hastalıkla ilişkisi rapor edilmiştir[76].
PRKCB1	Hücre içi sinyal yolağında meydana gelebilecek bozukluk hipertansiyonun da içinde bulunduğu birçok hastalığa sebep olmaktadır. PRKCB1 geni ise yolakta gerçekleşen reaksiyonlarda temel rol oynamaktadır[77, 78].
ELL2	Serum protein miktarlarındaki değişimler hipertansiyonla ilişkili olan böbrek ve kalp damar gibi hastalıklara neden olmaktadır. Serum-proteinlerinin ifadelenme düzeylerinin araştırıldığı bir çalışmada ELL2 geni hastalıklarla ilgili aday bölge olarak gösterilmiştir[79]. Ayrıca Bu genin SEC61B ile genetik etkileşimi vardır[80].
RPL7	Bu genin önerilen metot tarafından seçilen diğer 2 gen (SEC61B ve RPS17) ile aynı yolakta yer aldığı raporlanmıştır[81]. Bu genlerden SEC61B geni ise hipertansiyon için biyo-işaretçi olarak tanımlanan APOB geni ile ilişkilendirilmiştir[73]. Ayrıca RPL7 geni direk olarak ABOP geni ile de kolokalizedir[82]. Ayrıca bu genin PRKCB ile de genetik etkileşimi rapor edilmiştir[80]. PRKCB geni ise hipertansiyona yol açan yolakta önemli bir gen olarak bildirilmiştir.
CIB1	Kalp büyümesine neden olan en önemli faktörlerden birisi hipertansiyondur ve CIB1 geninin kalp kası büyümesi ile ilişkili olduğu bilinmektedir[83]. Ayrıca bu gen OMIM'de[84] birincil hipertansiyon fenotipi ile ilişkilendirilmiş AGTR1 ile genetik etkileşimi vardır[80]. Ek olarak SEC61B ile kolokalizedir[85].
S100A2	Hayvanlar üzerinde yapılan bir gen terapisi çalışmasında, S100A2 geni düzeltildiğinde, bozulan kalp kas dokusu hücrelerinin kasılma kusurlarının düzeldiği rapor edilmiştir[86].
MIST	MIST geninin SEC61B ve SELE genleri ile genetik etkileşimi rapor edilmiştir[80]. SELE geni ise OMIM'dekan basıncı düzenlenmesi (Blood pressure regulation QTL) ile ilişkilendirilmiş[84]. Hücre iletişiminde görev yapan bu genin böbrek hücre hattı olan HEK293'de ifadelendiği gösterilmiştir[87].
ITGB1	ITGB1 geninin OMIM'de hipertansiyon ile ilişkilendirilmiş 3 farklı gen ile etkileşimi rapor edilmiştir[84]. Bu genler ile etkileşim düzeyi ve genlerin hipertansiyon ile ilişkisi aşağıda özetlenmiştir: (i) Kan basıncı düzenlenmesi ile ilişkilendirilmiş SELE geni ile ortak protein domainine sahiptir(Punta et al., 2012). (ii) PTGIS ile genetik etkileşimi rapor edilmiştir[80] ve bu gen birincil hipertansiyon ile ilişkilendirilmiştir. (iii) Birincil hipertansiyona yatkınlık fenotipi ile ilişkilendirilmiş olan ECE1 geni ile kolokolizedir [82].

PSMB1	Deney fareleri üzerinde yapılan bir çalışmada, kronik olarak oksijen yetersizliğine maruziyetin akciğer hipertansiyon ile ilişkisinin genetik açıdan değerlendirilmesi yapılmış, PSMB 1 ve PSMB6'nın seviyelerinin kronik oksijen yetersizliği sonrasında arttığı tespit edilmiştir[88]. Birincil hipertansiyonunda de bu gen aday olarak değerlendirilebilir.
KIS (UHMK1)	Son zamanlarda yapılan bir çalışma da UHMK1 geninin hipertansiyonda etkili olan bu WNK1 geni ile birlikte ifadelendiğini rapor etmiştir[89]. WNK1 geniise hipertansiyon ile ilişkisi birçok çalışma tarafında gösterilmiştir. WNK1 geninin promotor bölgesi yakınlarında konumlanmış bir SNP hipertansiyonun şiddeti ile ilişkilendirilmiştir. WNK1 geninin ifadelenme düzeyindeki artışın birincil hipertansiyonun gelişim sürecinin tahminlenmesine katkıda bulunduğu raporlanmıştır[90]. Ayrıca WNK1 ve WNK4 genlerinde meydana gelen mutasyonların hipertansiyonunda eşlik ettiği PHAIIhastalığına sebep olduğu rapor edilmiştir [91].
RPS17	RPL7 geni ile birlikte ifadelenmektedir[89], gen genetik etkileşim halindedir[80] ve aynı yolakta yer almaktadır[81, 92]
FLJ35728	SEC61B ile genetik etkileşimi rapor edilmiştir[80].
RFX2	CIB1 ile genetik etkileşimi vardır[80].
STK25	UHMK1 geni ile ortak protein domainine sahiptir [93].
RER1	CIB1 ile kolokalizedir[85].

4.4 HGDP SONUÇLARI

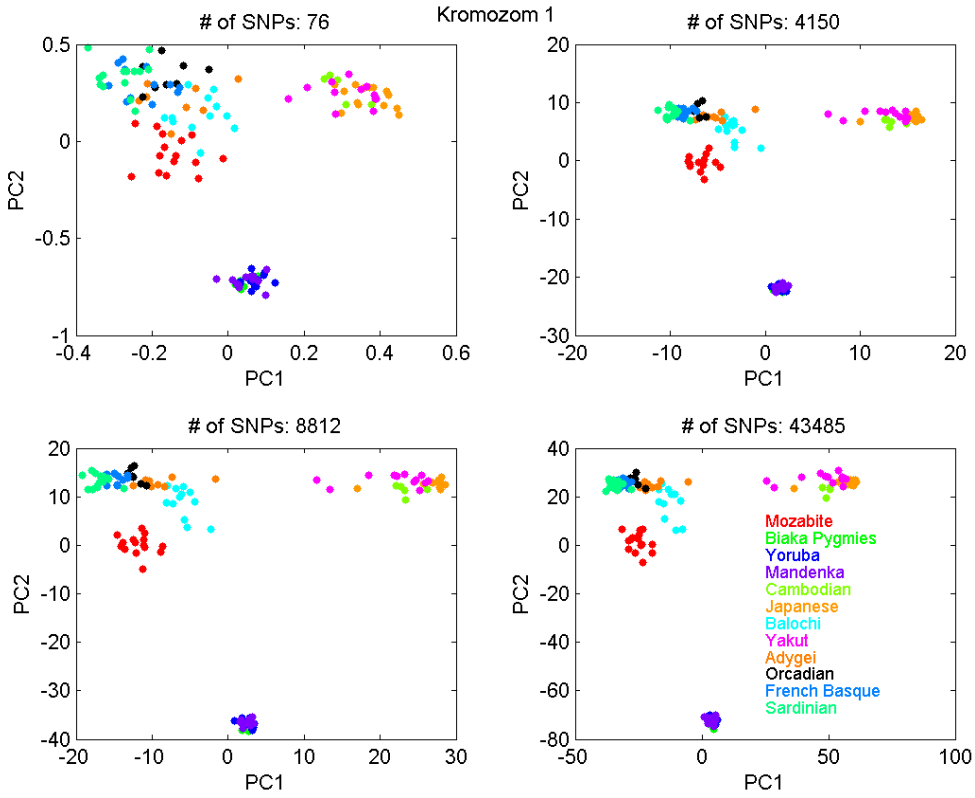
Yeni bir vaka çalışması, İnsan Genom Çeşitliliği Projesi (HGDP) verikümesinden 3 kıtaya yayılan (Afrika, Asya, Avrupa) 4 etnik kökenden,grup içi genetik çeşitliliği mümkün olduğunca az olan 12 grup üzerinde gerçeklenmiştir. Bu gruplar, **Tablo 3.3**'de coğrafî koordinatları ve boyutlarına göre listelenmiştir. Bu seçime ilişkin ana fikir, genellikle yakın grupların genetik olarak az farklılığa sahip olmasına dayanmaktadır. Bölüm 3'de belirtildiği üzere, bu çalışmanın amacı, yüksek sınıflandırma doğruluğu (amaç-1)ile genomik ve coğrafî mesafeler arasında yüksek korelasyon (amaç-2) sağlayan SNP'leri saptamaktır.Burada, her 2 amacı dadikkate alarak seçim yapmak için, çok kriterli karar verme yöntemi olan PO yaklaşımı kullanılmıştır.

Bölüm 3.7 de anlatıldığı gibi, birinci ve ikinci temel bileşenler (PC'ler) amaç 2'nin, MI ve Relieff değerleri ise amaç 1'in gerçeklenmesi için kullanılmıştır. Burada ortaya çıkarılan 4 boyutlu veri kümesine önerilen çok kriterli öznitelik seçim yöntemi uygulanmıştır.

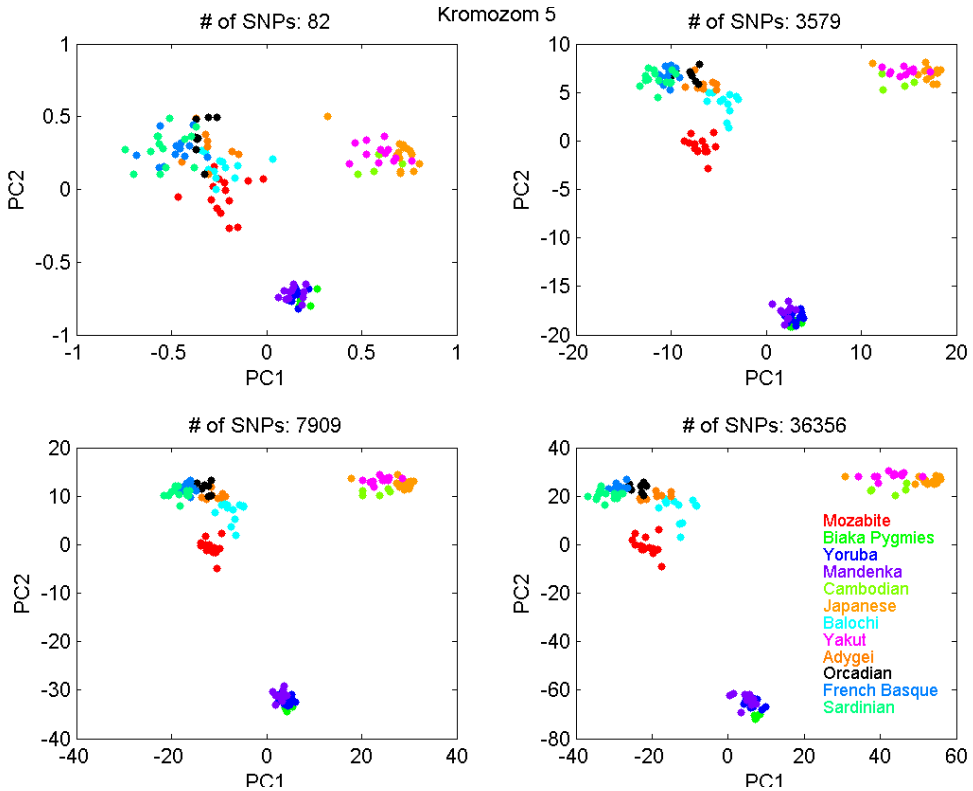
Verinin analizi için ilk olarak, her grubun yarısı eğitim kümesi; geri kalan yarısı test kümesi olarak ayrılmıştır. Her SNP'nin MI ve Relieff değeri, PC1 ve PC2 ağırlıkları eğitim kümesi kullanılarak hesaplanmıştır. Burada, mutlak değeri yüksek olan bir SNP, varyansın işaretini (pozitif veya negatif) etkilediğinden, PC yüklerinin mutlak değerleri kullanılmıştır.Bununla birlikte, 0'a yakın bir mutlak değeri olan bir SNP'nin toplam varyans ve dolayısıyla sınıflandırma üzerinde herhangi bir etkisi yoktur. 4 farklı kritere sahip (MI, Relieff, PC1, PC2) SNP'ler PO birleştirme yöntemi ile seçilmiştir.

PCA ile amaçlanan jeo-genomik dağılım arasındaki korelasyonun sağlandığının göstermek amacı ile PC1 ve PC2 düzlemlerindeki dağılımları gösterilmesidir. **Şekil 4.52-Şekil 4.56**'da sırasıyla 1,5,10,15,20. kromozomlar için örnek sonuçlar verilmiştir. Burada, PO seviyeleri kademe kademe arttırılıp, belli sayıdaki SNP'ler için genomik dağılım gösterilmiştir. Bu gösterimlerden, genomik dağılımların milletlerin coğrafik dağılımını yansıtma başarısı açıkça görülmektedir. Kullanılan SNP sayısı arttıkça harita gerçeğe yakınsamaktadır. Avrupa kıtasından seçilen Sardinian, Orcadian ve French_Basque örnekleri ile Asya kıtasından seçilen Yakut, Japanese ve Cambodian örnekleri ayrı iki küme oluştururken, Adygei ve Balochi grupları coğrafi yerleşimlerine paralel olarak bu iki küme arasında yer almışlardır. Bununla beraber Afrika kıtasından

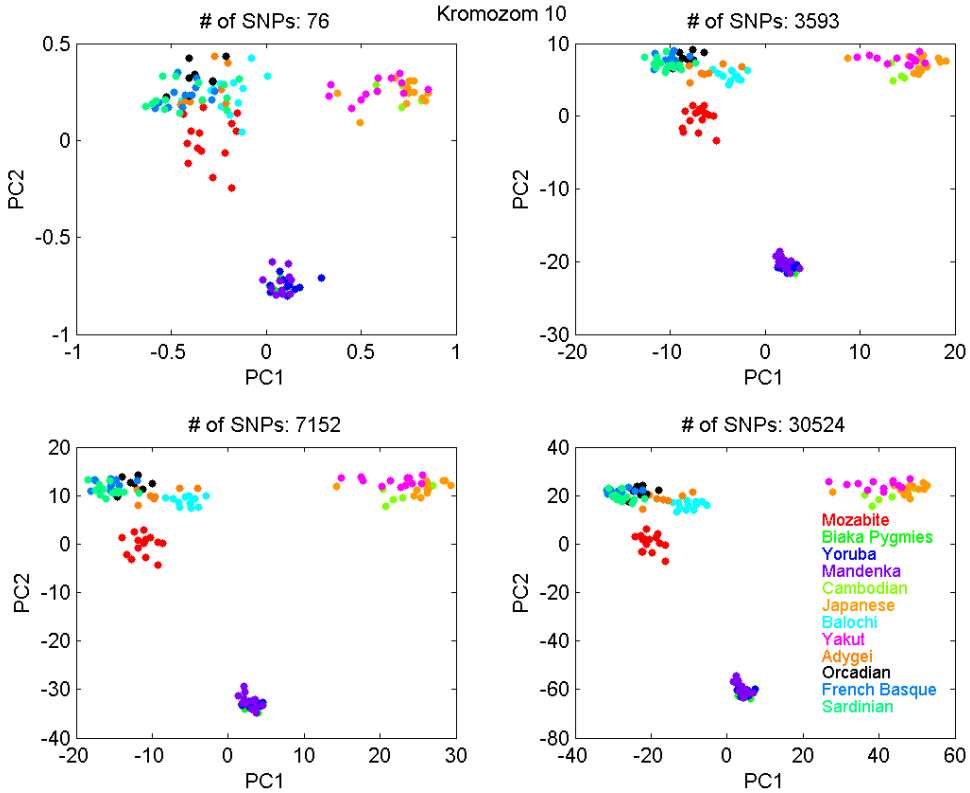
seçilen Yoruba, Mandenka ve Biaka_Pygmies grupları bir Afrika kümesi oluşturmuş ve bugünkü Fas bölgesinde yaşayan Mozabite grubu yine coğrafi konumuna paralel olarak Avrupa ve Afrika kümeleri arasında yer almıştır.



