T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOKTORA TEZİ

SYMPES YÖNTEMİYLE KANSER HASTALIKLARINDA ÖNEMLİ GENLERİN SAPTANMASI VE KANSER TÜRLERİNİN SINIFLANDIRILMASI

Ali SARIKAŞ

DANIŞMAN
Prof. Dr. Sıddık YARMAN

II. DANIŞMAN
Dr. Zeliha GÖRMEZ

Biyomedikal Mühendisliği Anabilim Dalı

Biyomedikal Mühendisliği Programı

İSTANBUL-2020

Bu çalışma, 22.07.2020 tarihinde aşağıdaki jüri tarafından Biyomedikal Mühendisliği Anabilim Dalı, Biyomedikal Mühendisliği Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Sıddık YARMAN(Danışman)
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Fırat KAÇAR
İstanbul Üniversitesi - Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Ali BULDU
Marmara Üniversitesi
Teknoloji Fakültesi

Prof. Dr. Osman Nuri UÇAN
Altınbaş Üniversitesi
Mühendislik ve Doğa Bilimleri Fakültesi

Doç. Dr. Yasin ÖZÇELEP
İstanbul Üniversitesi - Cerrahpaşa
Mühendislik Fakültesi

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi-Cerrahpaşa'nın abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Tez çalışmalarım süresince, karşılaştığım problemlerde tecrübesini ve desteğini hiçbir zaman esirgemeyen, yapıcı önerileriyle doğruları bulmamda beni sürekli cesaretlendiren, çalıştığımız süre boyunca kendisinden akademik ve insani olarak çok şey öğrendiğim çok değerli danışmanım Prof. Dr. Sıddık YARMAN'a en içten duygularımla teşekkür ederim.

Yine tez çalışmalarım boyunca, ihtiyaç duyduğum her zaman çözüm odaklı önerilerini ve yardımlarını esirgemeyen, insani olarak da kendisinden çok şey öğrendiğim eş danışmanım Dr. Zeliha GÖRMEZ'e teşekkürü bir borç bilirim.

Lisans öğrenimimden bu yana her zaman bilgi ve tecrübesinin yanında güler yüzü ile de her zaman yanımda olan, çalışmaktan inanılmaz mutluluk duyduğum ve bundan sonra da duyacağımdan çok emin olduğum, çok değerli hocam Prof. Dr. Ali BULDU'ya en içten dileklerimle teşekkür ederim.

Yine doktora öğrenimim süresince tanıdığım, ihtiyaç duyduğum her anda desteğini hissettiğim, akademik çalışmaları ve insani yönleriyle bana örnek olan, çok değerli hocam Prof. Dr. Fırat KAÇAR'a teşekkürü bir borç bilirim.

Son 1 yıldır tanıdığım, ihtiyaç duyduğum her zaman danışabildiğim, desteğini ve samimiyetini her zaman hissettiğim arkadaşım Osman KORKMAZ'a teşekkürü bir borç bilirim.

Hayatım boyunca maddi ve manevi desteklerini hissettiğim aileme ve sevdiklerime, özellikle tez dönemimde inanılmaz sabır gösteren ve destek veren, hayatıma anlam ve değer katan canım eşime, tez sürecimin son bir yılında hayatıma giren ve yaşama sevincimi, hayat tecrübemi inanılmaz artıran canım oğluma en içten duygularımla teşekkür ederim.

Temmuz 2020 Ali SARIKAŞ

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **viii**
TABLO LİSTESİ ..... **xii**
SİMGE VE KISALTMA LİSTESİ ..... **xv**
ÖZET ..... **xix**
SUMMARY ..... **xxi**
1. GİRİŞ **1**
1.1. LİTERATÜR ÖZETİ 2
1.2. TEZİN AMACI VE KATKILARI 7
2. GENEL KISIMLAR **8**
2.1. İNSAN GENOMU, GEN YAPISI VE İŞLEVİ 8
2.2. GEN İFADESİNİN TEMELLERİ VE SANTRAL DOGMA 9
2.3. KANSERDE GENETİK 12
2.4. MİKRODİZİ TEKNOLOJİSİ 14
2.5. GEN İFADESİ OMNİBUS 16
3. MALZEME VE YÖNTEM **18**
3.1. ÇALIŞMADA KULLANILAN VERİ KÜMELERİ 18
3.2. VERİ ÖN İŞLEME YÖNTEMLERİ 19
3.2.1. Eksik Değer Kestirimi 19
K-En Yakın Komşu Yöntemi *20*
Bayesçi Temel Bileşen Analizi Yöntemi *21*
Yerel En Küçük Kareler Yöntemi *21*
Ortalama ve Ortanca Değer Yöntemleri *21*
3.2.2. Veri Dönüşümü 21
Logaritmik Dönüşüm *22*
Box-Cox Dönüşüm Ailesi *22*
İlgin Dönüşüm *22*
3.2.3. Veri Normalizasyonu 23
Minimum – Maksimum Normalizasyonu *23*

Z-değeri Normalizasyonu *23*
Kantil Normalizasyon *24*
3.3. İSTATİSTİKSEL ÇÖZÜMLEME YÖNTEMLERİ 25
3.3.1. Kat Değişimi 25
3.3.2. Welch t – testi 26
3.3.3. Tek Yönlü Varyans Analizi 26
3.3.4. Wilcoxon Sıra Sayıları – Toplam Testi 28
3.3.5. Kruskal Wallis – H Testi 30
3.4. ÇOKLU KARŞILAŞTIRMA TESTLERİ 31
3.5. GEO2R PLATFORMU İLE FARKSAL OLARAK İFADE EDİLEN GENLERİN TESPİTİ 32
3.6. SINIFLANDIRMA YÖNTEMLERİ 37
4. GEN İFADE VERİLERİNİN MODELLENMESİNDE YENİ YAKLAŞIM: SYMPES 39
4.1. GEN İFADE VERİ KÜMELERİ İÇİN ÖNCEDEN TANIMLI ENERJİ VE BİLGİ BLOKLARININ BULUNMASI 39
4.1.1. SYMPES Kodlama Algoritmasının Aşamaları 50
4.1.2. SYMPES Kod Çözme Algoritmasının Aşamaları 54
4.2. ÖNCEDEN TANIMLI BİLGİ BLOKLARINDAN EVRENSEL ORTALAMA VE ORTANCA BİLGİ KÜTÜPHANELERİNİN OLUŞTURULMASI 55
4.3. SYMPES KÜTÜPHANELERİ KULLANILARAK YENİ BİR SINIFLANDIRMA YAKLAŞIMI 56
5. BULGULAR 59
5.1. RASGELE SEÇİLEN BİR DİZİNİN SYMPES YÖNTEMİ İLE DEĞERLENDİRİLMESİ 59
5.2. ÖNEMLİ GEN ÇERÇEVELERİNİN VE FARKSAL OLARAK İFADE EDİLEN GENLERİN BULUNMASI 60
5.2.1. Meme Kanseri Veri Kümesinin Sonuçları 61
5.2.2. Akciğer Kanseri Veri Kümesinin Sonuçları 74
5.2.3. Over Kanseri Veri Kümesinin Sonuçları 81
5.2.4. Prostat Kanseri Veri Kümesinin Sonuçları 89
5.2.5. Pankreas Kanseri Veri Kümesinin Sonuçları 97
5.3. SINIFLANDIRMA 111
5.3.1. Meme Kanseri Veri Kümesinin Sonuçları 111
5.3.2. Akciğer Kanseri Veri Kümesinin Sonuçları 112
5.3.3. Over Kanseri Veri Kümesinin Sonuçları 113

5.3.4. Prostat Kanseri Veri Kümesinin Sonuçları....114
5.3.5. Pankreas Kanseri Veri Kümesinin Sonuçları....115
6. TARTIŞMA VE SONUÇ....117
KAYNAKLAR....122
EKLER....127
EK-A: İSTATİSTİKSEL TESTLER....127
ÖZGEÇMİŞ....128

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1: DNA'nın yapısı [1]. 9

Şekil 2.2: Çift sarmal DNA'nın replikasyonu [1]. 10

Şekil 2.3: DNA'dan RNA'ya ve proteine bilgi akışı [1]. 12

Şekil 2.4: Affymetrix GeneChip mikrodizi ile bir deneyin akış şeması. 15

Şekil 2.5: GEO veri tabanının web arayüzü. 17

Şekil 3.1: İlgilenilen veri kümesinin kayıt numarasının girilmesi. 33

Şekil 3.2: Referans seri numarasının görüntülenmesi. 34

Şekil 3.3: GEO2R platfromunda referans seri numarasının girilmesi. 34

Şekil 3.4: Grupların tanımlanması. 35

Şekil 3.5: Gruplara örneklerin atanması. 35

Şekil 3.6: Ayarlar (isteğe bağlı). 36

Şekil 3.7: Farksal olarak ifade edilen ilk 250 genin hesaplanması. 36

Şekil 3.8: Farksal olarak ifade edilen ilk 250 genin listelenmesi 37

Şekil 4.1: Sürekli işaretin örneklenerek ayrık işarete dönüştürülmesi. 39

Şekil 4.2: Bir diziye ait gen ifade verisinin çerçevelere bölünmesi. 42

Şekil 4.3: Tüm veri kümesine ait gen ifade verisinin bloklarına ayrılması ile vektörel formda yeniden düzenlenmesinin gösterimi. 44

Şekil 4.4: PESB oluşturma basamakları akış şeması. 49

Şekil 4.5: SYMPES kodlama süreci akış şeması. 52

Şekil 4.6: SYMPES kod çözme süreci akış şeması. 53

Şekil 4.7: SYMPES ile sınıflandırma işlemi için akış şeması. 58

Şekil 5.1: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 66

Şekil 5.2: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 66

Şekil 5.3: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için) 67

Şekil 5.4: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için) 67

Şekil 5.5: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için) 68

Şekil 5.6: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için) 68

Şekil 5.7: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 5, \alpha = 0.05$). 70

Şekil 5.8: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 9, \alpha = 0.05$). 71

Şekil 5.9: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 15, \alpha = 0.05$). 72

Şekil 5.10: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 25, \alpha = 0.05$). 73

Şekil 5.11: Akciğer kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 77

Şekil 5.12: Akciğer kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 77

Şekil 5.13: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 5, \alpha = 0.05$). 78

Şekil 5.14: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 9, \alpha = 0.05$). 79

Şekil 5.15: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 15, \alpha = 0.05$). 80

Şekil 5.16: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 25, \alpha = 0.05$). 81

Şekil 5.17: Over kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 84

Şekil 5.18: Over kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 85

Şekil 5.19: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 5, \alpha = 0.05$). 86

Şekil 5.20: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 9, \alpha = 0.05$). 87

Şekil 5.21: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 15, \alpha = 0.05$). 88

Şekil 5.22: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 25, \alpha = 0.05$). 89

Şekil 5.23: Prostat kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 92

Şekil 5.24: Prostat kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 93

Şekil 5.25: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 5, \alpha = 0.05$). 94

Şekil 5.26: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 9, \alpha = 0.05$). 95

Şekil 5.27: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 15, \alpha = 0.05$). 96

Şekil 5.28: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 25, \alpha = 0.05$). 97

Şekil 5.29: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 103

Şekil 5.30: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için) 103

Şekil 5.31: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için) 104

Şekil 5.32: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için) 104

Şekil 5.33: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için) 105

Şekil 5.34: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için) 105

Şekil 5.35: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 5, \alpha = 0.05$). 107

Şekil 5.36: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 9, \alpha = 0.05$). 108

Şekil 5.37: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 15, \alpha = 0.05$) 109

Şekil 5.38: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($LF = 25, \alpha = 0.05$) 110

Şekil 5.39: Meme kanseri veri kümesi için sınıflandırma sonuçları. 112

Şekil 5.40: Akciğer kanseri veri kümesi için sınıflandırma sonuçları. 113

Şekil 5.41: Over kanseri veri kümesi için sınıflandırma sonuçları. 114

Şekil 5.42: Prostat kanseri veri kümesi için sınıflandırma sonuçları. 115

Şekil 5.43: Pankreas kanseri veri kümesi için sınıflandırma sonuçları. 116

TABLO LİSTESİ

Sayfa No

Tablo 1.1: Genetik hastalıkların sınıflandırılması ve örnek hastalıklar. 1

Tablo 2.1: Kanser adlandırmalarında kullanılan ön ekler ve anlamları. 13

Tablo 2.2: GEO veritabanından veri kümelerine erişim yöntemleri. 17

Tablo 3.1: Beş kanser türüne ait mikrodizi veri kümelerinin özellikleri. 19

Tablo 3.2: Tek yönlü ANOVA hesaplama sonuçları. 27

Tablo 3.3: Sıra Sayıları-Toplam testi için farklı yöntemler. 29

Tablo 3.4: Çerçeve boyutunun değişimi ile WRS testindeki kritik değerin değişimi. 30

Tablo 3.5: İstatistiksel bir hipotezin test edilmesinde olası durumlar. 31

Tablo 3.6: FWER değerinin kontrolünde kullanılan yöntemler. 32

Tablo 5.1: SYMPES yönteminin değişen çerçeve boyutlarına göre meme kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak). 61

Tablo 5.2: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 5\,, \alpha = 0.5,\ |\rho| \leq 0.1$) 62

Tablo 5.3: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 9\,, \alpha = 0.5, |\rho| \leq 0.1$). 63

Tablo 5.4: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 15\,, \alpha = 0.5,\ |\rho| \leq 0.1$). 64

Tablo 5.5: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 25\,, \alpha = 0.5, |\rho| \leq 0.1$). 65

Tablo 5.6: SYMPES yönteminin değişen çerçeve boyutlarına göre akciğer kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak). 74

Tablo 5.7: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 5\,, \alpha = 0.5,\ \rho < 0$). 75

Tablo 5.8: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 9\,, \alpha = 0.5,\ \rho < 0$). 75

Tablo 5.9: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 15\,, \alpha = 0.5,\ \rho < 0$). 76

Tablo 5.10: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 25\,, \alpha = 0.5,\ \rho < 0$). 76

Tablo 5.11: SYMPES yönteminin değişen çerçeve boyutlarına göre over kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak). 82

Tablo 5.12: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 5\,, \alpha = 0.5,\ \rho < 0$). 82

Tablo 5.13: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 9\,, \alpha = 0.5, \rho < 0$). 83

Tablo 5.14: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 15\,, \alpha = 0.5,\ \rho < 0$). 83