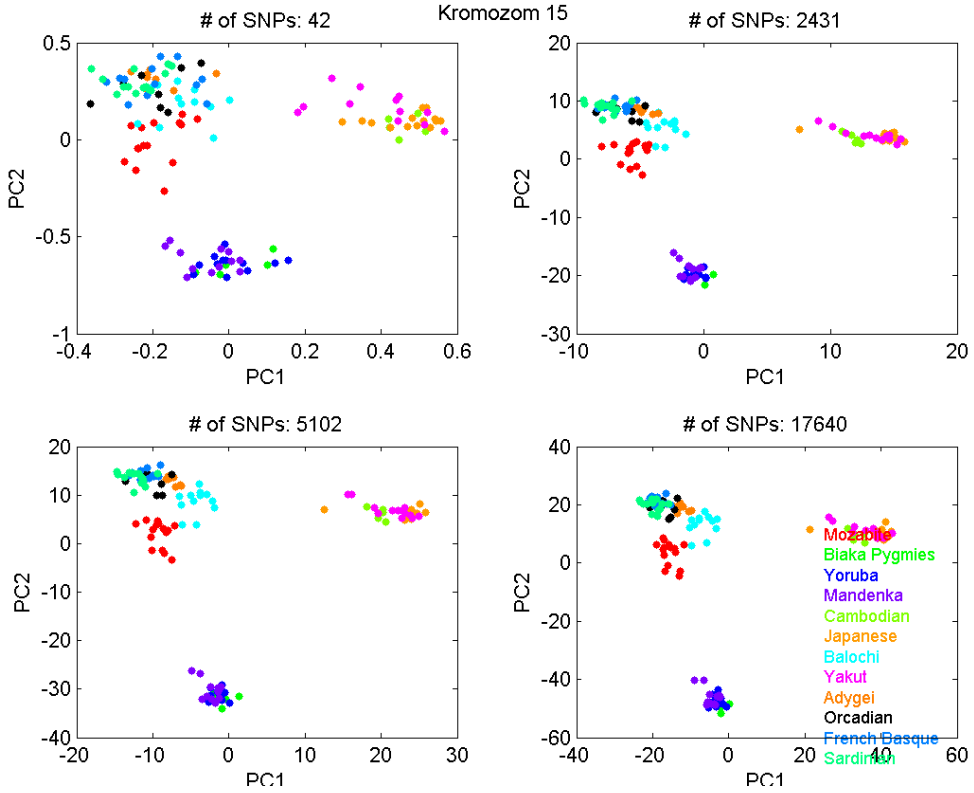
Şekil 4.52: PCA İle Genomik Dağılım (Koromozom 1)



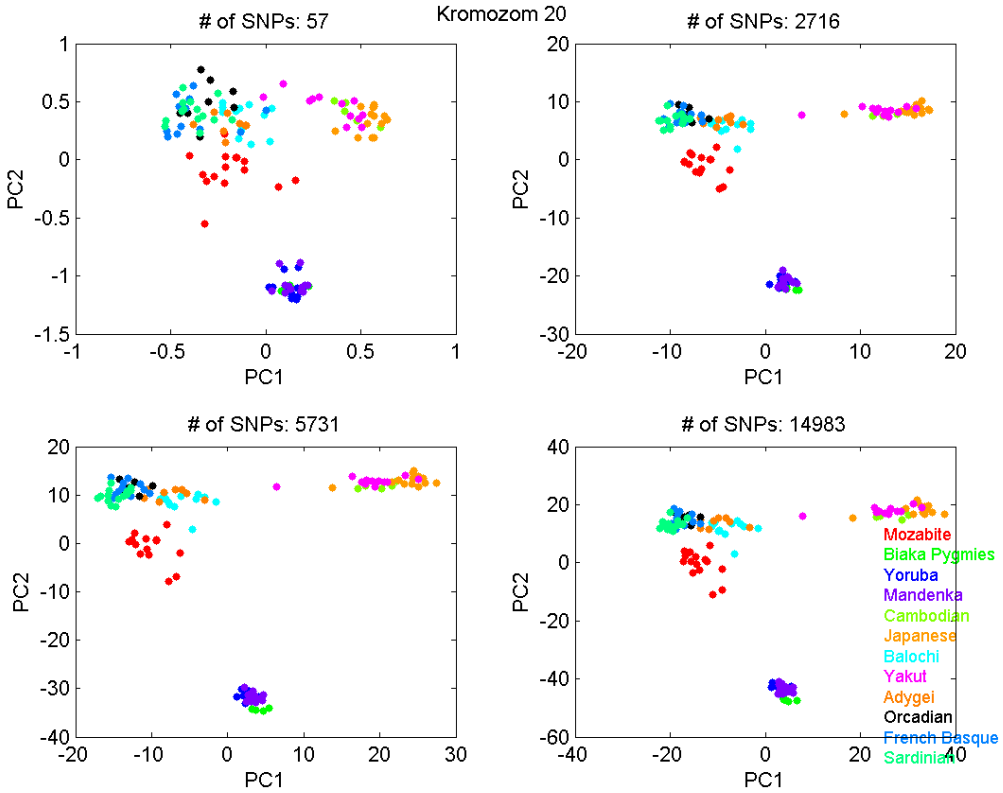
Şekil 4.53: PCA İle Genomik Dağılım (Koromozom 5)



Şekil 4.54: PCA İle Genomik Dağılım (Koromozom 10)

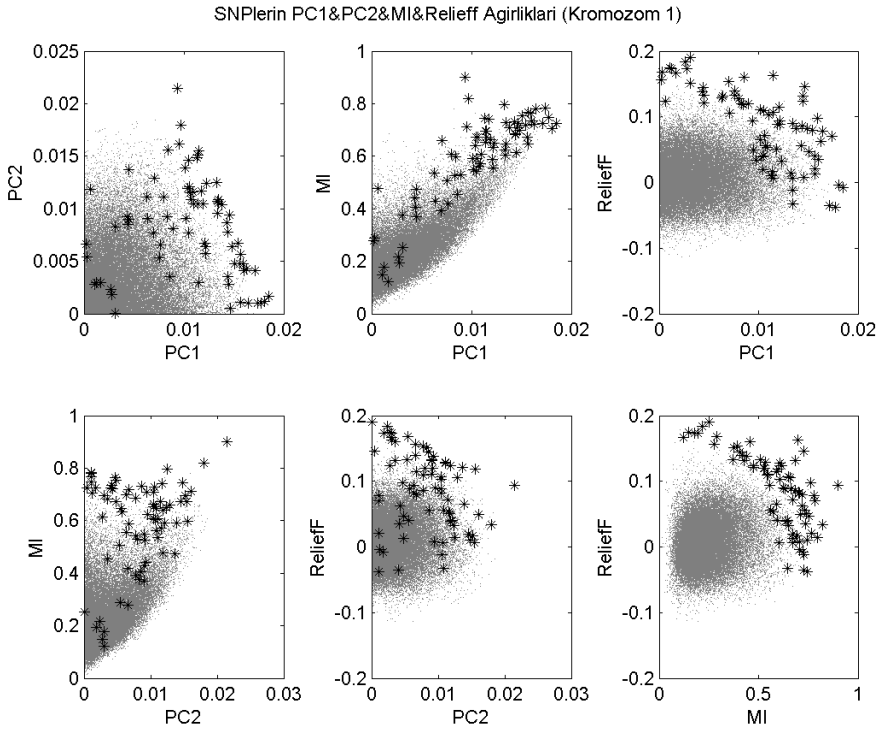


Şekil 4.55: PCA İle Genomik Dağılım (Koromozom 15)

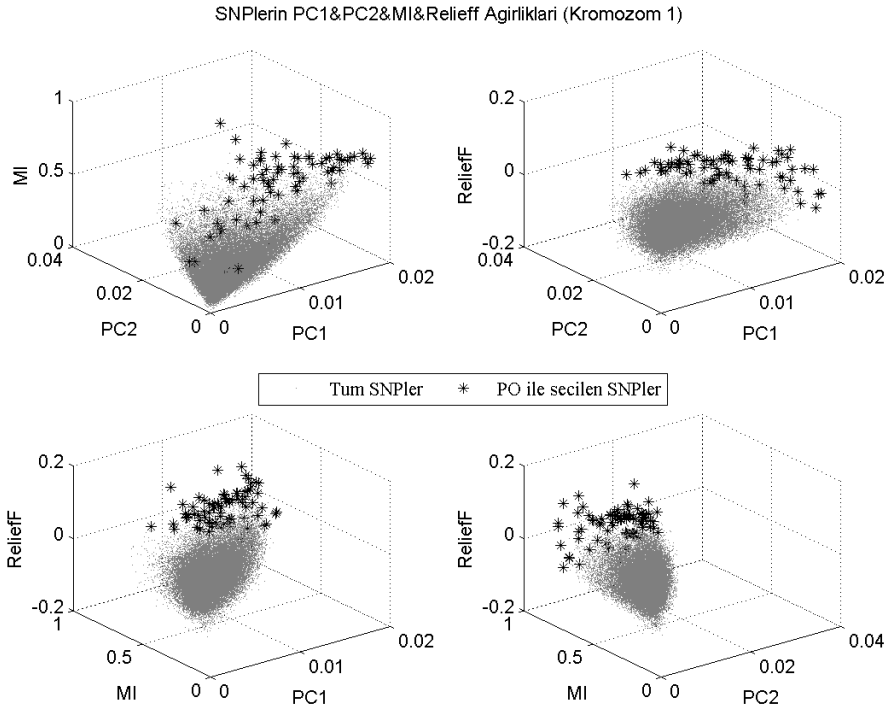


Şekil 4.56: PCA İle Genomik Dağılım (Kromozom 20)

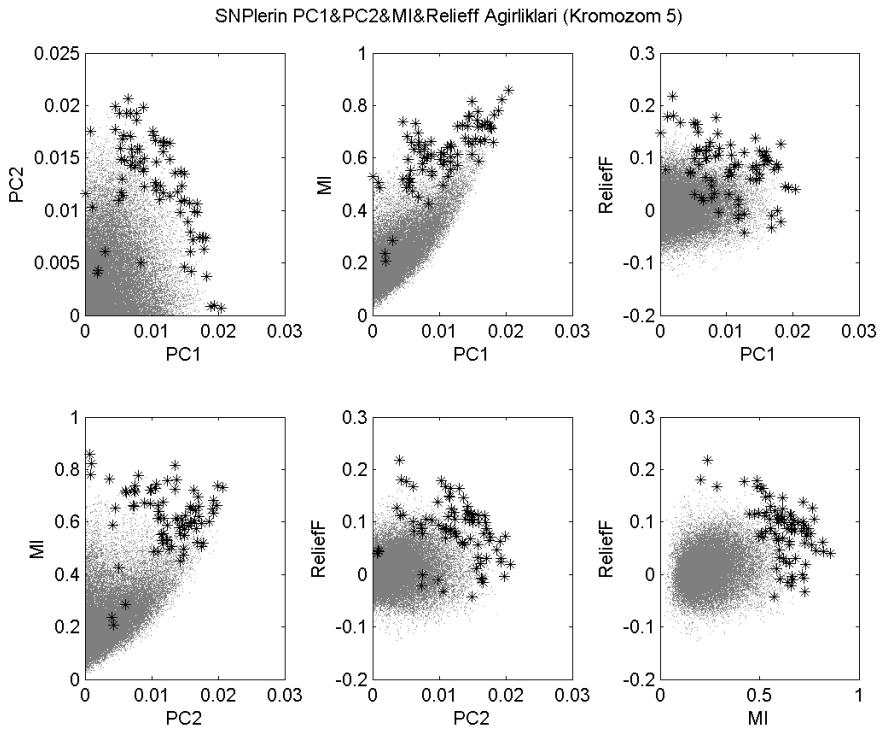
PO yönteminin çok kriterli SNP seçimineuyarlanması için, 2 farklı amaç gözetilerek 3 farklı metot (MI, Relieff ve PCA) ile oluşturulan 4 farklı kriter(MI, Relieff, PC1, PC2) kullanılmıştır. Tüm SNP'ler ve PO ile seçilen SNP'ler,kriterlerin farklı kombinasyonları ile 2 ve 3 boyutlu düzlemde **Şekil 4.57 - Şekil 4.66**'de kriterlerin tüm farklı kombinasyonları ile oluşturulan 2 ve 3 boyutlu düzlemlerde, 1, 5, 10, 15 ve 20. kromozomlar için tüm SNP'ler ve PO yöntemi kullanılarak seçilen SNP'ler aynı grafikte gösterilmiştir. Burada, tüm SNP'ler "." ile PO ile seçilen SNP'ler ise "*" ile işaretlenmiştir.**Şekil 4.57 - Şekil 4.66**'deki 2 ve 3 boyutlu görünümde seçilen SNP'lerin bazılarının bastırılamamış oldukları açıkça görülmektedir. Ancak seçim işlemi 4 boyutta yapıldığında dolayı2-3 boyutlu resimlerde eksik boyutlar olduğundan, seçilen bazı SNP'ler bastırılmış gibi görülmektedir. Sonuç olarak 2 ve 3 boyutlu gösterimler değerlendirilirken diğer boyutların varlığı unutulmamalıdır.



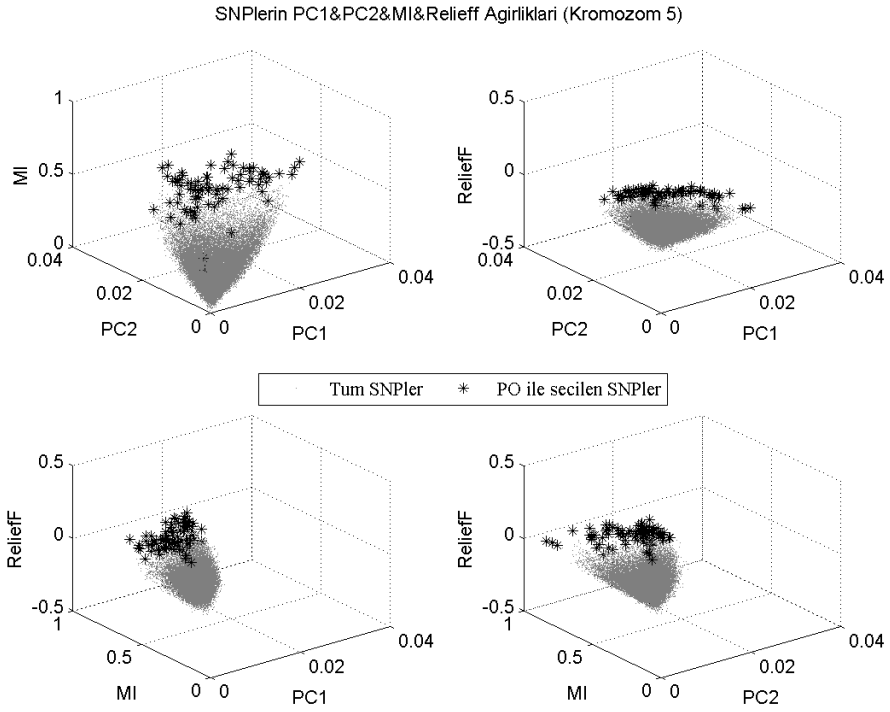
Şekil 4.57:PO İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr. 1)



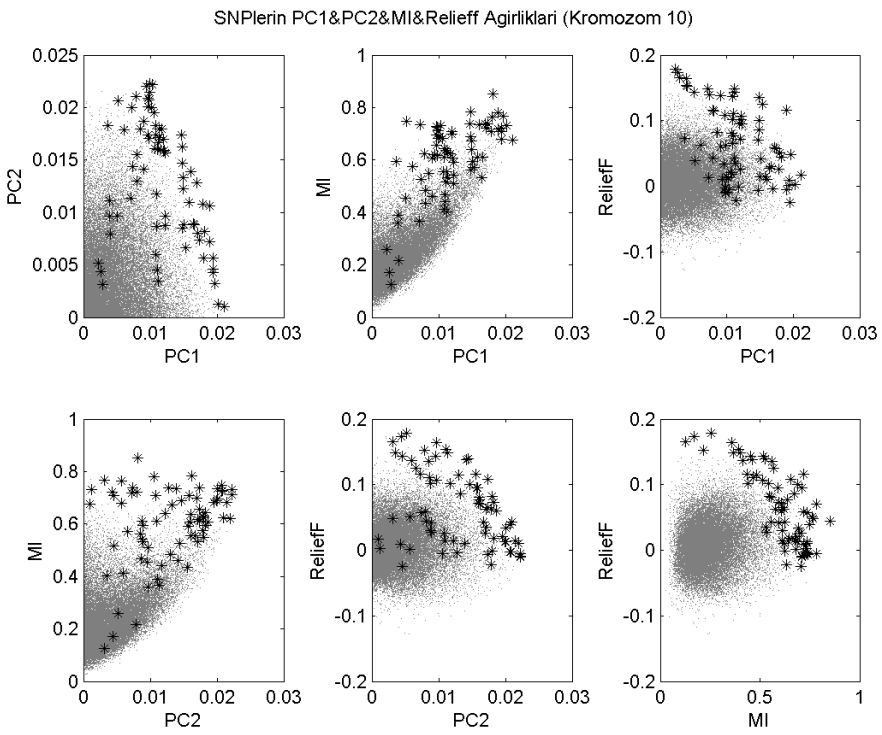
Şekil 4.58:PO İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr. 1)