Tablo 5.15: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 25\,, \alpha = 0.5,\ \rho < 0$). 84

Tablo 5.16: SYMPES yönteminin değişen çerçeve boyutlarına göre prostat kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak). 90

Tablo 5.17: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 5\,, \alpha = 0.5,\ \rho < 0$). 90

Tablo 5.18: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 9\,, \alpha = 0.5, \rho < 0$). 91

Tablo 5.19: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 15\,, \alpha = 0.5,\ \rho < 0$). 91

Tablo 5.20: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 25\,, \alpha = 0.5,\ \rho < 0$). 92

Tablo 5.21: SYMPES yönteminin değişen çerçeve boyutlarına göre pankreas kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak). 98

Tablo 5.22: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 5\,, \alpha = 0.5,\ |\rho| \leq 0.1$). 99

Tablo 5.23: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 9\,, \alpha = 0.5,\ |\rho| \leq 0.1$). 100

Tablo 5.24: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 15\,, \alpha = 0.5,\ |\rho| \leq 0.1$). 101

Tablo 5.25: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($LF = 25\,, \alpha = 0.5,\ |\rho| \leq 0.1$). 102

Tablo 5.26: Meme kanseri veri kümesi için sınıflandırma sonuçları. 111

Tablo 5.27: Akciğer kanseri veri kümesi için sınıflandırma sonuçları.....112

Tablo 5.28: Over kanseri veri kümesi için sınıflandırma sonuçları.....113

Tablo 5.29: Prostat kanseri veri kümesi için sınıflandırma sonuçları. 114

Tablo 5.30: Pankreas kanseri veri kümesi için sınıflandırma sonuçları. 115

SİMGE VE KISALTMA LİSTESİ

Simgeler	Açıklama
A	: Minimum – maksimum normalizasyonun alt sınır değeri
a	: Kaydırma sabiti
B	: Minimum – maksimum normalizasyonun üst sınır değeri
$\boldsymbol{E\{\cdot\}}$	: Beklenen değer
$\boldsymbol{e_{i,j}^{(k)}}$	: k. ncı bireye ait j. nci çervenin i. nci elemanı
$\boldsymbol{e_{x_i}, e_{y_i}}$	: X ve Y'nin i. nci bileşeninin içerdiği bilgi
$\boldsymbol{G_j}$	: GF_j'nin kazanç faktörü
$\boldsymbol{GF_J^{(k)}}$	: k. ncı bireye ait j. nci gen çerçevesi
$\boldsymbol{H_0, H_1}$	: Yokluk ve alternatif hipotezleri (sırasıyla)
$\boldsymbol{I}$	: Birim matris
$\boldsymbol{h}$	: Kruskal – Wallis H testinde hesaplanan değer
$\boldsymbol{J_t}$	: Toplam kareler hatası
$\boldsymbol{k}$	: Örneklem sayısı
$\boldsymbol{k_i}$	: i. nci özniteliğin kat değişimi oranı (*İng.* fold-change ratio)
$\boldsymbol{\widetilde{k}_i}$	: i. nci özniteliğin kat değişimi oranının dönüştürülmüş değeri
$\boldsymbol{\max(\omega_{gn})}$	: Azami gen ifadesi
$\boldsymbol{\min(\omega_{gn})}$	: Asgari gen ifadesi
$\boldsymbol{L_F}$	: Çerçeve boyutu
$\boldsymbol{n_A, n_B}$	: A ve B örneklemlerinin eleman sayısı
$\boldsymbol{N_F}$	: Çerçeve sayısı
$\boldsymbol{N_{IE}}$	: Önceden tanımlı bilgi bloğunun toplam indis sayısı
$\boldsymbol{N_{IS}}$	: Önceden tanımlı enerji bloğunun toplam indis sayısı
$\boldsymbol{p}$	: Olasılık değeri
$\boldsymbol{R_X}$	: Özilinti matrisi
$\boldsymbol{S_m(A)}$	: A dönüşüm matrisinin ilk m bileşeninde toplanan enerji miktarı
$\boldsymbol{s_{A,i}^2, s_{B,i}^2}$	: A ve B örneklemlerinin i. nci özniteliklerinin varyansı
$\boldsymbol{SS_A}$	: İşlem kareler toplamı (tek yönlü ANOVA)
$\boldsymbol{SS_E}$	: Hata kareler toplamı (tek yönlü ANOVA)

SS_T : Toplam kareler toplamı (tek yönlü ANOVA)

t_i : i. nci özniteliğin t değeri (Welch t testi)

u_1, u_2 : 1. ve 2. gözlemdeki Wilcoxon sıra sayıları toplam değerlerinin Mann-Whitney U testi ile ilişkilendirilmiş toplam değerleri

$u(k,l), u(l,k)$: k ve l öznitelikleri arasındaki uzaklık

v : Serbestlik derecesi

$\overline{x}_{A,i}, \overline{x}_{B,i}$: i. nci öznitelik için A ve B örneklemlerinin ortalama değerleri

χ^2_α : Ki-kare dağılımı kritik değeri

$y_{gn,i}$: i.nci dönüştürülmüş gen ifade verisi

y_{ij} : i.nci gruptaki j. nci gözlem (tek yönlü ANOVA)

$\overline{y}_{..}$: Tüm gözlemlerin ortalaması (tek yönlü ANOVA)

$\overline{y}_{i.}$: i.nci gruptaki tüm gözlemlerin ortalaması (tek yönlü ANOVA)

α : Anlamlılık seviyesi (Tip 1 hata değeri olarak da adlandırılır)

α' : Düzeltilen anlamlılık seviyesi

ε_t : Yaklaşık hata

λ_i : i. nci özdeğer

Λ : Özdeğer matrisi

$\tilde{\mu}_1, \tilde{\mu}_2$: 1. ve 2. örneklemin ortanca (medyan) değeri

$\mu^{(k)}_{e_{i,j,c}}$: c. nci sınıfa ait evrensel ortalama bilgi kütüphanesinin k. nci bireye ait j. nci çerçevesinin i. nci ortalama değeri

$\tilde{\mu}^{(k)}_{e_{i,j,c}}$: c. nci sınıfa ait evrensel ortanca bilgi kütüphanesinin k. nci bireye ait j. nci çerçevesinin i. nci ortanca değeri

$\mu_{UE,c}$: c. nci sınıfa ait evrensel ortalama bilgi kütüphanesi

$\tilde{\mu}_{UE,c}$: c. nci sınıfa ait evrensel ortanca bilgi kütüphanesi

$\mu_{\omega_{gn}}$: Gen ifade verilerinin ortalama değeri

$\sigma^{2\,(k)}_{e_{i,j,c}}$: c. nci sınıfa ait evrensel ortalama bilgi kütüphanesinin k. nci bireye ait j. nci çerçevesinin i. nci varyans değeri

ρ : Pearson rho değeri

ρ_{XY} : X ve Y örneklemlerinin Pearson özilinti katsayısı

$\sigma_{\omega_{gn}}$: Gen ifade verilerinin standart sapma değeri

ϕ_i : i. nci özvektör

Φ : Özvektör matrisi

ω_1, ω_2 : 1. ve 2. gözlemdeki sıra sayılarının toplamları (Wilcoxon Sıra Sayıları Toplam testinde)

$(.)^{T}$: Bir matrisin transpozu

$(.)^{-1}$: Bir matrisin tersi

$\|\cdot\|$: Bir matrisin uzunluğu

Kısaltmalar	**Açıklama**
ANOVA	: Varyans analizi
PEB	: Önceden tanımlı bilgi bloğu
PESB	: Önceden tanımlı enerji ve bilgi blokları
PSB	: Önceden tanımlı enerji bloğu
DNA	: Deoksiribonükleik asit
DT	: Karar ağacı (*İng.* Decision Tree)
FDR	: Yanlış bulgu oranı (*İng.* false discovery rate)
FWER	: Ailesel hata oranı (*İng.* family-wise error rate)
GBM	: Gradyan arttırma makineleri (*İng.* Gradient Boosting Machines)
GEO	: Gen İfadesi Omnibus **(***İng.* Gene Expression Omnibus**)**
GF	: Gen çerçevesi (*İng.* gene frame)
KLD	: Karhunen-Loeve Dönüşümü
KNN	: k-en yakın komşu (*İng.* k-nearest neighbor)
LR	: Lojistik ilkelleme (*İng.* Logistic regression)
LMS	: En küçük ortalama kareler (*İng.* least mean squares)
MIAME	: Bir Mikrodizi Deneyi Hakkında Minimum Bilgi (*İng.* Minimum Information About a Microarray Experiment)
MLP	: Çok Katmanlı Algılayıcı (*İng.* Multi Layer Perceptron)
NB	: Sade Bayes (*İng*. Naive Bayes)
NCBI	: Ulusal Biyoteknoloji Bilgi Merkezi (*İng.* National Center for Biotechnology Information)
PCC	: Pearson Özilinti Katsayısı (*İng.* Pearson Correlation Coefficient)
RF	: Rasgele Orman (*İng.* Random Forest)
RNA	: Ribonükleik asit
SAM	: Mikrodizi önemlilik analizi (*İng.* Significance Analysis of Microarray)
$SF_{min\text{-}max}$	: SYMPES yöntemi ile asgari – azami kat değişimi oranının birleşiminden oluşan yeni yaklaşım
$SPCCF_{min\text{-}max}$	: SYMPES yöntemi ile Pearson özilinti katsayısı ve asgari – azami kat değişimi oranının birleşiminden oluşan yeni yaklaşım
SVM	: Destek vektör makineleri (*İng.* Support vector machines)

SWRS : SYMPES yöntemi ile Wilcoxon sıra sayıları – toplam testinin birleşiminden oluşan yeni yaklaşım

SWRS_cor : SYMPES yöntemi ile Wilcoxon sıra sayıları – toplam testinin birleşiminden oluşan yeni yaklaşım (FWER düzeltmesi yapan biçimi)

SWT : SYMPES yöntemi ile Welch – t testinin birleşiminden oluşan yeni yaklaşım

SWT_cor : SYMPES yöntemi ile Welch – t testinin birleşiminden oluşan yeni yaklaşım (FWER düzeltmesi yapan biçimi)

TBG : Tümör Baskılayan Gen

ÖZET

DOKTORA TEZİ

SYMPES YÖNTEMİYLE KANSER HASTALIKLARINDA ÖNEMLİ GENLERİN SAPTANMASI VE KANSER TÜRLERİNİN SINIFLANDIRILMASI

Ali SARIKAŞ

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Biyomedikal Mühendisliği Anabilim Dalı

Danışman : Prof. Dr. Sıddık YARMAN

II. Danışman : Dr. Zeliha GÖRMEZ

Mikrodizi teknolojisi, organizmaların genomik dizi bilgilerini kullanarak, sağlık ve hastalık durumlarının daha iyi anlaşılması, daha iyi teşhisler koyulması, yeni terapötik müdahaleler sunacak işlevsel biyolojik mekanizmalara dönüştürülmesi gibi görevleri yerine getirecek eşsiz niteliklere sahip olması sayesinde bilimsel araştırmaların önemli bir aracı haline gelmiştir. Mikrodiziler, giderek artan genomik bilgiyi etkili bir şekilde ölçeklendiren güvenilir, hızlı ve uygun maliyetli bir yöntem sağlarlar. Kanserlerin gelişiminde bir çok genin rol aldığı bilinmektedir. Mikrodizi teknolojileri ile aynı anda ve tek seferde tüm genomun analizinin gerçekleştirilmesinin mümkün olmasıyla, sağlıklı ve hasta bireylerin gen ifade profilleri çıkarılarak hastalık durumunu etkileyen genlerin belirlenmesinde önemli bir adım atılmıştır.

Bu tezde, Affymetrix gen dizileme cihazından elde edilen 54.675 adet proba karşılık gelen aynı sayıda gen dizilimi SYMPES yöntemi ile modellenmiştir. Geliştirilen SYMPES modeli kullanılarak 5 farklı kanser türünün gelişmesinde etkili olduğu varsayılan farksal olarak ifade edilen genlerin (*İng.* differentially expressed genes) tespiti yapılmıştır.

Affymetrix gen çipi kişilerden temin edilen ve kişilere ilişkin gen ihtiva eden sıvı örnekler üzerinde çalışır. Her bir gen, adına hibridizasyon dediğimiz bir prosese tabi tutularak parlaklık düzeyine göre cihaz tarafından tanımlanır. Başka bir deyişle, art arda cihaz tarafından dizilen genlerin yaklaşık 55.000 gene ilişkin parlaklık düzeyleri bir çıktı olarak cihaz kullanıcısına

sunulur. Buna göre, kullanıcı cihaz tarafından dizilimi sabitlenen genleri teker teker $N = 1,2,\ldots,54675$'e kadar örneklenmiş olduğunu kabul eder ve her gene ilişkin parlaklık düzeyi kayıt altına alınır. Böylesi bir yaklaşımda, gen dizisindeki genler n ile ifade edilirken, bunlara karşı gelen parlaklık düzeyleri $X(n)$ ile ifade edilir. Böylece $X(n) = f(n)$ gibi ayrık zaman domeninde ifade edilen bir fonksiyon olarak ifade edilmiş olur. Böylece SYMPES modelleme yöntemi doğrudan $X(n) = f(n)$ üzerinde uygulanır.

Literatüre kazandırılmış önceki çalışmalar incelendiğinde, SYMPES yönteminin biyomedikal alanındaki biyopotansiyel işaretler (EMG, EKG, EEG), konuşma işaretleri, tıbbi cihazlardan alınan görüntüler gibi tek ve iki boyutlu işaretlerin modellenmesinde kullanıldığı görülmüştür.