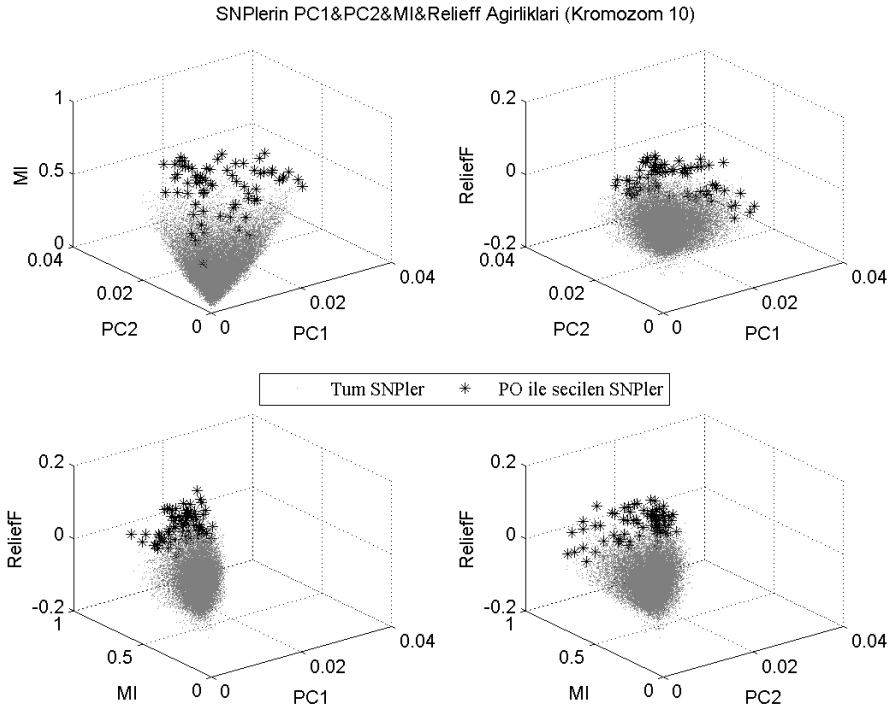
Şekil 4.59: PO İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr. 5)



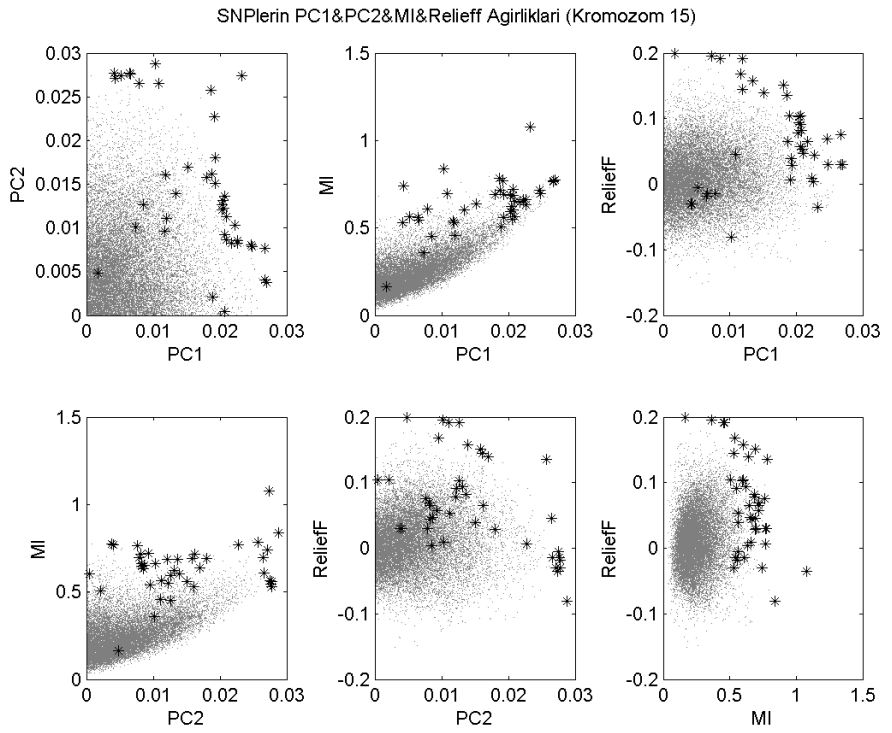
Şekil 4.60: PO İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr. 5)



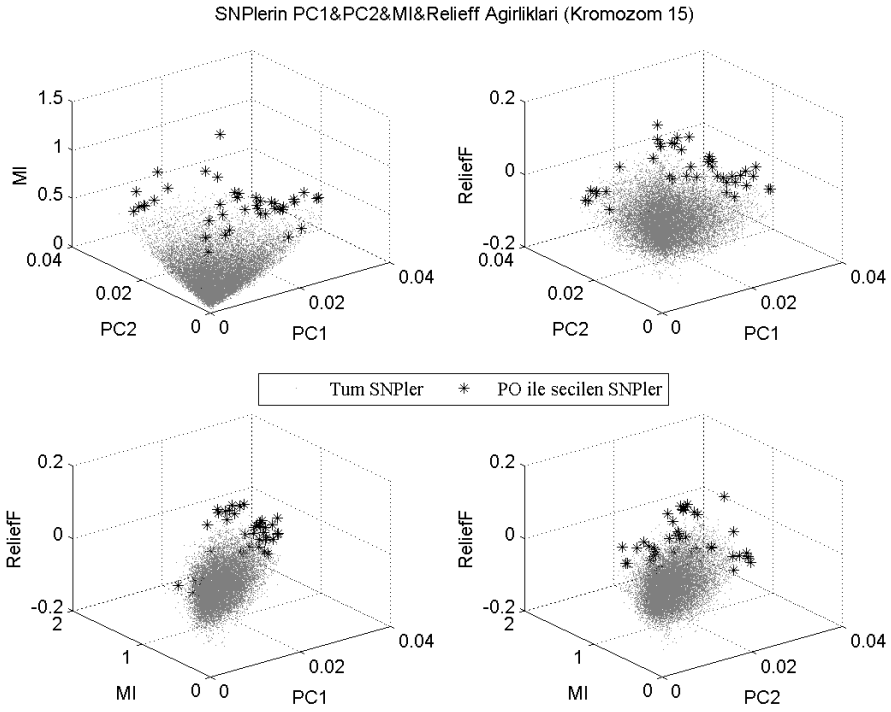
Şekil 4.61: PO İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr. 10)



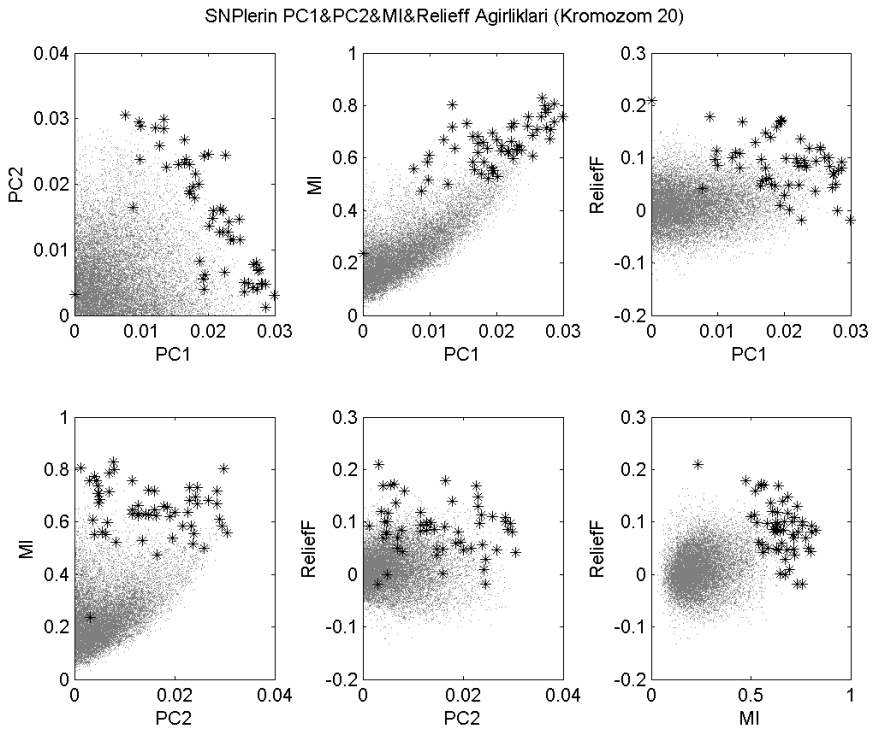
Şekil 4.62: PO İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr 10)



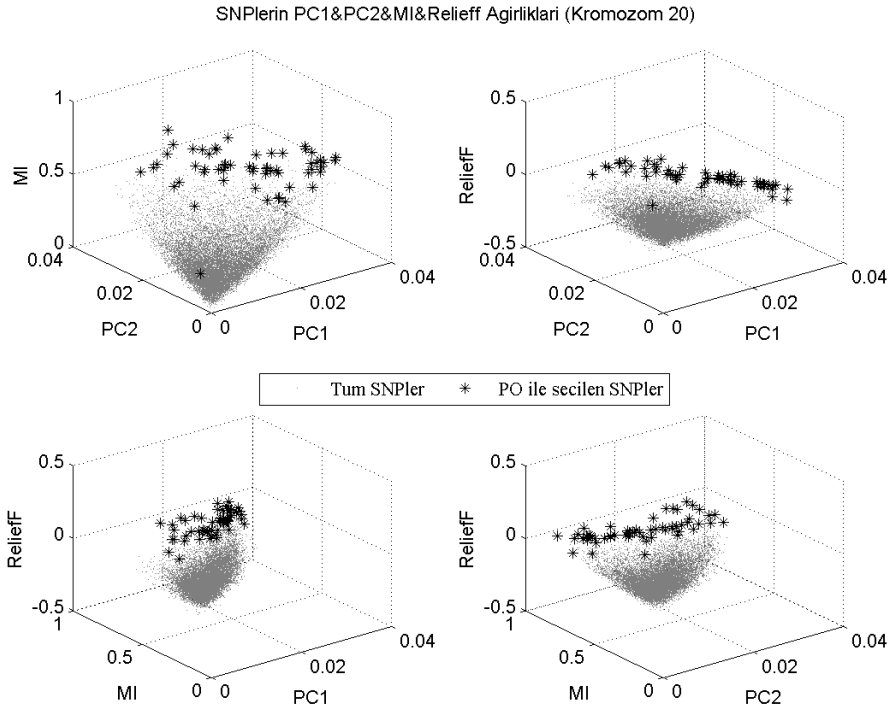
Şekil 4.63: PO İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr 15)



Şekil 4.64: PO İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr 15)



Şekil 4.65: PO İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr 20)



Şekil 4.66: PO İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr 20)

PO tarafından seçilen N adet SNP kullanılarak test kümesi üzerinde sınıflandırma doğruluğu (Amaç 1) ve jeo-genomik korelasyon (Amaç 2) değerleri hesaplanmıştır. Bu uygulamada, sınıflandırıcı olarak k-en yakın komşu (k=3) algoritması kullanılmıştır. Ayrıca jeo-genomik korelasyonun hesaplanması için, her grup çiftine ait ortalama genomik mesafeler (öklit mesafeleri) ve coğrafî mesafeler kullanılmıştır.

Sonuçlar **Tablo 4.15**'de sunulmuştur. Tablodaki ilk kolon (Kr.) koromozom bilgisini, 2 kolon (#S) ilgili kromozom için PO ile seçilen SNP sayısını,3 ve 5. kolonlar sırasıyla, PO-(PC1&PC2) kriteri ile PO tarafından seçilen SNP'leri kullanarak elde edilen sınıflandırma doğruluklarını ve test grupları arasındaki genomik ve coğrafî mesafeler arasındaki korelasyonu; 4. ve 6. kolonlar ise sırasıyla, MI kriteri ile elde edilen sınıflandırma doğruluğunu ve her kromozomda PC yükleri yerine MI skorlarına göre aynı miktarda SNP seçimi sonucu elde edilen korelasyonu göstermektedir. Tabloda en iyi sonuçlar altı çizili olarak gösterilmiştir.

Beklendiği gibi, MI tarafından seçilen SNP altkümelerinin PCA kullanılarak seçilen SNP altkümelerine göre sınıflandırma başarısının (amaç-1) daha yüksek olduğu görülmektedir. Aynı SNP altkümeleri, jeo-genomik korelasyonu için (amaç-2), ortalama

korelasyon değerleri göz önüne alındığında, öngörüldüğü gibi PO&PC kriteri MI kriterine göre (aynı miktarda SNP kullanılarak)daha iyi performans sağlamıştır. Bu sonuçlar herbir amaç için doğru öznitelik seçim yönteminin kullanıldığını göstermektedir. Sonuç olarak, **Tablo 4.15**'de görüldüğü gibi, PO&PC'nin kendisi, yalnız MI ile karşılaştırıldığında daha iyi korelasyon, MI da PO&PC ye göre daha iyi doğruluk değeri veren SNP'ler seçmektedir.

Tablo 4.15:2 Kriterli Seçim Doğruluk ve Korelasyon (%)

		Doğruluk (%)		**Korelasyon (%)**	
Kr	**#S**	**PO-PC1&PC2**	**MI**	**PO-PC1&PC2**	**MI**
1	13	41	45	51	45
2	16	36	45	42	55
3	15	36	44	44	39
4	19	43	31	46	35
5	17	37	43	65	64
6	14	31	43	78	65
7	21	36	41	47	35
8	16	39	37	62	49
9	8	29	40	36	30
10	18	33	51	53	59
11	14	33	41	53	46
12	12	41	39	47	43
13	14	32	39	31	40
14	14	33	29	42	70
15	7	24	45	51	46
16	16	35	44	53	62
17	10	36	33	53	39
18	17	32	29	51	47
19	15	36	42	60	35
20	14	35	31	35	37
21	14	33	32	55	37
22	13	34	43	62	46
23	17	27	36	39	35
Ort	**15**	**34**	**39**	**50**	**46**

Bununla birlikte, hem korelasyonu hemde doğruluğu da iyileştirebilmek için**Tablo 4.16**'de görüldüğü üzere Pareto Optimal ile 3 kriterli seçim yapılmıştır.Böylece Amaç 1 ve Amaç 2'nin eşzamanlı olarak optimize edilmesi düşünülmüştür. Bu iki amacı göz önünde bulundurarak seçim yapması amacı ile önerilen çok kriterli öznitelik seçim yönteminin**sadece** MI kullanımına göre, etnik grupları daha iyi ayıran SNP

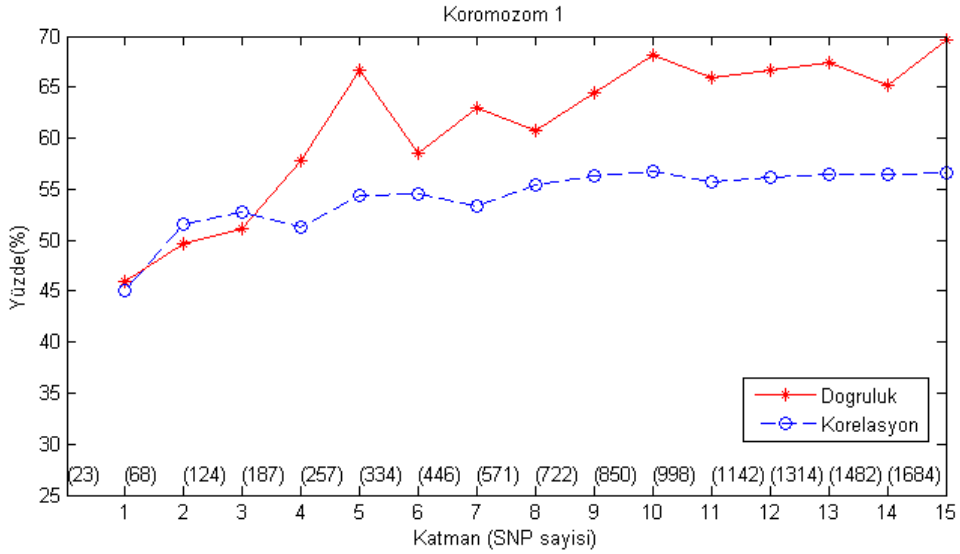
altkümelerini ortaya çıkardığı, bu arada sadece PC1&PC2 kullanıma göre de daha yüksek korelasyon değerleri de elde ettiği gösterilmiştir.

Tablo 4.16:Çoklu Amaç İle Doğruluk ve Korelasyon (%)

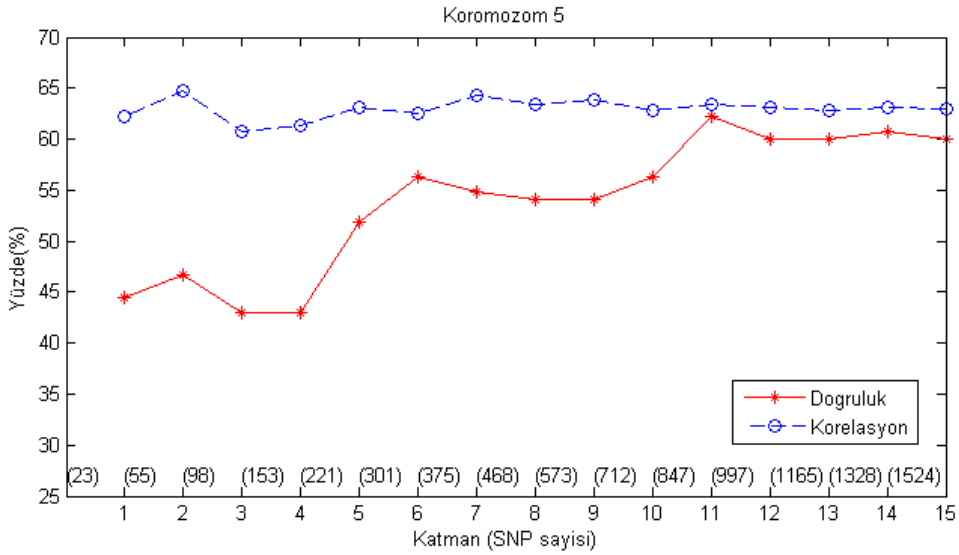
		Doğruluk (%)		Korelasyon (%)	
Kr	**#S**	**PO-PC&MI**	**MI**	**PO-PC&MI**	**PO-PC1&PC2**
1	23	45.93	45.19	45	51
2	19	39.26	38.52	52	42
3	24	45.19	40.74	50	44
4	28	43.70	40.74	51	46
5	23	44.44	42.96	62	65
6	17	37.78	40.00	77	78
7	27	42.22	36.30	49	47
8	31	38.52	38.52	64	62
9	11	30.37	33.33	37	36
10	24	41.48	52.59	63	53
11	21	33.33	37.04	49	53
12	24	33.33	38.52	52	47
13	18	37.04	36.30	34	31
14	18	37.04	31.11	49	42
15	7	23.70	33.33	51	51
16	26	45.19	46.67	56	53
17	13	34.07	29.63	56	53
18	23	39.26	28.89	56	51
19	16	35.56	39.26	64	60
20	16	40.74	28.89	36	35
21	17	34.07	28.15	50	55
22	22	42.96	40.00	69	62
23	24	39.26	37.04	39	39
Ort	**21**	**38.45**	**37.55**	**53**	**50**

Bu uygulamada incelenen 12 sınıflı bir problem için, kromozom başınaortalama 15-21 SNP seçimi ile yüksek başarı beklemek mümkün değildir. Paschou ve arkadaşlarının çalışmasında[48] yalnızca 2 milleti ayırmak (2 sınıflı problem) için yaklaşık 100 SNP kullanıldığı ifade edilmiştir. Diğer yandan, çalışmamızda temel aldığımız PO yönteminin doğası gereği seçilecek SNP sayısı belirlenememektedir. Bu problemin çözümü için, bu çalışmada çok katmanlı PO yönteminin kullanılması önerilmiştir. Bu amaçla1. seviyede Pareto Optimal kümeye dâhil edilmeyen SNP'ler çözüm kümesinden çıkarılarak, geriye kalan SNP'ler yeniden yarıştırılmış ve 2. seviye SNP'ler elde edilmiştir. Bu işlem 15 katmana kadar devam ettirilmiştir. Çok katmanlı PO sonuçları sırasıyla 1, 5, 10, 15 ve 20. kromozomlar ve tüm kromozomların ortalaması için **Şekil 4.67** - **Şekil 4.72**'da sunulmuştur. Sonuçlar analiz edildiğinde, sınıflandırma başarısının

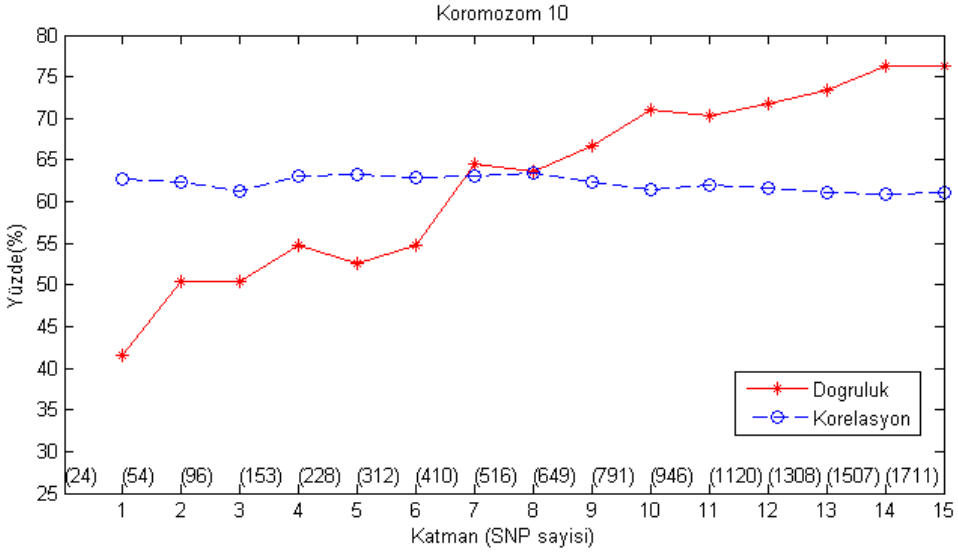
SNP sayısı ile orantılı olarak sürekli bir artış göstermesine rağmen jeo-genomik korelasyonun belli bir noktadan sonra sabit kaldığı görülmektedir.



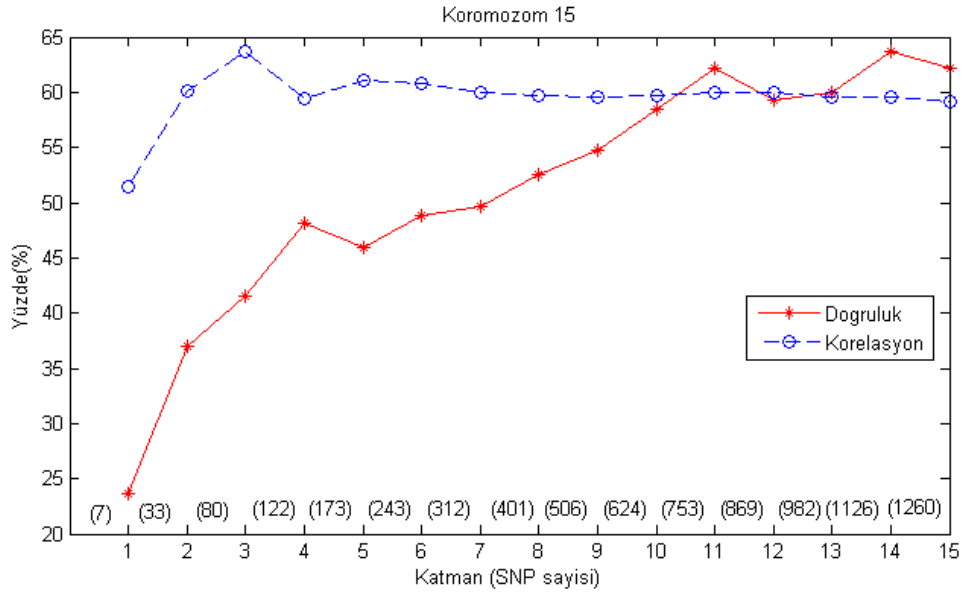
Şekil 4.67: Çok Katmanlı Po Sonuçları (Kromozom 1)



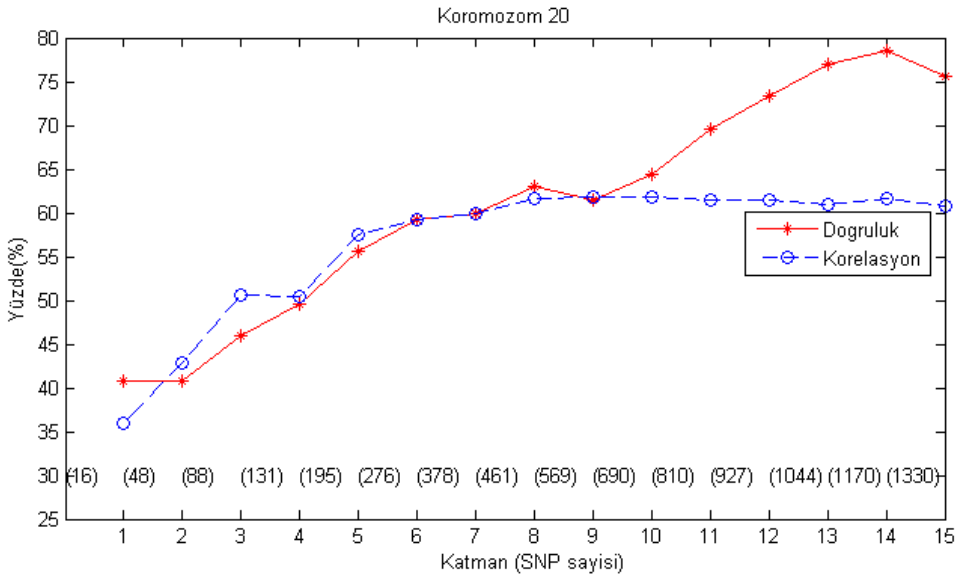
Şekil 4.68: Çok Katmanlı Po Sonuçları (Kromozom 5)



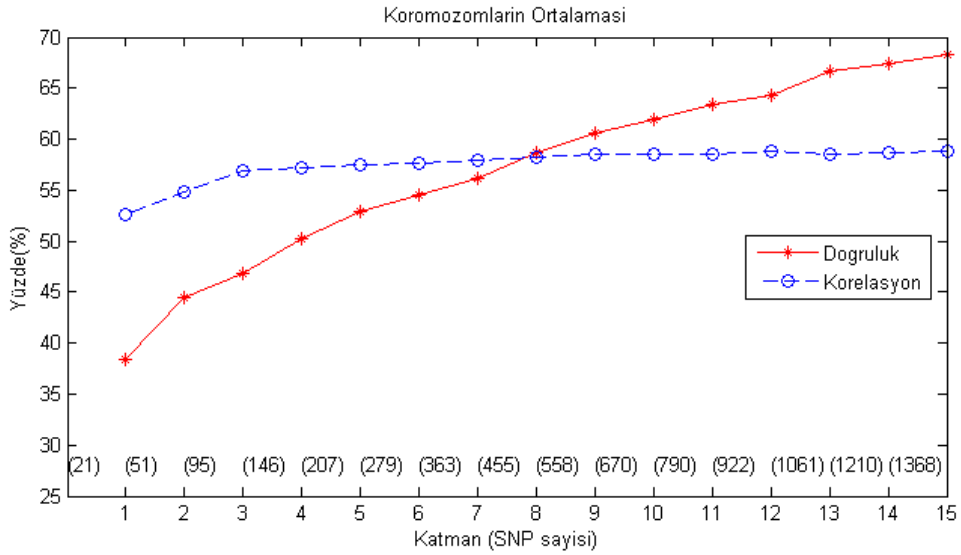
Şekil 4.69: Çok Katmanlı Po Sonuçları (Kromozom 10)



Şekil 4.70: Çok Katmanlı Po Sonuçları (Kromozom 15)



Şekil 4.71: Çok Katmanlı Po Sonuçları (Kromozom 20)



Şekil 4.72: Çok Katmanlı PO Sonuçları (Tüm Kromozomlar İçin Ortalama)

Çok amaçlı SNP seçimi için, sınıflandırma başarısını arttırmak amacı ile yeni bir kriter (Relief-F skoru) daha eklenerek toplam kriter sayısı 4'e çıkartılmıştır. Bunun yanısıraçok kriterli seçim için, Condercet, MC4 ve AHP metotları da kullanılarak, PO yöntemi ile birlikte toplam 4 birleştirme metotu uygulanmıştır. Böylece, yöntemlerin karşılaştırılması da sağlanmıştır.

Verinin analizi için yine ilk olarak, her grubun yarısı eğitim kümesi; geri kalan yarısı test kümesi olarak ayrılmıştır. SNP'lerin MI değeri, Relief-F skoru, PC1 ve PC2

ağırlıkları eğitim kümesi kullanılarak hesaplanmıştır. Böylelikle 4 farklı kritere (MI, Relieff, PC1, PC2) göre SNP'ler 4 ayrı birleştirme yöntemi kullanılarak yeniden analiz edilmiştir PO doğası gereği, adayları birbiri ile yarıştırmakta ve yenilemeyen tüm adayları derecelendirme yapmaksızınsadece seçmektedir. Condercet, MC4 ve AHP yöntemleri ise mevcut kriterleri kullanarak yeni bir konsensüs skor oluşturarak SNP'leri sıralamaktadır. Bu yöntemlerin sıralama sonuçlarına göre, en yüksek skora sahip N adet (N; PO tarafından seçilen SNP sayısı) SNP seçilmiştir.

Yine sınıflandırma doğruluğu (Amaç 1) ve jeo- genomik korelasyon (Amaç 2), birleştiriciler tarafından seçilen N adet SNP kullanılarak test kümesi üzerinde hesaplanmıştır. Sınıflandırma başarının hesaplanması için k-en yakın komşu (k=3) algoritması kullanılmıştır. Gruplar arası ortalama genomik mesafeler (öklit mesafeleri) ve coğrafî mesafeler jeo-genomik korelasyonun hesaplanmasında kullanılmıştır. Üç birleştirici yöntem tarafından seçilen SNP'ler üzerinde hesaplanan doğruluk ve korelasyon değerleri**Tablo 4.17**'da sunulmuştur.

İlk olarak her iki amaca hizmet edecek SNP'lerin ayrı olarak seçim işlemi yapılmış daha sonra iki amacı aynı anda sağlayacak SNP'ler seçilmiştir. **Tablo 4.17**' de sırasıyla seçilen SNP'lerin sınıflandırma başarıları ve jeo-genomik korelasyon değerleri gösterilmiştir. İlk kolon koromozom bilgisini içermektedir. #S ilgili koromozom için PO ile seçilen SNP sayısını göstermektedir. Tabloda en iyi sonuçların altı çizii gösterilmiştir.

2. bölümde her iki amacı gerçeklemek amacı ile dört kriteri de göz önünde bulunduran PO sonuçları listelenmiştir. 3. bölüm sadece sınıflandırma başarısını yüksek kılmayı hedefleyerek MI ve Relieff değerlerini kullanan PO sonuçlarını göstermektedir. 4. bölüm jeo-genomik korelasyonu yüksek kılmayı hedefleyerek PC1 ve PC2 ağırlıklarını kullanan PO sonuçlarını göstermektedir. Sonuçlar 2 amacı birlikte hedefleyen SNP'lerin iyi bir başarı elde ettiğini göstermektedir. Birçok durumda her iki amaçta da artış sağlanmıştır. Ancak 1. amacın 2 amaca negatif etki ettiği de görülmektedir. Sınıflandırma başarısı sonuçlarına bakıldığında, heriki amacı gözeten sonuçların sadece 1. amacı gözeten sonuçlardan herzaman daha başarılı olduğu görülmektedir. Burada ilginç olan,bazı durumlarda 2. amacı hedefleyen seçimlerin en yüksek başarıyı vermesidir. Sonuçlar temel bileşen analizinin sadece jeo-genomik mesafe

korelasyonunu sağlamakla kalmayıp, yüksek sınıflandırma başarısı verdiğinide göstermektedir. Ayrıca bu sonuçlar temel bileşen analizinin sadece var olan boyutlardan yeni boyutlar çıkarmak için değil, boyutlar arasında seçim yapmak içinde kullanılabileceğini göstermiştir.

Tablo 4.17: 4 Kriterli Seçim Doğruluk Ve Korelasyon (%)

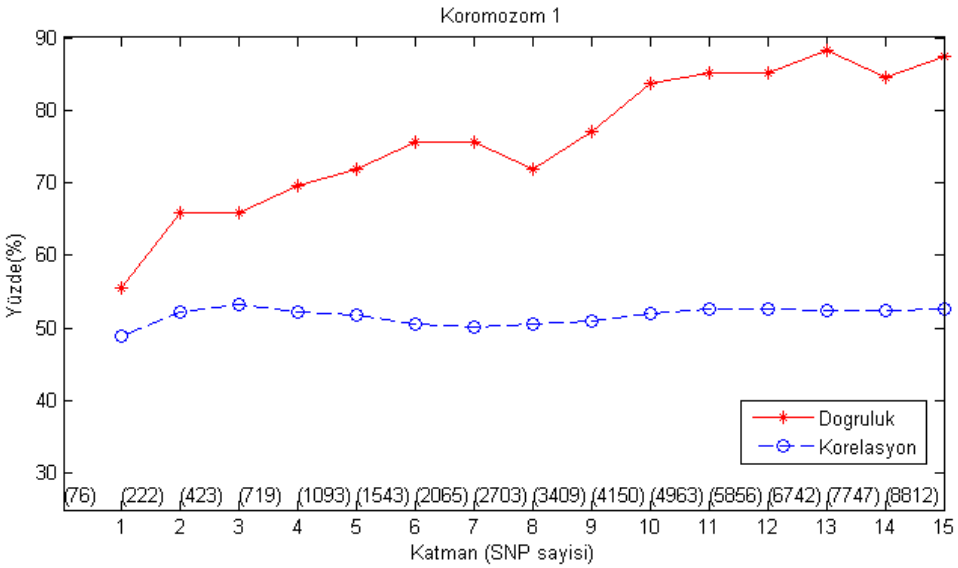
	PO_amac1&amac2			**PO_amac1 (MI&Relieff)**			**PO_amac2 (PC1&PC2)**		
Kr	**#S**	**Doğruluk**	**Korelasyon**	**#S**	**Doğruluk**	**Korelasyon**	**#S**	**Doğruluk**	**Korelasyon**
1	76	56	49	6	41	45	13	51	45
2	87	57	56	18	36	45	16	42	55
3	103	53	50	7	36	44	15	44	39
4	66	50	54	8	43	31	19	46	35
5	82	49	54	12	37	43	17	65	64
6	53	38	66	8	31	43	14	78	65
7	106	63	49	8	36	41	21	47	35
8	56	46	55	8	39	37	16	62	49
9	81	49	46	11	29	40	8	36	30
10	76	47	63	11	33	51	18	53	59
11	81	57	52	7	33	41	14	53	46
12	67	49	49	5	41	39	12	47	43
13	46	48	41	3	32	39	14	31	40
14	54	41	46	6	33	29	14	42	70
15	42	40	55	8	24	45	7	51	46
16	45	44	54	3	35	44	16	53	62
17	36	42	51	5	36	33	10	53	39
18	51	50	52	6	32	29	17	51	47
19	35	34	56	6	36	42	15	60	35
20	57	49	48	10	35	31	14	35	37
21	42	40	49	6	33	32	14	55	37
22	44	44	61	7	34	43	13	62	46
23	79	41	51	6	27	36	17	39	35
Ort	**64**	**47**	**52**	**8**	**34**	**39**	**15**	**50**	**46**

Sınıflandırma doğruluğu ve coğrafî-genomik korelasyonu arttırmak amacı ile öznitelik kümesine daha çok SNP eklenmiştir. Bunun için çok katmanlı Pareto Optimal kullanılmıştır.**Tablo 4.18**'da tüm kromozomlar için, PO elde edilen ortalama doğruluk ve korelasyon değerleri sunulmuştur.Burada, Pareto Optimal'in ilk 15 seviyesi kullanılmıştır.