Bu tezde, SYMPES yöntemi Affymetrix gen dizileme cihazından alınan gen dizilerini modellemede kullanılmadan önce, gen ifade verileri 0 ila 1 arasında kalacak şekilde normalize edilir. Ayrık değerler niteliğinde kaydedilmiş $N \times 1$ boyutlu sayısal değerler içeren gen ifade veri kümesi, $GF(n)$ olarak verilsin. Bu durumda, SYMPES yöntemine göre, her biri eşit L_F uzunluğunda çerçevelere bölünen işaretin her bir j. çerçevesi için $GF_j \cong G_j E_{IE} S_{IS}$, $(j = 1,2,\ldots,N_F)$ şeklinde ifade edilebilir. Burada, G_j çerçeve kazancını ifade eden reel bir sabit; $S_{IS}^T = \left[S_{IS1}\ S_{IS2}\ \ldots S_{ISL_F}\right]$ işaretin istatistiksel özelliklerini yansıtan ve en küçük ortalama karesel hata (LMS) anlamında $G_j S_{IS}$ ifadesi ile çerçevenin neredeyse tüm enerjisini içeren bir imza (veya enerji) vektörü; E_{IE} ise $L_F \times L_F$ boyutunda köşegen matris olup $G_j S_{IS}$ ifadesinin bir zarf (veya bilgi) vektörüdür. Önceden tanımlı bilgi ve enerji vektörleri sınıflandırma çalışmalarında etkin olarak kullanılmıştır. $L_F = \{5,9,15,25\}$ olarak yapılandırılan çerçeve uzunlukları için SYMPES yönteminin sınıflandırma başarımları irdelenmiş, buna ek olarak k-en yakın komşu (*İng.* K-nearest neighbors, KNN), destek vektör makineleri (*İng.* Support vector machines, SVM), rasgele orman (*İng.* Random forest, RF) ve sade bayes (*İng.* Naive Bayes, NB) sınıflandırma yöntemleri ile bu yöntemin karşılaştırmaları da yapılmıştır. Bunun yanı sıra, her bir grup için evrensel ortalama ve ortanca bilgi kütüphaneleri de farksal olarak ifade edilen genlerin tespit edilmesinde kullanılmak üzere oluşturulmuştur. SYMPES yönteminin farksal olarak ifade edilen genleri tespit edebilmesindeki başarımını arttırmak üzere, evrensel bilgi kütüphaneleri Wilcoxon sıra sayıları-toplam testi, Welch t-testi, kat değişimi (*İng.* Fold change), Pearson özilinti katsayısı (*İng.* Pearson correlation coefficient) ve ailesel ortak hata oranı (İng. Family-wise errror rate, FWER) düzeltme yöntemleri kullanılmak suretiyle 6 farklı yaklaşım önerilmiştir. Bunların içerisinde, $SPCCF_{min\text{-}max}$ yöntemi veri kümelerinin hemen hepsinde tüm ikili gruplar arasında %100'e varan oranda önemli genleri tespit edebilmiştir. SYMPES yöntemi, bu yaklaşımların her biri için yeni önemli gen listesi oluşturur ve GEO2R platformu ile elde edilen referans gen listesini baz alarak ortak genlerin sayısını ortaya çıkarır.

Tezde elde edilen sonuçlara göre, SYMPES yönteminin kanser hastalıklarının türlerinin tespitinde ve bu hastalıklara ilişkin etkin genlerin belirlenmesinde yüksek başarı oranı ile kullanılabileceği gösterilmiştir.

Temmuz 2020, 150 sayfa.

Anahtar kelimeler: Mikrodizi gen ifade verisi, farksal olarak ifade edilen gen analizi, kanser sınıflandırma, önceden tanımlı enerji ve bilgi blokları, evrensel bilgi kütüphaneleri.

SUMMARY

Ph.D. THESIS

DETECTION OF SIGNIFICANT GENES IN CANCER DISEASES AND CLASSIFICATION OF CANCER TYPES BY SYMPES METHOD

Ali SARIKAŞ

Istanbul University-Cerrahpasa

Institute of Graduate Studies

Department of Biomedical Engineering

Supervisor : Prof. Dr. Sıddık YARMAN

Co-Supervisor : Dr. Zeliha GÖRMEZ

Microarray technology has become an important tool of scientific research, using the genomic sequence information of organisms, having unique qualities to perform tasks such as better understanding of health and disease conditions, better diagnoses, and transformation into functional biological mechanisms that will provide new therapeutic interventions. Microarrays provide a reliable, fast and cost-effective method that effectively scales increasing genomic information. It is known that many genes play a role in the development of cancers. With the ability to perform the analysis of the entire genome at once with microarray technologies, an important step was taken to identify the genes affect the disease state by revealing the gene expression profiles of healthy and unhealthy ones.

In this thesis, the same number of gene sequences corresponding to 54,675 probes obtained from the Affymetrix gene sequencing device were modeled with the SYMPES method. By using the developed SYMPES model, differentially expressed genes, which are assumed to be effective in the development of 5 different types of cancer, were detected.

The Affymetrix GeneChip works on liquid samples that are available from individuals and contain gene related to individuals. Each gene is identified by the device according to the

brightness level by subjected to a process called hybridization. In other words, the brightness levels of about 55,000 genes of the genes sequenced by the device are presented to the user of the device as an output. Accordingly, the user assumes that the genes whose sequences are fixed by the device are sampled one by one until $N = 1,2,\ldots,54675$ and the brightness level for each gene is recorded. In such an approach, the genes in the gene sequence are expressed by n, while their corresponding brightness levels are expressed by $X(n)$. Thus, it is expressed as a function in a discrete time domain such as $X(n) = f(n)$. Thus, the SYMPES modeling method is applied directly on $X(n) = f(n)$.

When the previous studies in the literature were examined, it was seen that SYMPES method was used in the modeling of one and two-dimensional signals such as biopotential signals (EMG, ECG, EEG), speech signals and images captured from medical devices.

In this thesis, before the SYMPES method is used to model gene sequences from the Affymetrix gene sequencing device, gene expression data is normalized to remain between 0 and 1. The gene expression dataset containing $N \times 1$ dimensional numerical values recorded as discrete values is given as $GF(n)$. In this case, according to the SYMPES method, each frame of the signal divided into frames of equal L_F length can be expressed as $GF_j \cong G_j E_{IE} S_{IS}$ $, (j = 1,2,\ldots,N_F)$. Here, G_j is a real constant that expresses the frame gain; $S_{IS}^T = [S_{IS1}\ S_{IS2} \ldots S_{ISL_F}]$ expression is a signature (or energy) vector that reflects the statistical properties of the signal and contains almost all the energy of the frame with the expression $G_j S_{IS}$ in the sense of the least mean square error (LMS); E_{IE} is a diagonal matrix of size $L_F \times L_F$ which is an envelope (or information) vector of the $G_j S_{IS}$ expression. Predefined information and energy vectors have been used effectively in classification studies. Classification performances of SYMPES method for frame lengths configured as $L_F = \{5,9,15,25\}$ are examined, in addition k-nearest neighbor (KNN), support vector machines (SVM), random forest (RF) and Naive Bayes (NB) classification methods were also compared with this method. In addition, universal mean and median information libraries for each group were created to be used to detect genes that are expressed differentially. To increase the performance of the SYMPES method in detecting differentially expressed genes, a total of 6 new methods have been proposed using Wilcoxon rank-sum test, Welch's t-test, fold-change, Pearson correlation coefficient, and family-wise error rate correction. Among these, the $SPCCF_{min\text{-}max}$ approach was able to detect up to 100% important genes among all binary groups in almost all datasets. The SYMPES method creates a new significant gene list for each of these approaches and reveals the number of common genes based on the reference gene list obtained by the GEO2R platform.

According to the results obtained in the thesis, it has been shown that SYMPES method can be used to classify the types of cancer diseases and to determine the significant genes related to these diseases with high success rate.

July 2020, 150 pages.

Keywords: Microarray gene expression data, differentially expressed gene analysis, cancer classification, predefined envelope and signature blocks, universal information libraries.