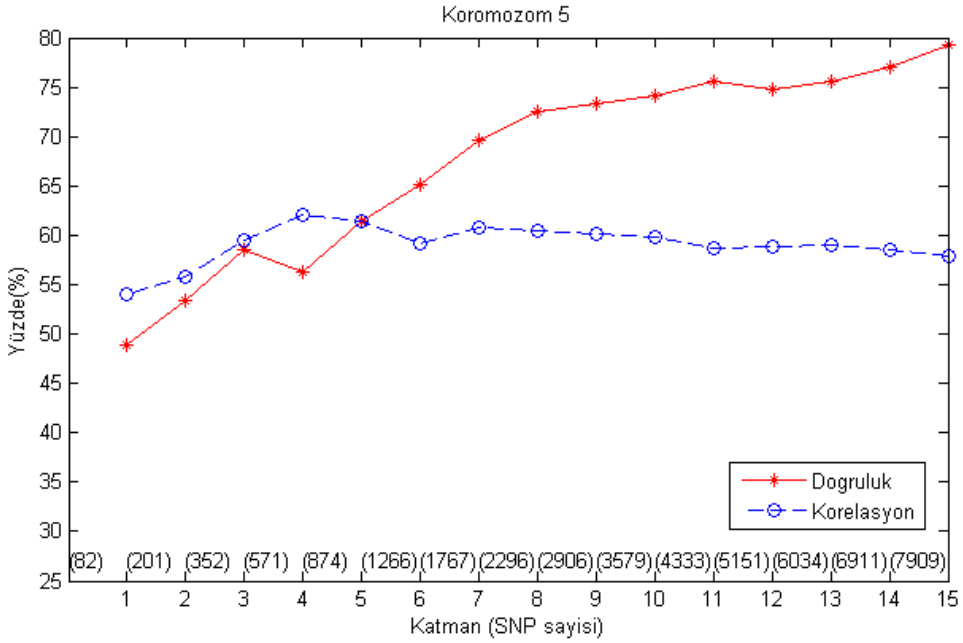
Tablo 4.18: Artan Seviye (Snp Sayısı) İçin Ortalama Doğruluk/Korelasyon Değerleri (%)

Katman	#S	Doğruluk	Korelasyon
1	64	47	52
2	176	54	54
3	346	59	54
4	565	63	54
5	833	66	54
6	1157	69	54
7	1549	71	54
8	1995	73	54
9	2476	75	54
10	3014	77	55
11	3598	78	55
12	4238	80	55
13	4923	81	55
14	5634	81	55
15	6382	82	55

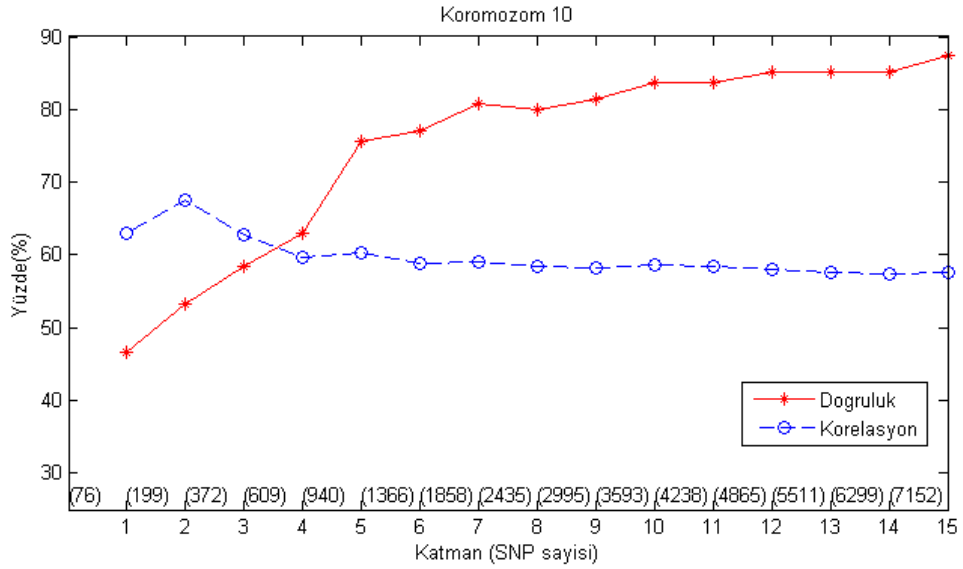
Tablo 4.18'de görüldüğü üzere, seviye sayısı arttıkça, sınıflandırmanın ortalama doğruluğu da artmaktadır. Bununla birlikte, bu artış, ortalama korelasyon değerleri için sınırlıdır.**Şekil 4.73-Şekil 4.78** pareto seviyenin artışı ile sırasıyla sınıflandırma doğruluğunun ve coğrafî-genomik korelasyonun artışını göstermektedir.



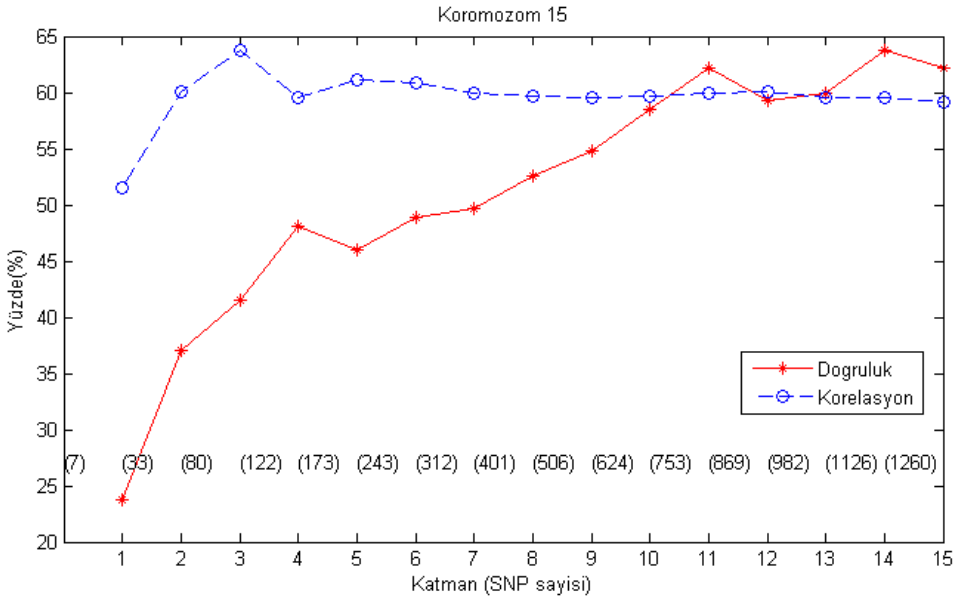
Şekil 4.73: Pareto Seviyesine Karşılık Doğruluk Ve Korelasyonun Artışı (Kr. 1)



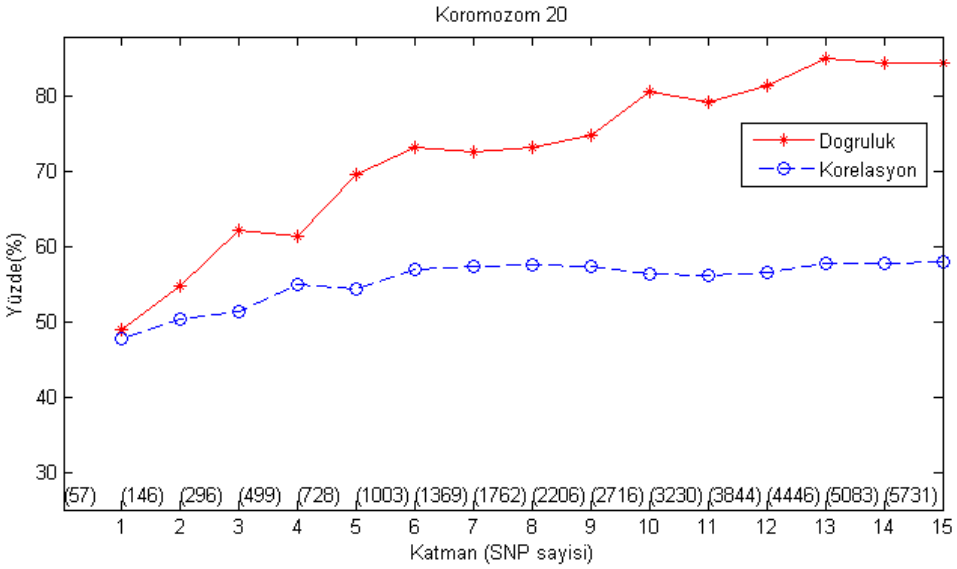
Şekil 4.74: Pareto Seviyesine Karşılık Jeo Doğruluk Ve Korelasyonun Artışı (Kr. 5)



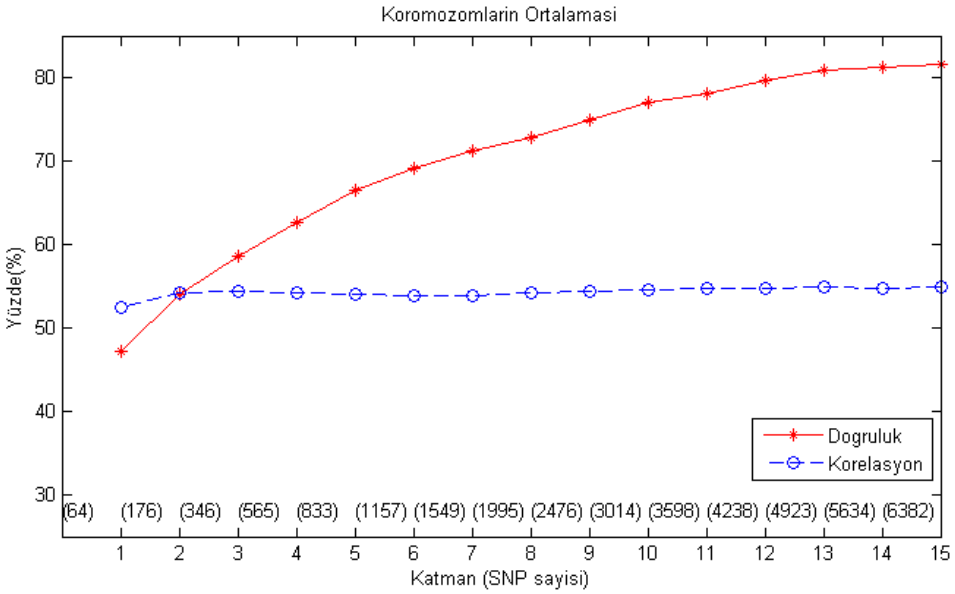
Şekil 4.75: Pareto Seviyesine Karşılık Doğruluk Ve Korelasyonun Artışı (Kr. 10)



Şekil 4.76: Pareto Seviyesine Karşılık Jeo Doğruluk Ve Korelasyonun Artışı (Kr. 15)



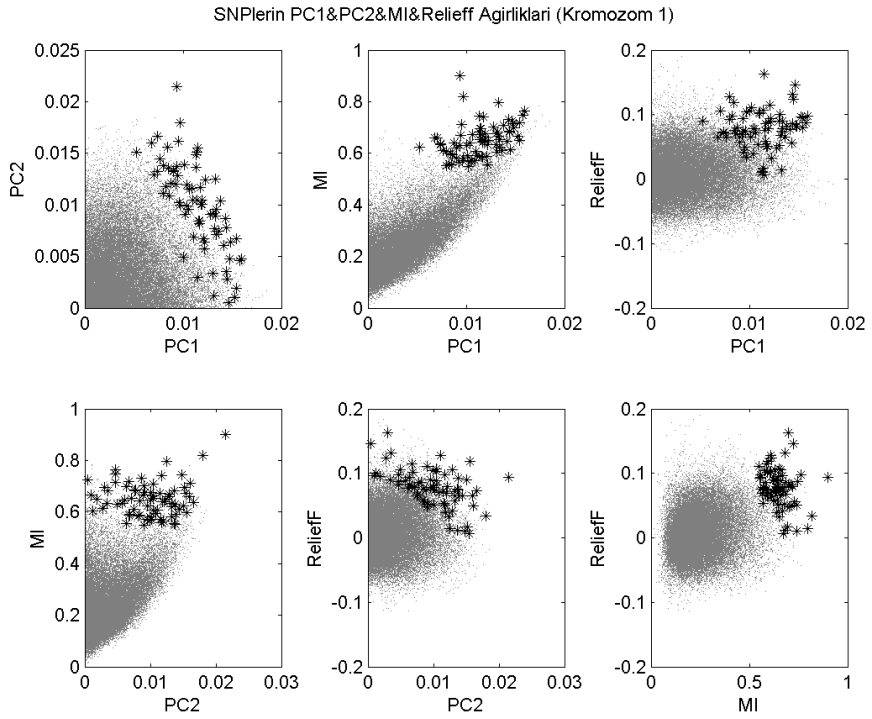
Şekil 4.77: Pareto Seviyesine Karşılık Doğruluk Ve Korelasyonun Artışı (Kr. 20)



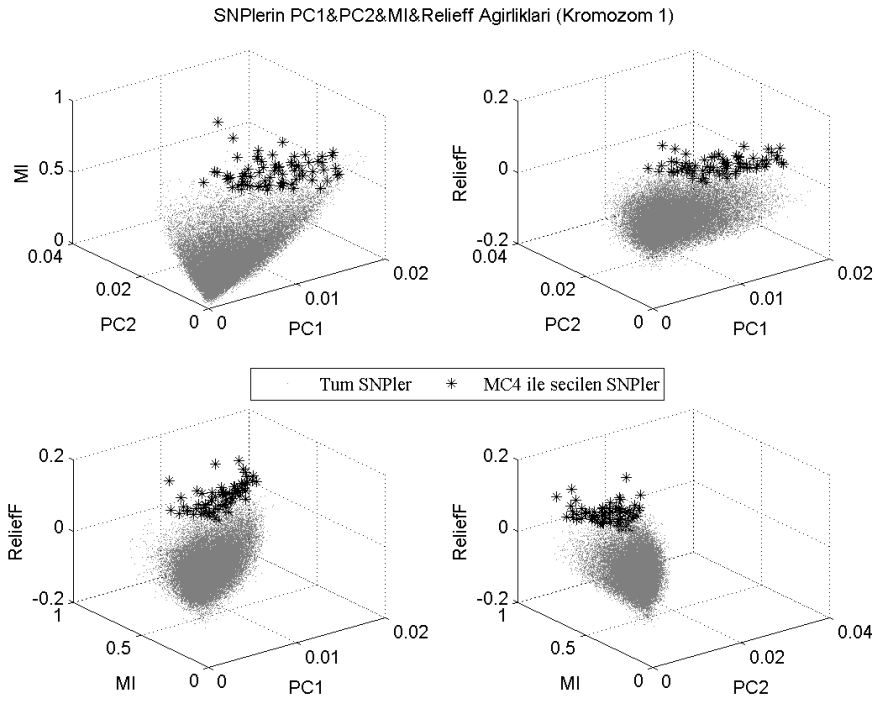
Şekil 4.78: Pareto Seviyesine Karşılık Doğruluk Ve Korelasyonun Artışı (Ortalama)

HGDP verisi ile yaptığımız son çalışmada farklı çok kriterli öznitelik seçim yöntemlerinin karşılaştırması yapılmıştır. Bu amaçla PO yöntemine ek olarak Condorcet, MC4 ve AHP yöntemleri ile SNP'lerin 4 skoru (MI, Relieff, PC1, PC2) kullanılarak SNP seçim işlemi yapılmıştır. Yine 2 amacın bir arada optimize edilmesi hedeflenmiştir.