1. GİRİŞ

İnsan Genom Projesi'nin 2000'li yılların başında tamamlanması ve beraberinde geliştirilen teknolojiler genetik hastalıkların tanımında yeni bir dönem açmıştır. Bugün gelinen noktada, hemen hemen her hastalığın gen ve çevrenin birlikte etki etmesi sonucu ortaya çıktığı, ancak genetik bileşenin göreceli rolünün büyük ya da küçük olabildiği kanısı hakimdir. Kısmen ya da tamamen genetik faktörler nedeniyle görülen hastalıklar kromozomal hastalıklar, tek-gen hastalıkları (Mendelyen) ve multifaktöriyel hastalıklar olmak üzere üç kategoriye ayrılır. Kromozomal hastalıklarda, bozukluk (veya defekt) genomdaki tek bir hatadan değil, tüm bir kromozomda ya da bir kromozomal bölgede yer alan genlerin eksikliğine ya da fazlalığına bağlıdır. Tek gen hastalıkları, tek bir gendeki patojenik mutasyonlar sonucu görülen hastalıklardır. Tek gen hastalıkları genellikle ailelerde klasik kalıtım kalıplarından biri ile kalıtılır (otozomal resesif/çekinik, otozomal dominant/baskın, X'e bağlı kalıtım). Multifaktöriyel hastalıklar tüm popülasyonun çok daha büyük bir kısmını etkisi altına alan (%60'tan fazla), tek gen hastalıklarında tipik olarak gözlenen kalıtım kalıplarına uymayan, farklı genlerin etkisinin birleşimi sonucu görülen hastalıklardır. Multifaktöriyel hastalıklara insan kompleks hastalıkları ya da çoklu genetik hastalıklar adı da verilmektedir. Multifaktöriyel hastalıklar çevresel faktörler, yaşam biçimi, çoklu genetik gibi faktörlerle etkileşim içerisinde olup, bu tür hastalıkların kalıtımına kompleks kalıtım adı verilir. Mendelyen hastalıklarda tek genin büyük katkısı söz konusu iken, multifaktöriyel hastalıklarda bir çok genin – küçük oranlarda olsa bile – katkısı söz konusudur. Tablo 1.1'de genetik hastalık türlerine ait bilinen hastalıklar örnek olarak verilmiştir [1].

Tablo 1.1: Genetik hastalıkların sınıflandırılması ve örnek hastalıklar.

Kromozomal Hastalıklar	**Tek Gen Hastalıkları**	**Multifaktöriyel Hastalıklar**
Trizomi 13	Kistik fibrozis	Şizofreni
Trizomi 18	Orak hücreli anemi	Otizm
Trizomi 21 (Down Sendromu)	Marfan Sendromu	Manik-depresif (bipolar) hastalık
Williams Sendromu	Polikistik böbrek hastalığı	Tip 1 diabetes mellitus
Prader-Willi Sendromu	Ailesel hiperkolesterolemi	Crohn hastalığı
DiGeorge Sendromu	Miyotonik distrofi	Multipl sklerozis
Angelman Sendromu	Huntington hastalığı	Romatoid artrit
Klinefelter Sendromu	Akondroplazi	Travma dışı epilepsi
Turner Sendromu	Hemofili-A	Koroner arter hastalığı
Frajil X Sendromu	Rett Sendromu	Alzheimer hastalığı

Klinik tıptaki en yaygın ve ciddi hastalıklardan biri de kanserdir. Giderek bir kitleye veya tümöre (neoplazm) dönüşecek şekilde kontrolsüz hücre büyümesi sürecinin (neoplazi) daha virulan formu olan kanserin, köken aldığı bölgeyi çevreleyen komşu dokuları işgal edebilme ve daha uzak bölgelere yayılabilme (metastaz) yeteneği vardır. Hem kansere kalıtsal yatkınlığı olanlar için, hem de genel toplum için kanserin erken tanısı ve tedavisi hayati önem taşımaktadır ve bunların her ikisi de genom dizilemesi ve gen ifadesi incelemelerindeki ilerlemelere giderek daha fazla bağımlı olmaktadır. Birçok genin rol alabildiği kanser gelişiminin incelenmesinde, gen ifade profilinin çıkarılması amacıyla etkin kullanılan bir araç olan mikrodizi teknikleri, hücreleri ve kromozomları yerinde tek seferde incelemek yerine, aynı anda örtüşen veya düzenli aralıklarla yer alan deoksiribonükleik asit (DNA) parçaları içeren, mikroskop slayt üzerindeki düzenli sıralı genomik segmentler olarak temsil edilen tüm genomun analiz edilmesine olanak sağlamaktadır. Buradan hareketle, sağlıklı ve hasta bireylerin mikrodizi gen ifade profilleri çıkarılarak hastalık durumunda etkilenen genlerin belirlenmesinde önemli bir adım atılmış olur.

1.1.LİTERATÜR ÖZETİ

Tez çalışmaları kapsamında mikrodizi veri kümeleri üzerinde farksal olarak gen ifade analizi ve sınıflandırma çalışmaları yapılmıştır. Bu doğrultuda literatürde yer alan, önceden yapılan çalışmalar incelenmiş ve bu bölümde aktarılmıştır.

Mutch ve ark. sınırlı kat değişimi modelini kullanarak mikrodizi veri kümesinde farksal olarak ifade edilen genlerin seçilmesi üzerine çalışma yapmışlardır. Mevcut gen seçimi modelini geliştirmek için, aykırı değerlerin seçilmesinin geçerliliği, en üstteki %5'i bir karşılaştırma ölçütü olarak kullanan çok çeşitli genler için değerlendirmişler, model eğilimlerini %1'den %10'a kadar incelemişlerdir. Geliştirdikleri modelin %85.7 oranında başarım sağladığını belirtmişlerdir [2].

Dalman ve ark. anlamlılık düzeyini ifade eden p-değerleri (kısaca, olasılıksal sayısal bir değer olup bir olayın gerçekleşmesi ihtimalini ifade etmektedir.) ile keyfi kat değişimi parametrelerini kullanmanın mikrodizi veri kümelerinin yorumlanmasını değiştirdiğini yaptıkları çalışma ile göstermişlerdir. Buna rağmen kullandıkları veri kümesinde keyfi kat değişimi oranının 2'den büyük (logaritma 2 tabanında bu değer, karşılaştırılan örnekler arasındaki oranın 1 e eşit olması anlamına gelmektedir.) ve p-değerinin 0.002'den küçük olarak belirlenen ölçütleri temel alındığında farksal olarak ifade edilen genlerin gözlemlenebildiğni belirtmişlerdir [3].

Butte ve ark. logaritmik kat değişimi katsayıları hesaplayarak, bu ölçütleri farksal olarak ifade edilen genlerin tayinini yapmak için kullanmışlardır [4].

Pavlidis farksal olarak ifade edilen genlerin tespit edilmesinde klasik istatistiksel yöntem olan varyans analizinin (ANOVA) kat-değişimi yöntemine göre daha etkili ve basit bir alternatif olduğunu ileri sürmüştür. Anlamlılık düzeyi olarak kullanılan p-değerlerini ve ardından yanlış bulgu keşif oranını düzeltme yöntemini kullanarak anlamlı derecede farklılık gösteren gerçek pozitif genleri tespit etmiştir [5].