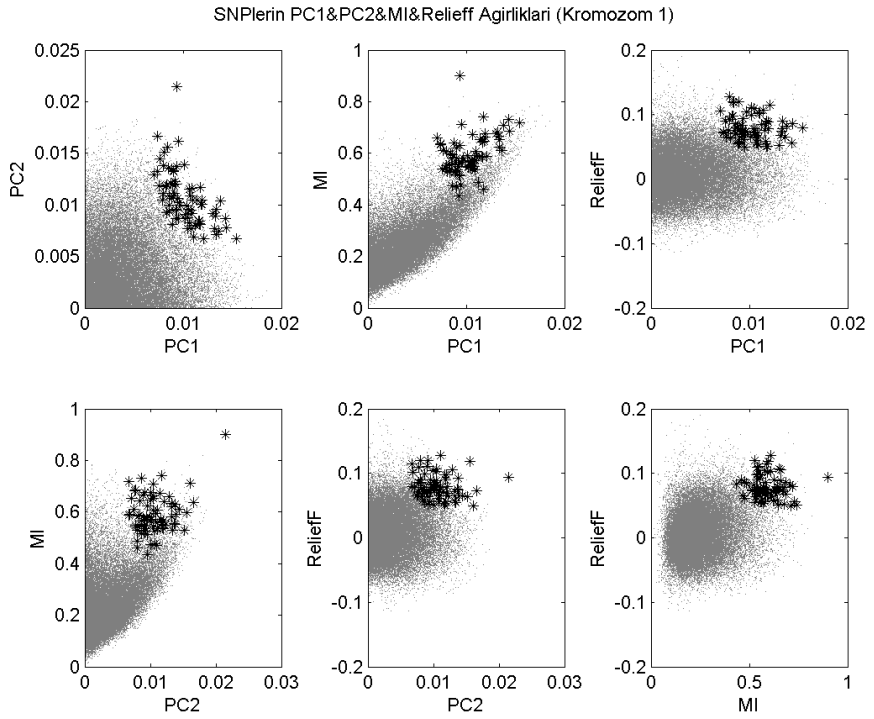
İlk olarak tesmil olması açısından 3 yöntemle kromozom 1 için seçilen SNP'lerinkriterlerin farklı kombinasyonları ile 2 ve 3 boyutlu düzlemdeki görünümleri elde edilmiştir **(Şekil 4.79-Şekil 4.84)**. Seçim işlemi 4 boyutta yapılmıştır. Dolayısı ile 2 ve 3 boyutlu resimlerde eksik boyutlar olduğundan, 2 ve 3 boyutlu gösterimler değerlendirilirken diğer boyutların varlığı unutulmamalıdır.



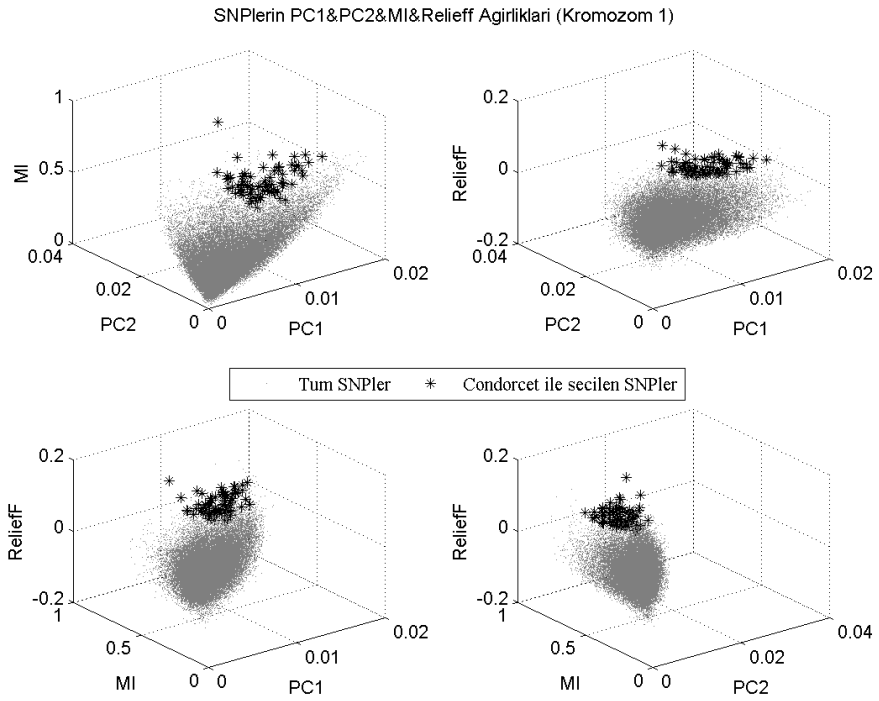
Şekil 4.79: MC4 İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr. 1)



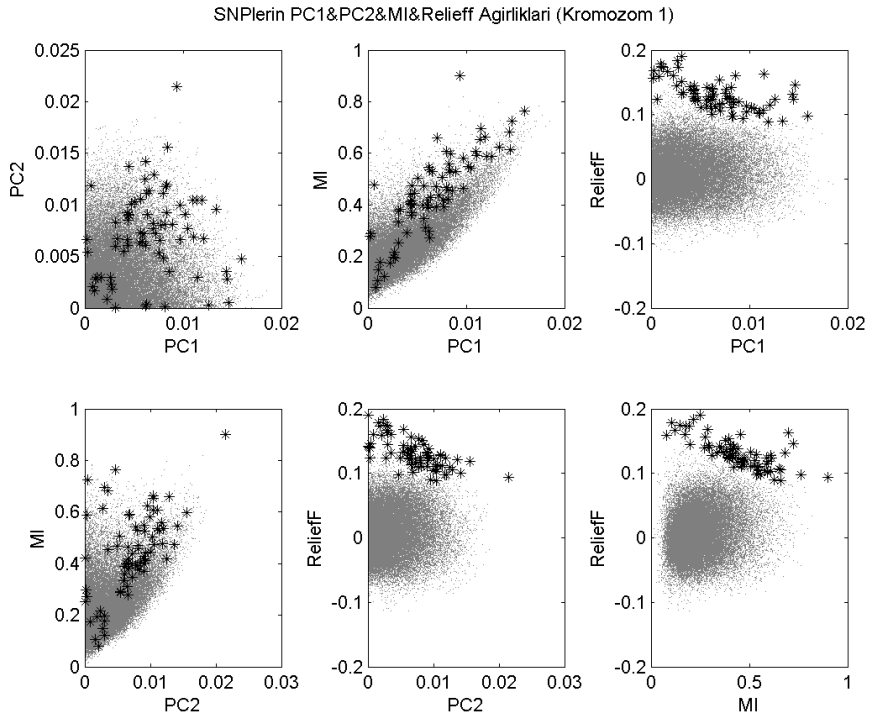
Şekil 4.80: MC4 İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr 1)



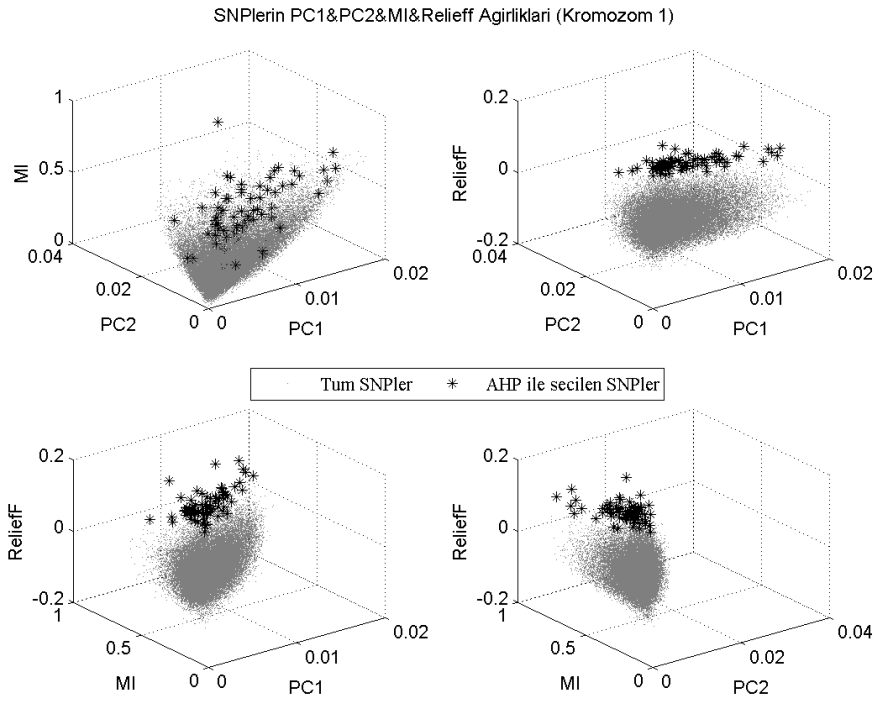
Şekil 4.81: Condorcet İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr. 1)



Şekil 4.82: Condorcet İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr. 1)



Şekil 4.83: AHP İle Seçilen SNP'lerin 2 Boyutlu Gösterimi (Kr 1)



Şekil 4.84:AHP İle Seçilen SNP'lerin 3 Boyutlu Gösterimi (Kr 1)

PO, Condorcet, MC4 ve AHP yöntemlerince seçilen SNP'lerin 1. seviye PO de seçilen öznitelik sayısı kadar SNP ile doğruluk ve korelasyon değerleri **Tablo 4.19**'de sunulmuştur. Sonuçlar doğruluk açısından metotların birbirne yakın olduğunu göstermektedir. Yine de bir sıralama yapılacak olursa MC4 9, PO 8, Condorcet 4, AHP 3 kromozom için en iyi sonucu vermiştir. Ancak korelasyon açısından bakıldığında en iyi sonucun Condorcet ile alındığı, AHP ile alınan sonuçların ise en kötü olduğu görülmektedir.

Tablo 4.19: 4 Farklı Seçim Metodu için 1. Seviye Doğruluk/Korelasyon Değerleri (%)

Seviye1		Doğruluk (%)				Korelasyon (%)			
Kr	**#S**	PO	Condorcet	MC4	AHP	PO	Condorcet	MC4	AHP
1	76	57	44	54	50	49	69	56	19
2	87	52	47	60	51	56	65	61	54
3	103	47	47	50	47	50	72	65	40
4	66	46	51	47	44	54	72	66	37
5	82	44	47	41	53	54	71	61	15
6	53	34	48	41	42	66	75	66	55
7	106	63	43	47	52	49	67	56	39
8	56	41	43	48	41	55	68	65	48
9	81	49	47	47	42	46	61	48	42
10	76	50	42	44	44	63	63	58	24
11	81	51	43	53	44	52	61	68	41
12	67	46	44	41	45	49	48	45	19
13	46	45	42	44	43	41	59	58	1
14	54	41	50	43	39	46	75	68	-5
15	42	37	46	52	44	55	41	40	-9
16	45	44	45	50	50	54	63	53	25
17	36	40	41	41	42	51	62	52	17
18	51	47	49	51	44	52	57	53	39
19	35	35	37	44	36	57	72	47	35
20	57	48	50	46	39	48	59	54	38
21	42	41	40	34	40	49	63	56	48
22	44	42	47	45	45	61	82	76	27
23	79	43	39	31	36	51	59	42	37
Ort	**64**	**45**	**45**	**46**	**44**	**52**	**65**	**57**	**30**

Aynı sonuçlar PO'nun 15. seviyesinde seçilen SNP sayısı kadar SNP ile tekrar hesaplanmıştır. Sonuçlar **Tablo 4.20**'da listelenmiştir. Burada, beklendiği gibi artan SNP sayısı ile sınıflandırma başarısının eş zamanlı olarak arttığı gözlemlenmiştir. Korelasyon açısından bakıldığında ise ters orantılı olarak düşüşün olduğu ve 3 metodun

%55'e yakınsadığı, AHP'nin ise yine en kötü sonuçları ortaya çıkardığı görülmektedir. Sınıflandırma başarısı açısından benzer metotlar benzer sonuçlar üretmiştir. En iyi sonucu vermeleri açısından bakıldığında MC4 9, PO 8, Condorcet 7, AHP 5 olarak sıralanırlar.

Tablo 4.20: 4 Farklı Seçim Metotu için 15. Seviye Doğruluk/Korelasyon Değerleri (%)

Seviye15		Doğruluk (%)				Korelasyon (%)			
Kr	**#S**	PO	Condorcet	MC4	AHP	PO	Condorcet	MC4	AHP
1	8812	87	88	93	81	53	55	55	40
2	9865	84	85	87	90	60	61	61	59
3	8283	79	81	81	79	58	60	59	57
4	8155	89	86	87	86	54	54	54	52
5	7909	79	79	81	71	58	57	58	42
6	7132	88	87	90	79	58	60	61	45
7	8365	82	81	81	90	55	55	56	55
8	6829	83	84	81	85	52	54	54	52
9	6719	86	84	85	87	57	55	56	56
10	7152	90	84	87	81	58	59	59	40
11	7907	86	90	88	87	57	57	58	57
12	6721	84	85	88	76	59	52	53	36
13	5668	87	84	83	75	55	56	57	40
14	4149	74	78	76	68	51	54	54	39
15	5102	79	81	81	76	53	53	54	35
16	6171	88	88	88	85	52	51	52	40
17	5128	84	83	82	81	52	56	56	50
18	6172	79	81	81	77	50	49	50	40
19	4073	88	86	87	79	50	55	55	44
20	5731	84	84	84	86	58	56	56	57
21	3792	79	77	76	74	56	54	54	45
22	4192	79	75	76	73	58	58	58	49
23	2750	58	61	61	59	49	45	45	44
Ort	**6382**	**82**	**82**	**83**	**79**	**55**	**55**	**55**	**47**

Herbir kromozomların 15 seviye için ortalama değerleri ise **Tablo 4.21** özetlenmiştir. Sınıflandırmadaki doğruluk değeri açısından bakıldığında metotların benzer sonuçlar ürettiği görülmektedir. AHP ortalama olarak başarısız görünse de 7, 11 ve 21 için en iyi başarıyı vermiştir. Korelasyon açısından bakıldığında Condercet en iyi, AHP en kötü başarıyı vermiştir.

Tablo 4.21: Tüm Seviyelerin Ortalama Doğruluk/Korelasyon Değerleri (%)

Ortalama		Doğruluk (%)				Korelasyon (%)			
Kr	**#S**	PO	Condorcet	MC4	AHP	PO	Condorcet	MC4	AHP
1	337	77	73	76	71	52	58	57	34
2	365	76	78	79	77	58	64	63	57
3	324	67	70	70	69	56	65	63	51
4	302	75	77	75	71	56	59	57	46
5	295	68	69	69	63	59	61	59	35
6	259	70	73	73	69	57	65	62	46
7	325	75	73	73	75	54	57	56	49
8	258	72	74	75	72	53	55	56	48
9	262	72	73	73	71	52	53	53	52
10	283	76	75	76	70	60	56	57	30
11	312	75	72	75	76	57	58	59	52
12	258	69	72	72	65	56	53	53	27
13	224	71	69	70	64	53	57	58	29
14	154	64	66	66	62	50	59	57	31
15	200	68	68	71	64	54	47	49	19
16	235	75	73	74	71	53	55	55	34
17	202	70	71	72	69	50	59	56	42
18	229	70	69	71	69	47	51	50	31
19	164	71	70	72	68	52	58	55	39
20	221	72	74	74	71	55	55	54	52
21	153	64	66	65	66	56	56	54	40
22	177	67	66	67	64	59	63	61	41
23	126	52	54	53	50	49	48	45	41
Ort	**246**	**70**	**71**	**71**	**68**	**54**	**57**	**56**	**40**

Dört farklı seçim metodunun sonuçlarına genel anlamda bakıldığında herbir metodun farklı veriler için başarılı olduğu açıkça görülmektedir. Bu sonuçlar başta belirtiğimiz öznitelik seçim işleminin metot bağımlılığını ortaya koymaktadır. Bu problemin tamamen ortadan kaldırılması mümkün değildir. Ancak farklı amaçlara hizmet eden metotların birleştirilmesini önerdiğimiz bu çalışma ile bu sorunun çözümüne bir katkı sağlanmıştır. Hernekadar PO yöntemi Condercet ve MC4 yöntemlerini tamamen geçemesede PO'nun artı özellikleri dikkate alındığında tercih sebebi olacaktır. Örneğin öznitelik seçiminde en büyük sorulardan biri en iyi-k için *k* değerinin belirlenmesidir. Oysa PO ile öznitelikler bir yarış içine girdiğinden en iyi-k yöntemce belirlenmektedir. Ayrıca çok farklı bir amaca hizmet eden bir yöntemden gelen kriterlerdiğer 2 yöntemde

etkisiz elemana dönüşürken PO bunu özellikle dikkate almaktadır. MC4 iyi bir başarı sergilemektedir. Bu tercih sebebi olabilir ancak MC4 özdeğer hesabı gerektirdiğinden işlem zamanı uzundur ve yüksek kapasiteli bilgisayar gerektirir. Burada AHP yönteminde tüm kriterlerden ortak sıralama çıkarmak için en basit normalizasyon yöntemi uygulanmıştır. Ancak literatürde özvektör hesabı gibi daha kapsamlı hesaplama yöntemleri de önerilmiştir. Böyle metotlar uygulandığında AHP'nin de başarısı artabilir. Buna karşılık, bu tarz bir hesaplamada AHP yöntemi de yüksek boyutlu öznitelikleri hesaplamak için fazla bellek harcadığından uygulamada kısıtlar oluşturacaktır. Condercet ise PO gibi uygulaması kolay, yüksek kapasite gerektirmeyen bir yöntemdir. Çok kriterli seçimde tercih edilebilecek bir yöntem olarak görülmektedir.

5. TARTIŞMA VE SONUÇ

Makina öğrenmesi bilgisayarların örnek veri yada geçmiş deneyimler yardımıyla daha önce tecrübe edilmemiş bir durumu tahminlenmesidir. Günlük hayatımızda gözlenmiş verilerden yeni bir durumun analiz edilmesi işletme, sağlık, haberleşme gibi bir çok alanda ihtiyaç duyulan bir konudur. Tahminlenmesi istenilen durum kimi zaman müşterilerin ne alacağı, kimi zaman kredi değerlendirmesi kimi zaman ağ yönetimi kimi zaman tanı olsade problem temelde aynıdır: gözlemlenmiş gerçek verileri kullanarak öğrenen ve yeni durum için çıkarma yapabilen bir sistem. Böyle bir sistem için en temel gereksinim öğrenme için kullanılacak büyük veri kümesidir. Bilişim teknolojilerindeki gelişmeler ihtiyaç duyulan bu verinin hem üretilmesini hemde saklanıp işlenmesini mümkün kılmıştır. Her alanda olduğu gibi biyoloji alanında da gün geçtikçe üretilen veri miktarı ve verinin karmaşıklığı artmaktadır. Ancak yeni durumların analiz edilmesi için öğrenme verisinin artması kadar bu devasa veriyi kullanacak yeni durumunu tahminleyebilecek yöntemler gerekmektedir. Verinin karmaşıklığı, biyolojik verilerin klasik istatistiksel yöntemlerin dışında, veriden öğrenen ve öğrendiği ile tahminler yapabilen makina öğrenme yöntemlerinin varlığını ortaya koymuştur. Bu gereksinim biyolojik verilerin bilgisayarlar yardımı ile analizi olarak özetlenebilecek biyoenformatik alanını doğurmuştur.

Biyoenformatik alanındaki çalışmalarda makina öğrenme yöntemleri ile biyolojik verilerde gizli olan örüntüler ve biyo-işaretçiler keşfedilmeye çalışılmaktadır. Bu amaçla, verilerin analizleri için boyutlarının azaltılması, özniteliklerin bir alt kümenin seçilmesi ve verinin sınıflandırılması, kümelenmesi, yeni durumlarının tahminlenmesi yaklaşımları uygulanmaktadır.

Bu çalışmada özniteliklerden bir alt küme seçilmesi işlemi üzerine çalışmalar yürütülmüştür. Boyut azaltma bölümünde tartışıldığı üzere yüksek boyutlu verilerin analizinde öznitelik seçimi çok önemli bir adımdır. Zira verinin sahip olduğu tüm öznitelikler problemle alakalı olmayacağı gibi, yüksek boyutlu verilerle çalışan makina

öğrenme yöntemlerinin de bazı kısıtları olacaktır. Tüm bu problemlerin üstesinden gelmek için veri analizinden önce boyut azaltma işlemi yapılmalıdır [4].

Gen ifadelenme verilerinden oluşan kanser, hipertansiyon gibi hastalık veri kümelerinde ki amaç hastalığı açıklayabilecek en az sayıda gen içeren alt kümeyi bulmaktır. Zira laboratuvar araştırmaları ile tüm genlerin analizinin yapılması ve hastalık ile ilişkisinin ortaya konması insan ömrünü açabilecek bir durumdur. Bu sebeple araştırmacılara öncelik verebilecekleri bir alt küme sunmak temel hedeftir. Bu hedef makina öğrenme yöntemlerinin öznitelik seçimi problemi ile birebir örtüşmektedir. Ancak öznitelik seçimi kendi içerisinde birçok problem barındıran bir süreçtir. Yapılan çalışmalar seçim işleminin seçim metotlarından ve veri kümesindeki çeşitlikten bağımsız olmadığını göstermiştir. Bu çalışmada öznitelik seçim yönteminin bu bağımlıklarını azaltan yöntemler geliştirilmesi hedeflenmiştir.

Öncelikle seçim işleminin bağımlıkları bir örnek veri kümei üzerinde gözlemlenmiştir. Sonuçlar,tüm veri kümesininbir örneğinin dışarıda bırakılıp, diğerinin dâhil edilmesi ile oluşturulan neredeyse özdeş altkümelerinde bile öznitelik seçim yöntemlerinin farklı seçimler üretebileceğini göstermiştir. Gözlem sonuçları bölüm 4.1 de raporlanmıştır. Yaptığımız deneyler, farklı veri kümeleri ve farklı metotların farklı öznitelikler seçmesinden dolayı tek bir yöntem yada tek bir eğitim kümesi ile elde edilen sonuçların güvenilmez olduğunu ortaya koymuştur. Bu probleme çözüm üretme amacı ile öznitelik seçim yöntemlerini geliştirecek çok kriterli yaklaşımlar önerilmiştir. Önerilen yöntemler farklı biyolojik veri kümelerine uygulanmıştır.

Kanser veri kümeleri ile yaptığımız çalışmada, öznitelik seçim ve derecelendirme işlemi için çok kriterlikarar verme yöntemi olan PO ve AHP birleştirilerek karma bir yöntem önerilmiştir.Veri kümesindeki çeşitliliğin iyi bir temsili gerçekleştirilerek, öznitelik derecelendirme ve seçim yöntemlerinin başarısı arttırılmış.Önerilen karma yaklaşım, iki sınıflı üç kanser veri kümesine uygulanmıştır. Tüm genler, sıkça kullanılan beş derecelendirme yöntemi ile derecelendirilmiş ve gen seçiminde veri kümesi çeşitliliğine bağımlılıktan kaçınmak için çoklu eğitim kümeleri kullanılmıştır.Veri kümesi N kez eğitim ve test olmak üzere iki farklı kümeye bölünmüş ve tüm özniteliklerin derecelendirilmesi için yalnızca eğitim kümesi kullanılmıştır. N tane derecelendirme değerinden oluşan ve tüm amaçları en büyük kılmaya çalışan PO yöntemi ile en önemli

genler seçilmiştir. Ardından bu genler, AHP ile önceliklendirilmiştir. PO&AHP karma yöntemiyle seçilen genlerin sınıflandırma başarısını gösterebilmek için, iki basit sınıflandırıcı ile üç gerçek veri kümesi ele alınmıştır. AHP tabanlı olarak sıralanmış gen altkümelerinin, artırımlı altküme başarımları elde edilmiş ve son olarak, daha az sayıda gen içeren alt kümelerin, basit ve temel sınıflandırıcılar ile elde edilen başarısı, çok daha fazla gene sahip alt kümelerin gelişmiş sınıflandırıcılar ile başarısıyla karşılaştırılmıştır.

Önerilen çok kriterlikarma yöntem benzer ancak farklı bir yaklaşımla hipertansiyon veri kümesine de uygulanmıştır. Yine PO yöntemi ile çoklu öğrenme kümeleri kullanılarak veri seti varyasyonuna bağımlığı en aza indirgenmiştir. Ayrıca 5 farklı metodun birleşim kümesi kullanılarak öznitelik seçim işleminin metot bağımlığının azaltılması amaçlanmıştır. AHP ile de çoklu kriter yardımıyla genler önceliklendirilmiştir. Seçtiğimiz genlerin veri kümesini toplayan çalışma ve genepattern analizi tarafından da önemli görülmesi sonuçlarımızı doğrulamaktadır. Alınan sonuçlar daha ilişkili genlerin AHP skorlarının varsayılan öncelikten daha yüksek olduğunu göstermiştir.

Sonuç olarak, çok kriterli karar verme işleminde, yerel uzmanları birleştiren PO&AHP karma yaklaşımının, kanser ve hipertansiyon hastalığıyla en çok ilgili genlerin seçiminde kullanılabileceği görülmüştür. Buna ek olarak, PO, genleri seçmek için çoklu kriterler kullanarak veri kümesi çeşitliliğine bağımlılığı en düşük seviyeye indirgemiştir. Ayrıca, PO, en başarılı gen altkümesini seçerken, seçilecek eleman sayısı ya da seçim eşiği gibi parametrelere ihtiyaç duymadığında alt kümenin sayısı yöntemce belirlenmiştir. AHP ise, çoklu kriterler üzerinden genlere birer öncelik atar. AHP sonucunda, hastalıkla yakından ilgili genler önceden atanan önceliklerinden daha büyük önceliklere sahip olurlar. PO ve AHP'yi sıralı bir biçimde uygulamak, öznitelik seçim yöntemlerinin başarısını arttırır. Yerel yöntemlerin çok kriterli karar verme metotları ile başarılı bir şekilde birleştirilmesi, eş zamanlı olarak farklı amaçlara hizmet eden yerel öznitelik seçim metotlarının da birleştirilebileceğini düşündürmüştür. Bu eş zamanlı birleştirme ileriki çalışmalar olarak hedeflenmiştir.

Mikrodizi verilerinde, önerilen karma yöntem klasik mikrodizi yöntemlerince seçilen nispeten çok sayıdaki genin daha aza indirilmesi amacı ile de kullanılabilir. Buradan hareketle, önerdiğimiz yöntemin mikrodizi analizlerine uygulanmasında, genlerin

makine öğrenme yöntemleri ile sıralanması değil, klasik yöntemlerce sıralanıp önerilen karma yöntemle seçime tabi tutulması ileriki çalışmalar olarak düşünülmüştür.

Genom boyu ilişkilendirme çalışmaları, biyo-işaretçiler(örneğin SNP) ve fenotipler arasındaki ilişkiyi bulmayı amaçlar. Bu ilişki bazen açıktır ve basit istatistikler kullanılarak ortaya çıkarılabilir. Bununla birlikte, bazen çoklu amaçların gerektiği durumlar mevcut olabilir. Bu yüzden SNP seçimi, hastalıklarla ilişki, etnik gruplama, jeo-genomik korelasyon, göç yolları gibi birçok amaçla ilişkili olabilecek küçük SNP alt kümelerinin seçimine hizmet edebilir. Böylesine farklı amaçlar için pek çok farklı SNP derecelendirmeleri kullanılarak SNP'lerin ilgili bir altkümesini oluşturabilmek amacıyla Pareto Optimal, Condorcet, MC4 ve AHP gibi farklı birleştirme yöntemleri kullanılabilir. Bu çalışmada, seçilen SNP altkümelerinin faydaları değerlendirilirken iki amaç ele alınmıştır: etnik grupların sınıflandırılma doğruluğu ve jeo-genomik mesafe korelasyonu. SNP'lerin bu amaçlara ne kadar hizmet ettiği karışılıklı bilgi(MI) ve Relief-F skoru (ilk amaca hizmet edecek şekilde) ve temel bileşen yükleri (ikinci amaca hizmet edecek şekilde) ile ölçülmüştür.

Özetle bu çalışmada öznitelik seçim işleminin veri kümesindeki değişintiden ve seçim metodundan etkilendi ortaya konmuştur. Veri kümesindeki değişintiye bağımlılığın çözümü için ilk önce çoklu öğrenme kümesi ile öznitelik seçimi önerilmiştir. Bu amaçla iki çok kriterli karar verme yöntemini birleştiren karma bir model ortaya konmuştur. Model içerisinde kullanılması önerilen PO yöntemi biyoenformatik alanında bizim kullandığımız anlamda öznitelik seçimi için ilk defa kullanılmıştır. Literatürde genellikle en az sayı-en çok başarı getirecek öznitelik grubunu farklı metotlarca seçmek şekilde özetlenebilecek 2 amaçlı seçim problemlerinde kullanılmıştır. Biz veriyi farklı açılardan temsil edecek birden çok öğrenme kümesinin sonuçlarını birleştirmek amacı ile PO yöntemini kullandık. PO yönteminin seçtiği öznitelikleri sıralamıyor olması ise karşımıza çıkan bir problem olmuştur. Bu problemin üstesinden gelmek için ise seçilen öznitelikleri sıralamak için bir çok kriterli karar verme problemi olan AHP yönteminin kullanılması önerilmiştir. Yöneylem araştırmalarında sıkça kullanılan bu yöntem biyoenformatik alanında tanınmamaktadır. Bu çalışma ile AHP yöntemini sayıca az olan öznitelikleri birçok kriteri göz önünde bulundurmak sureti ile sıralamak için kullandık. Alınan sonuçlar AHP tabanlı sıralamanın başarılı olduğunu göstermektedir.

Öznitelik seçiminde bir diğer sorun seçim metoduna bağımlılıktır. Bu problemin çözümü için ise farklı amaçlara hizmet eden metotların yine çok kriterli karar verme yöntemleri ile birleştirilmesi önerilmiştir. Bunoktada PO yöntemi yine problemimize çözüm sunmaktadır ve amaçlanan hedeflere eş zamanlı hizmet edebilecek öznitelikleri seçebilmiştir.

KAYNAKLAR

[1]. S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *Lancet,* vol. 365, pp. 488-492, Feb 5 2005.

[2]. F. Yang and K. Z. Mao, "Robust Feature Selection for Microarray Data Based on Multicriterion Fusion," *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 8, pp. 1080-1092, Jul-Aug 2011.

[3]. Z. Gormez, O. Kursun, A. Sertbas, N. Aydin, and H. Seker, "Statistical bias and variance of gene selection and cross validation methods: A case study on hypertension prediction," in *Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on*, 2012, pp. 616-619.

[4]. E. Alpaydın, *Yapay öğrenme*: Boğaziçi Üniversitesi Yayınevi, 2011.

[5]. Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, pp. 2507-2517, Oct 1 2007.

[6]. J. J. Chen, C. A. Tsai, S. L. Tzeng, and C. H. Chen, "Gene selection with multiple ordering criteria," *Bmc Bioinformatics,* vol. 8, Mar 5 2007.

[7]. Y. Feng and K. Z. Mao, "Improving robustness of gene ranking by resampling and permutation based score correction and normalization," in *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, 2010, pp. 444-449.

[8]. L. Y. Chuang, C. H. Yang, K. C. Wu, and C. H. Yang, "A hybrid feature selection method for DNA microarray data," *Computers in Biology and Medicine,* vol. 41, pp. 228-237, Apr 2011.

[9]. S. Niijima and S. Kuhara, "Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE," *Bmc Bioinformatics,* vol. 7, Dec 25 2006.

[10]. L. K. Luo, L. J. Ye, M. X. Luo, D. F. Huang, H. Peng, and F. Yang, "Methods of forward feature selection based on the aggregation of classifiers generated by single attribute," *Computers in Biology and Medicine,* vol. 41, pp. 435-441, Jul 2011.

[11]. P. H. Lee, J. Y. Jung, and H. Shatkay, "Functionally Informative Tag SNP Selection Using a Pareto-Optimal Approach," *Advances in Computational Biology,* vol. 680, pp. 173-180, 2010.

[12]. O. Uncu and I. B. Turksen, "A novel feature selection approach: Combining feature wrappers and filters," *Information Sciences,* vol. 177, pp. 449-466, Jan 15 2007.

[13]. E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach," *Evolutionary Computation, IEEE Transactions on,* vol. 3, pp. 257-271, 1999.

[14]. K. Deb and A. R. Reddy, "Reliable classification of two-class cancer data using evolutionary algorithms," *Biosystems,* vol. 72, pp. 111-129, Nov 2003.

[15]. A. J. Soto, R. L. Cecchini, G. E. Vazquez, and I. Ponzoni, "Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach," *Qsar & Combinatorial Science,* vol. 28, pp. 1509-1523, Dec 2009.

[16]. C. K. Ting, W. T. Lin, and Y. T. Huang, "Multi-objective tag SNPs selection using evolutionary algorithms," *Bioinformatics,* vol. 26, pp. 1446-1452, Jun 1 2010.

[17]. S. Sabzevari and S. Abdullah, "Gene Selection in Microarray Data From Multi-Objective Perspective," *2011 3rd Conference on Data Mining and Optimization (Dmo),* pp. 199-207, 2011.

[18]. J. Handl, D. B. Kell, and J. Knowles, "Multiobjective optimization in bioinformatics and computational biology," *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 4, pp. 279-292, Apr-Jun 2007.

[19]. L. P. Dinu and M. Popescu, "A Multi-Criteria Decision Method Based on Rank Distance," *Fundamenta Informaticae,* vol. 86, pp. 79-91, 2008.