Mikrodizi veri kümelerinin normal dağılımlı olmamasından dolayı parametrik olmayan testlerin de kullanımı söz konusu olmuştur. Zhao ve Pan, mikrodizi deneylerinde farksal olarak ifade edilen genlerin tespit edilmesinde geliştirdikleri önceki yöntemi iki farklı şekilde modifiye ederek hipotez testlerinde Tip-1 hatanın azaltılmasına yönelik çalışma yapmışlardır. Sonuç olarak, z-değerini esas alan iki farklı reddetme bölgesi belirleyerek farksal olarak ifade edilen genleri tespit etmişlerdir [6].

Chen ve ark. mikrodizi veri analizinde farksal olarak ifade edilen genlerin seçiminde kullanılan parametrik ve parametrik olmayan yöntemlerden t-testi, SAM testi[1], U- testi ve M testini[2] belirledikleri eşik p değerini esas alarak değerlendirmişlerdir. Bunun yanında tip 1 hatanın azaltılmasına yönelik olarak çoklu karşılaştırma testlerinin de uygulamasını yapmışlardır [7].

Troyanskaya ve ark. parametrik olmayan t-testi, Wilcoxon sıra sayıları – toplam testi ve yüksek Pearson özilinti katsayısına dayanan sezgisel bir yöntem olmak üzere, bu üç yöntemin mikrodizi verilerinde farksal olarak ifade edilen genlerin keşfi için uygulamasını yapmışlardır. Sıra sayıları toplam testinin üzerinde çalıştıkları tüm veri kümelerinde daha iyi sonuçlar verdiğini ama diğer yöntemlerin de oldukça düşük sahte pozitif oranı vererek başarımlarının iyi olduğunu vurgulamışlardır [8].

Parrish ve Spencer III normalizasyonun oligonükleotid mikrodizi verilerinde etkisini inceledikleri çalışmada, kantil normalizasyon yönteminin medyan-IQR normalizasyonu yöntemine göre daha etkili sonuçlar doğurduğu belirtilmiştir [9].

[1] EK-A'da verilmiştir.
[2] EK-A'da verilmiştir.

Smyth deneysel Bayes yaklaşımını kullanarak ve doğrusal modeller geliştirerek "limma" adını verdiği, günümüzde halen yaygın kullanımı olan ve oldukça güvenilir bir farksal olarak gen ifade analizi yöntemi ortaya koymuştur. Bu yöntem Benjamini – Hochberg'in yanlış bulgu oranı (*İng.* false discovery rate, FDR) düzeltmesinin yanında, ilerleyen bölümlerde de bahsedilen, FWER düzeltme yöntemlerinden bazılarını kullanıcı isteğine göre yapabilmektedir [10].

Ritchie ve ark. "limma" yönteminin ribonükleik asit-dizileme (RNA-Seq) ve mikrodizi deneyleri için farksal olarak ifade analizini güçlendirdiğini ileri sürmüşlerdir. Deneysel Bayes ve doğrusal modellerin yanında, çoklu karşılaştırma testlerini de kullanarak minimum sahte pozitif oranını sağlayacak yeni yöntemleri "limma" yöntemini güncelleyerek ortaya çıkarmışlar ve oldukça başarılı sonuçlar elde ederek diğer yöntemlere üstün gelmişlerdir [11].

Rapaport ve ark. RNA-Seq verisi üzerinde Cuffdiff, DESeq, edgeR, limmaQN, limmaVoom, PoissonSeq ve baySeq farksal olarak gen ifade analiz yöntemlerinin başarımlarını karşılaştırmışlardır. Her birinin iyi sonuçlar vermesine karşılık limmaQN yönteminin daha iyi sonuçlar verdiği gözlemlenmiştir [12].

Robinson ve ark. mikrodizi veri kümelerinin negatif binom dağılımına sahip olma potansiyelinden dolayı yine deneysel Bayes yaklaşımını kullanan edgeR adını verdikleri yeni bir farksal olarak gen ifade analizi yöntemi önermişlerdir [13].

Statnikov ve ark. mikrodizi kullanılan kanser çalışmalarında SVM ve RF yöntemlerinin sınıflandırma başarımlarını gözlemlemişlerdir. SVM yönteminin RF yöntemine üstün geldiğini ikili ve çoklu sınıftan oluşan çok sayıda veri kümesi üzerinde test ederek göstermişlerdir [14].

Mei ve ark. oldukça güçlü sınıflandırma yöntemlerinden ikisi olan KNN ve SVM yönteminin birleşiminden hibrit bir yöntem ortaya çıkarmışlardır. İki veri kümesi üzerinde yaptıkları çalışmalarda bu hibrit yöntemin KNN ve SVM yöntemlerine göre daha iyi sonuçlar alındığını göstermişlerdir [15].

Turgut ve ark. meme kanseri için mikrodizi verisi üzerinde yaptıkları çalışmada SVM, KNN, MLP, DT, RF, LR, Adaboost ve GBM yöntemlerini sınıflandırma başarımlarını ölçmek için kullanmışlardır. Özellik seçimi yöntemini uygulamadan önce RF yöntemi en başarılı olarak

bulunmasına karşın, özellik seçimi yöntemi uygulandıktan sonra SVM yöntemi diğerleri arasında en başarılı yöntem olarak bulunmuştur [16].

Aydadenta'nın iki sınıflı beş kenser mikrodizi verisi için SVM, ANN, NB, KNN, C4.5 ve RF yöntemlerini karşılaştırdığı çalışmasında RF'nin diğer yöntemlere üstün geldiğini göstermiştir [17].

Wang ve Simon, 11 veri kümesi üzerinde geliştirdikleri tek gen sınıflayıcıları modellerini (SGC-t, SGC-WMW) NB, KNN, SVM ve RF yöntemleri ile kıyaslamış, bir çok veri kümesinde yöntemlerinin genel olarak kıyaslanan bu 4 yöntem kadar iyi sonuçlar verdiğini göstermişlerdir [18].

Bu tezde, Prof. Dr. Sıddık YARMAN tarafından geliştirilen ve literatüre kazandırılan SYMPES (*İng.* Systematic Procedure for Predefined Envelope and Signature Sequences) adını verdiği yeni bir sistematik işaret modelleme yöntemi kullanılmıştır. Bu yöntem, özgün kodlama ve kod çözme algoritmalarını barındırmakta, bu algoritmalar ile sürekli ve ayrık işaretleri temel tanım dizileri olarak da adlandırılan özgün fonksiyonlar (önceden tanımlanmış zarf ve imza dizileri) olarak yeniden tanımlamaktadır. SYMPES yöntemi ile bir işaretin çerçeve kazancı, zarf (veya bilgi) fonksiyonu ve imza (veya enerji) fonksiyonu olmak üzere üç bileşen ile ifade edilebileceği gösterilmiştir.