[20]. S. Prakash, P. Kumar, B. V. N. S. Prasad, and A. Gupta, "Pareto optimal solutions of a cost-time trade-off bulk transportation problem," *European Journal of Operational Research,* vol. 188, pp. 85-100, Jul 1 2008.

[21]. I. Kacem, S. Hammadi, and P. Borne, "Pareto-optimality approach for flexible job-shop scheduling problems: hybridization of evolutionary algorithms and fuzzy logic," *Mathematics and Computers in Simulation,* vol. 60, pp. 245-276, Sep 30 2002.

[22]. R. Tavakkoli-Moghaddam, M. Azarkish, and A. Sadeghnejad-Barkousaraie, "A new hybrid multi-objective Pareto archive PSO algorithm for a bi-objective job shop scheduling problem," *Expert Systems with Applications,* vol. 38, pp. 10812-10821, Sep 2011.

[23]. H. Yumurtaci-Aydogmus, "Variable Neighbourhood Search Heuristic Method and an Application in A Supply Chain Management," in *Institute of Science*. vol. Phd Turkey: Istanbul University, 2011.

[24]. J. Gao, "Model and Algorithm of Vehicle Routing Problem with Time Windows in Stochastic Traffic Network," *Proceedings of 2010 International Conference on Logistics Systems and Intelligent Management, Vols 1-3,* pp. 848-851, 2010.

[25]. G. Fleury, A. Hero, S. Yoshida, T. Carter, C. Barlow, and A. Swaroop, "Pareto analysis for gene filtering in microarray experiments," in *European Signal Processing Conference (EUSIPSO 2002)* Toulouse, France, 2002.

[26]. E. Bagheri, M. Asadi, D. Gasevic, and S. Soltani, "Stratified Analytic Hierarchy Process: Prioritization and Selection of Software Features," *Software Product Lines: Going Beyond,* vol. 6287, pp. 300-315, 2010.

[27]. S. Koul, A. Saraswat, and R. Verma, "Evaluation and ranking of supplier at a service firm using analytic hierarchy process," in *Information and Communication Technologies (WICT), 2011 World Congress on*, 2011, pp. 922-927.

[28]. A. W. Labib, "A supplier selection model: a comparison of fuzzy logic and the analytic hierarchy process," *International Journal of Production Research,* vol. 49, pp. 6287-6299, 2011.

[29]. S. H. Amin and G. Q. Zhang, "An integrated model for closed-loop supply chain configuration and supplier selection: Multi-objective approach," *Expert Systems with Applications,* vol. 39, pp. 6782-6791, Jun 15 2012.

[30]. M. Danner, J. M. Hummel, F. Volz, J. G. van Manen, B. Wiegard, C. M. Dintsios, H. Bastian, A. Gerber, and M. J. IJzerman, "Integrating patients' views into health technology assessment: Analytic hierarchy process (AHP) as a method to elicit patient preferences," *International Journal of Technology Assessment in Health Care,* vol. 27, pp. 369-375, Oct 2011.

[31]. N. Basoglu, T. U. Daim, and U. Topacan, "Determining Patient Preferences for Remote Monitoring," *Journal of Medical Systems,* vol. 36, pp. 1389-1401, Jun 2012.

[32]. M. Sadeghi and A. Ameli, "An AHP decision making model for optimal allocation of energy subsidy among socio-economic subsectors in Iran," *Energy Policy,* vol. 45, pp. 24-32, Jun 2012.

[33]. C. Shin, J. Cho, J. G. Kim, and B. Lee, "An AHP-based resource management scheme for CRRM in heterogeneous wireless networks," *Annals of Telecommunications-Annales Des Telecommunications,* vol. 67, pp. 511-522, Dec 2012.

[34]. T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGraw Hill. International, 1980.

[35]. O. Kadioglu, G. Ustunkar, and Y. A. Son, "GWAS with AHP based SNP prioritization approach to identify SNP biomarkers for Alzheimer's disease," in *Health Informatics and Bioinformatics (HIBIT), 2012 7th International Symposium on*, 2012, pp. 7-10.

[36]. K. S. Lynn, L. L. Li, Y. J. Lin, C. H. Wang, S. H. Sheng, J. H. Lin, W. Liao, W. L. Hsu, and W. H. Pan, "A neural network model for constructing endophenotypes of common complex diseases: an application to male young-onset hypertension microarray data," *Bioinformatics,* vol. 25, pp. 981-988, Apr 15 2009.

[37]. C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J Bioinform Comput Biol,* vol. 3, pp. 185-205, Apr 2005.

[38]. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 96, pp. 6745-6750, Jun 8 1999.

[39]. M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 98, pp. 11462-11467, Sep 25 2001.

[40]. M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, pp. 68-74, Jan 2002.

[41]. L. L. Cavalli-Sforza, "The Human Genome Diversity Project: past, present and future," *Nat Rev Genet*, vol. 6, pp. 333-340, 04//print 2005.

[42]. T. Vincenty, "{Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations}," *Survey Review*, vol. 22, pp. 88-93, // 1975.

[43]. A. Gupta, K. G. Mehrotra, and C. Mohan, "A clustering-based discretization for supervised learning," *Statistics & Probability Letters*, vol. 80, pp. 816-824, // 2010.

[44]. C. M. Bishop, *Neural Networks for Pattern Recognition*: Oxford University Press, Inc., 1995.

[45]. M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, pp. 23-69, Oct-Nov 2003.

[46]. I. Kononenko, "Estimating attributes: Analysis and extensions of RELIEF," in *Machine Learning: ECML-94*. vol. 784, F. Bergadano and L. Raedt, Eds.: Springer Berlin Heidelberg, 1994, pp. 171-182.

[47]. J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante, "Genes mirror geography within Europe," *Nature*, vol. 456, pp. 98-101, 11/06/print 2008.

[48]. P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas, "PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations," *PLoS Genet*, vol. 3, p. e160, 2007.

[49]. E. Gumus, Z. Gormez, and O. Kursun, "Multi objective SNP selection using pareto optimality," *Computational Biology and Chemistry*, vol. 43, pp. 23-28, Apr 2013.

[50]. B. Srdjevic, "Combining different prioritization methods in the analytic hierarchy process synthesis," *Computers & Operations Research*, vol. 32, pp. 1897-1919, Jul 2005.

[51]. T. L. Saaty, "How to make a decision: The analytic hierarchy process," *European Journal of Operational Research*, vol. 48, pp. 9-26, 1990.

[52]. E. Triantaphyllou and S. H. Mann, "Using the Analytic Hierarchy Process for Decision Making in Engineering Applications: Some Challenges," *International Journal of Industrial Engineering Applications and Practice*, vol. 2, pp. 35-44, 1995.

[53]. J. A. N. de Caritat marquis de Condorcet, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*: De l'Imprimerie royale, 1785.

[54]. R. C. Prati, "Combining feature ranking algorithms through rank aggregation," *2012 International Joint Conference on Neural Networks (Ijcnn),* 2012.

[55]. V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach," *Bioinformatics,* vol. 23, pp. 1607-1615, Jul 1 2007.

[56]. R. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni, "Combining results of microarray experiments: A rank aggregation approach," *Statistical Applications in Genetics and Molecular Biology,* vol. 5, Jul 2 2006.

[57]. T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on,* vol. 13, pp. 21-27, 1967.

[58]. S. Baek, C. A. Tsai, and J. J. Chen, "Development of biomarker classifiers from high-dimensional data," *Briefings in Bioinformatics,* vol. 10, pp. 537-546, Sep 2009.

[59]. Y. Zhang, C. Ding, and T. Li, "Gene selection algorithm by combining reliefF and mRMR," *Bmc Genomics,* vol. 9, 2007.

[60]. Z. Gormez, E. Gumus, A. Sertbas, and O. Kursun, "Comparison of aggregators for multi-objective SNP selection," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 3062-3065.

[61]. M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, "GenePattern 2.0," *Nat Genet,* vol. 38, pp. 500-1, May 2006.

[62]. M. G. Butler, "Genetics of hypertension. Current status," *J Med Liban,* vol. 58, pp. 175-8, Jul-Sep 2011.

[63]. J. Lee, N. Batnyam, and S. Oh, "RFS: efficient feature selection method based on R-value," *Computers in Biology and Medicine,* vol. 43, pp. 91-9, Feb 2013.

[64]. M. Pirooznia, J. Y. Yang, M. Q. Yang, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data," *Bmc Genomics,* vol. 9 Suppl 1, p. S13, 2008.

[65]. W. Zhou and J. A. Dickerson, "A novel class dependent feature selection method for cancer biomarker discovery," *Computers in Biology and Medicine,* vol. 47, pp. 66-75, Apr 2014.

[66]. J. M. Arevalillo and H. Navarro, "Exploring correlations in gene expression microarray data for maximum predictive-minimum redundancy biomarker selection and classification," *Computers in Biology and Medicine,* vol. 43, pp. 1437-43, Oct 2013.

[67]. R. P. Lifton, "Molecular genetics of human blood pressure variation," *Science,* vol. 272, pp. 676-80, May 3 1996.

[68]. O. A. Carretero and S. Oparil, "Essential hypertension. Part I: definition and etiology," *Circulation,* vol. 101, pp. 329-35, Jan 25 2000.

[69]. A. Binder, "A review of the genetics of essential hypertension," *Curr Opin Cardiol,* vol. 22, pp. 176-84, May 2007.

[70]. D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research,* vol. 38, pp. W214-20, Jul 2010.

[71]. R. Mani, R. P. St Onge, J. L. t. Hartman, G. Giaever, and F. P. Roth, "Defining genetic interaction," *Proc Natl Acad Sci U S A,* vol. 105, pp. 3461-6, Mar 4 2008.

[72]. M. Michaut and G. D. Bader, "Multiple genetic interaction experiments provide complementary information useful for gene function prediction," *PLoS Comput Biol,* vol. 8, p. e1002559, 2012.

[73]. Y. Chen, F. Le Caherec, and S. L. Chuck, "Calnexin and other factors that alter translocation affect the rapid binding of ubiquitin to apoB in the Sec61 complex," *J Biol Chem,* vol. 273, pp. 11887-94, May 8 1998.

[74]. S. Razick, G. Magklaras, and I. M. Donaldson, "iRefIndex: a consolidated protein interaction database with provenance," *BMC Bioinformatics,* vol. 9, p. 405, 2008.

[75]. P. M. Frossard, E. N. Obineche, and G. G. Lestringant, "Association of an apolipoprotein B gene marker with essential hypertension," *Hypertension,* vol. 33, pp. 1052-6, Apr 1999.

[76]. B. Das, N. Pawar, D. Saini, and M. Seshadri, "Genetic association study of selected candidate genes (ApoB, LPL, Leptin) and telomere length in obese and hypertensive individuals," *BMC Med Genet,* vol. 10, p. 99, 2009.

[77]. K. Lorenz, M. J. Lohse, and U. Quitterer, "Protein kinase C switches the Raf kinase inhibitor from Raf-1 to GRK-2," *Nature,* vol. 426, pp. 574-9, Dec 4 2003.

[78]. D. J. Kelly, Y. Zhang, C. Hepper, R. M. Gow, K. Jaworski, B. E. Kemp, J. L. Wilkinson-Berka, and R. E. Gilbert, "Protein kinase C beta inhibition attenuates the progression of experimental diabetic nephropathy in the presence of continued hypertension," *Diabetes,* vol. 52, pp. 512-8, Feb 2003.

[79]. N. Franceschini, F. J. van Rooij, B. P. Prins, M. F. Feitosa, M. Karakas, J. H. Eckfeldt, A. R. Folsom, J. Kopp, A. Vaez, J. S. Andrews, J. Baumert, V. Boraska, L. Broer, C. Hayward, J. S. Ngwa, Y. Okada, O. Polasek, H. J. Westra, Y. A. Wang, M. F. Del Greco, N. L. Glazer, K. Kapur, I. P. Kema, L. M. Lopez, A. Schillert, A. V. Smith, C. A. Winkler, L. Zgaga, S. LifeLines Cohort, S. Bandinelli, S. Bergmann, M. Boban, M. Bochud, Y. D. Chen, G. Davies, A. Dehghan, J. Ding, A. Doering, J. P. Durda, L. Ferrucci, O. H. Franco, L. Franke, G. Gunjaca, A. Hofman, F. C. Hsu, I. Kolcic, A. Kraja, M. Kubo, K. J. Lackner, L. Launer, L. R. Loehr, G. Li, C. Meisinger, Y. Nakamura, C. Schwienbacher, J. M. Starr, A. Takahashi, V. Torlak, A. G. Uitterlinden, V. Vitart, M. Waldenberger, P. S. Wild, M. Kirin, T. Zeller, T. Zemunik, Q. Zhang, A. Ziegler, S. Blankenberg, E. Boerwinkle, I. B. Borecki, H. Campbell, I. J. Deary, T. M. Frayling, C. Gieger, T. B. Harris, A. A. Hicks, W. Koenig, O. D. CJ, C. S. Fox, P. P. Pramstaller, B. M. Psaty, A. P. Reiner, J. I. Rotter, I. Rudan, H. Snieder, T. Tanaka, C. M. van Duijn, P. Vollenweider, G. Waeber, J. F. Wilson, J. C. Witteman, B. H. Wolffenbuttel, A. F. Wright, Q. Wu, Y. Liu, N. S. Jenny, K. E. North, J. F. Felix, B. Z. Alizadeh, L. A. Cupples, J. R. Perry and A. P. Morris,

"Discovery and fine mapping of serum protein loci through transethnic meta-analysis," *Am J Hum Genet,* vol. 91, pp. 744-53, Oct 5 2012.

[80]. A. Lin, R. T. Wang, S. Ahn, C. C. Park, and D. J. Smith, "A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes," *Genome Res,* vol. 20, pp. 1122-32, Aug 2010.

[81]. E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander, "Pathway Commons, a web resource for biological pathway data," *Nucleic Acids Res,* vol. 39, pp. D685-90, Jan 2011.

[82]. E. E. Schadt, S. W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K. W. Hart, A. Russell, G. Li, G. Cavet, J. Castle, P. McDonagh, Z. Kan, R. Chen, A. Kasarskis, M. Margarint, R. M. Caceres, J. M. Johnson, C. D. Armour, P. W. Garrett-Engele, N. F. Tsinoremas, and D. D. Shoemaker, "A comprehensive transcript index of the human genome generated using microarrays and computational approaches," *Genome Biol,* vol. 5, p. R73, 2004.

[83]. J. Heineke, M. Auger-Messier, R. N. Correll, J. Xu, M. J. Benard, W. Yuan, H. Drexler, L. V. Parise, and J. D. Molkentin, "CIB1 is a regulator of pathological cardiac hypertrophy," *Nat Med,* vol. 16, pp. 872-9, Aug 2010.

[84]. V. A. McKusick, "Mendelian Inheritance in Man and its online version, OMIM," *Am J Hum Genet,* vol. 80, pp. 588-604, Apr 2007.

[85]. J. M. Johnson, J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker, "Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays," *Science,* vol. 302, pp. 2141-4, Dec 19 2003.

[86]. W. Wang, M. L. Asp, G. Guerrero-Serna, and J. M. Metzger, "Differential effects of S100 proteins A2 and A6 on cardiac Ca(2+) cycling and contractile performance," *J Mol Cell Cardiol,* vol. 72, pp. 117-25, Jul 2014.

[87]. R. Higdon, E. Stewart, L. Stanberry, W. Haynes, J. Choiniere, E. Montague, N. Anderson, G. Yandl, I. Janko, W. Broomall, S. Fishilevich, D. Lancet, N. Kolker, and E. Kolker, "MOPED enables discoveries through consistently processed proteomics data," *J Proteome Res,* vol. 13, pp. 107-13, Jan 3 2014.

[88]. J. Wang, L. Xu, X. Yun, K. Yang, D. Liao, L. Tian, H. Jiang, and W. Lu, "Proteomic analysis reveals that proteasome subunit beta 6 is involved in hypoxia-induced pulmonary vascular remodeling in rats," *PLoS One,* vol. 8, p. e67942, 2013.

[89]. T. M. Bahr, G. J. Hughes, M. Armstrong, R. Reisdorph, C. D. Coldren, M. G. Edwards, C. Schnell, R. Kedl, D. J. LaFlamme, N. Reisdorph, K. J. Kechris, and R. P. Bowler, "Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease," *Am J Respir Cell Mol Biol,* vol. 49, pp. 316-23, Aug 2013.

[90]. C. L. Huang, E. Kuo, and R. D. Toto, "WNK kinases and essential hypertension," *Curr Opin Nephrol Hypertens,* vol. 17, pp. 133-7, Mar 2008.

[91]. K. A. Choate, K. T. Kahle, F. H. Wilson, C. Nelson-Williams, and R. P. Lifton, "WNK1, a kinase mutated in inherited hypertension with hyperkalemia, localizes

to diverse Cl- -transporting epithelia," *Proc Natl Acad Sci U S A,* vol. 100, pp. 663-8, Jan 21 2003.

[92]. G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biol,* vol. 11, p. R53, 2010.

[93]. M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The Pfam protein families database," *Nucleic Acids Res,* vol. 40, pp. D290-301, Jan 2012.