Günümüz bilgi çağında, birçok hastalığın tanı ve tedavisinde yaygın kullanılan biyomedikal işaretlerin işlenmesi, kaydedilmesi ve iletilmesi oldukça önem kazanmıştır. Bazı işaretlerin büyük veri içermesi, işaret işleme, kaydetme, iletme ve yeniden oluşturma gibi işlemlerin belirli kısıtlamalarla karşılaşmasına sebep olabilmektedir. Bu probleme çözüm getiren bir yaklaşım veri sıkıştırma işlemidir. Sıkıştırılan işaretlerde kabul edilebilir bir düzeyde kayıp kaçınılmazdır. SYMPES yöntemi veriyi çok az kayıpla sıkıştırmakta, veri iletimini yaparken işareti değil işaretin belirli parametrelerini göndererek kendiliğinden güvenli bir haberleşme sistemi ortaya koymaktadır. SYMPES yöntemi kullanılarak bir çok işaretin modellemesi yapılmış, literatüre kazandırılmıştır. Bunlardan ilki, elektrokardiyogram (EKG) işaretinin modellenmesi olmuştur. Yarman ve ark. biyopotansiyel işaretlerden biri olan EKG işaretini SYMPES yöntemini kullanarak önceden tanımlı kişiselleştirilmiş bilgi ve enerji fonksiyonları ile modellemişlerdir. Sıkıştırılan EKG verilerini alt çerçeve boyutunu $L_F = 16$ ve sıkıştırma

oranını 8:1 olarak yapılandırarak orijinal veriyi 0.001'den küçük hata ile yeniden oluşturabilmişlerdir [19].

Yine EKG verilerinin modellenmesi için yapılan bir çalışmada [20], araştırmacılar farklı alt çerçeve boyutları için iletilecek verinin bit uzunluğu, sıkıştırma oranı (*İng.* compression ratio, CR), yüzdelik olarak sıkıştırma oranı ve yüzdelik olarak ortalama kare farkının kökü (*İng.* percent root mean square difference, PRD) sonuçlarını sunmuşlardır. Optimum çerçeve uzunluğu $L_F = 16$, CR ve ortalama PRD değerleri karşılaştırılarak belirlenmiştir. Bu durumda, önerilen yöntem PRD %3.5 ile 9.6:1 (%90) sıkıştırma oranına ulaşmıştır. Böylelikle yeni sıkıştırma yöntemi ile, yeniden yapılandırılmış EKG sinyalindeki teşhis bilgileri korunurken, düşük seviyeli geri çatma hatası ile iyi sıkıştırma oranları elde edilmiştir.

SYMPES yöntemi ile konuşma işaretlerinin modellenmesi üzerine de çok sayıda çalışma yapılmıştır. Güz ve ark. SYMPES yöntemi ile geliştirdikleri yeni konuşma modelini ticari olarak piyasada bulunan yöntemlerle karşılaştırmışlardır. Aynı sıkıştırma oranı veya daha iyisi için, SYMPES'in 16Kbps'de G.726 ve 2.4Kbps sesli uyarılmış LPC-IOE gibi kodlayıcılar üzerinde önemli ölçüde daha iyi işitme kalitesi sağladığı gösterilmiştir [21], [22].

Güz ve ark. konuşma işaretlerinin modellenmesi üzerine yaptıkları bir başka çalışmada, yeni, sağlam ve değişken uzunluklu çerçeve tabanlı konuşma modelleme yöntemini tanıtmışlardır. Bu yöntem sırasıyla gürültü giderme algoritması, kodlama ve kod çözme algoritmalarını içeren üç ana adımdan oluşmaktadır. Kodlama ve kod çözme parçaları, SYMPES yöntemi temel alınarak geliştirilmiştir. Bu setler, konuşma sinyallerinin algısal kalitesini korurken daha verimli kodlama stratejisi ve daha yüksek sıkıştırma oranı elde etmek için konuşma sinyallerinin sesli ve sessiz bölümlerini ayrı ayrı temsil eden iki tipte geliştirilmiştir. Önceki çalışmalarının [21], [22], [23] bir uzantısı olarak, sadece temiz konuşma sinyallerinin kodlanmasını değil, aynı zamanda gürültülü konuşma sinyallerinin de kodlanmasını dikkate almaktadırlar. Gürültüye dayanıklı yeni modül, gürültüyü bastırmakta ve yeni tasarlanan modelleme parçasına temiz konuşma sinyalini göndermektedir. Modelleme kısmı, sesli ve sessiz çerçeveleri dikkate alarak daha uygun önceden tanımlanmış dizi tiplerine geçerek daha yüksek sıkıştırma oranları vaat etmektedir [24].

Konuşma işaretlerinin modellenmesi üzerine yaptıkları bir başka çalışmada Yarman ve ark., kodlanmış IP tabanlı bir güvenli sesli iletişim sistemi kurmuşlardır [25].

Gürkan ve ark., elektroensefalogram (EEG) işaretlerinin modellenmesinde SYMPES yöntemini kullanmış, önerdikleri yöntemin, yeniden yapılandırılmış EEG sinyalindeki teşhis bilgilerini korurken, düşük seviyeli yeniden yapılandırma hatası ile iyi sıkıştırma oranları elde ettiğini göstermişlerdir [26].

Gökbay ve ark. sınıflandırılmış enerji ve desen bloklarına (*İng.* classified energy and pattern blocks, CEPB) dayalı yeni bir biyomedikal görüntü sıkıştırma yöntemi önermiştir. CEPB tabanlı sıkıştırma yöntemi bilgisayarlı tomografi (BT) görüntülerine özel olarak uygulanmış ve değerlendirme sonuçları sunulmuştur. Esasen, CEPB, kodlama ve kod çözme işlemlerini uygulamak için bir iletişim sisteminin hem vericisi hem de alıcı kısmında bulunan benzersiz bir şekilde tasarlanmış ve yapılandırılmış kod kütüphanesidir. Kodlama parametreleri, blok ölçeklendirme katsayısı (*İng.* block scaling coefficient, BSC) ve CEPB'ye dayalı olarak girdi görüntülerinin her bloğu için belirlenen enerji (IE) ve desen (IP) bloklarının indis değerleridir. Değerlendirme sonuçları yeni önerilen yöntemin önemli görüntü sıkıştırma oranları ve görüntü kalitesi sağladığını göstermektedir [27], [28].

1.2.TEZİN AMACI VE KATKILARI

Tezin amaçları, (a) gen ifade profillerinin çıkarılmasında yaygın kullanımı olan mikrodizi teknolojisinin ilgili kanser türleri ile ilgili verilerinden yararlanarak, bu alana ilk defa uygulanan SYMPES yöntemi ile kanserlerin gelişimini etkileyen çok sayıdaki farksal olarak ifade edilen genin tespitini yapmaktır ve (b) sağlıklı ve kanserli bireylerin verilerini içeren mikrodizi verilerinde kanser türlerinin sınıflandırmasını yapmak, bunun yanında literatürde yaygın kullanımı olan sınıflandırma yöntemleri ile tez kapsamında yeni geliştirilen önceden tanımlı enerji ve bilgi bloklarına dayalı yöntemin başarım yönünden karşılaştırmasını yapmaktır. Sonuçlar göstermiştir ki, yeni yöntem kanser türlerini sınıflandırmada yüksek başarım elde etmiş ve birçok yönteme üstün gelebilmiştir.

Sonuç olarak, bireylerden alınan DNA örnekleri ile – varsa – hastalığının tanısı, kromozom anomalileri, vb. gibi tanıları önceden koyarak hem bireyi hem de sağlık kuruluşunu gereksiz biyopsi alma, tıbbi görüntüleme cihazları ile tetkik etme, vb. gibi yüksek maliyetli işlemlerden koruyarak bir anlamda gereksiz harcanan zamandan ve mali yükten kurtarmak bu çalışmaların yapılma sebebi olarak görülmelidir. Ayrıca kanser hastalığının prognozu bu yöntemle takip edilebilecektir.

2. GENEL KISIMLAR

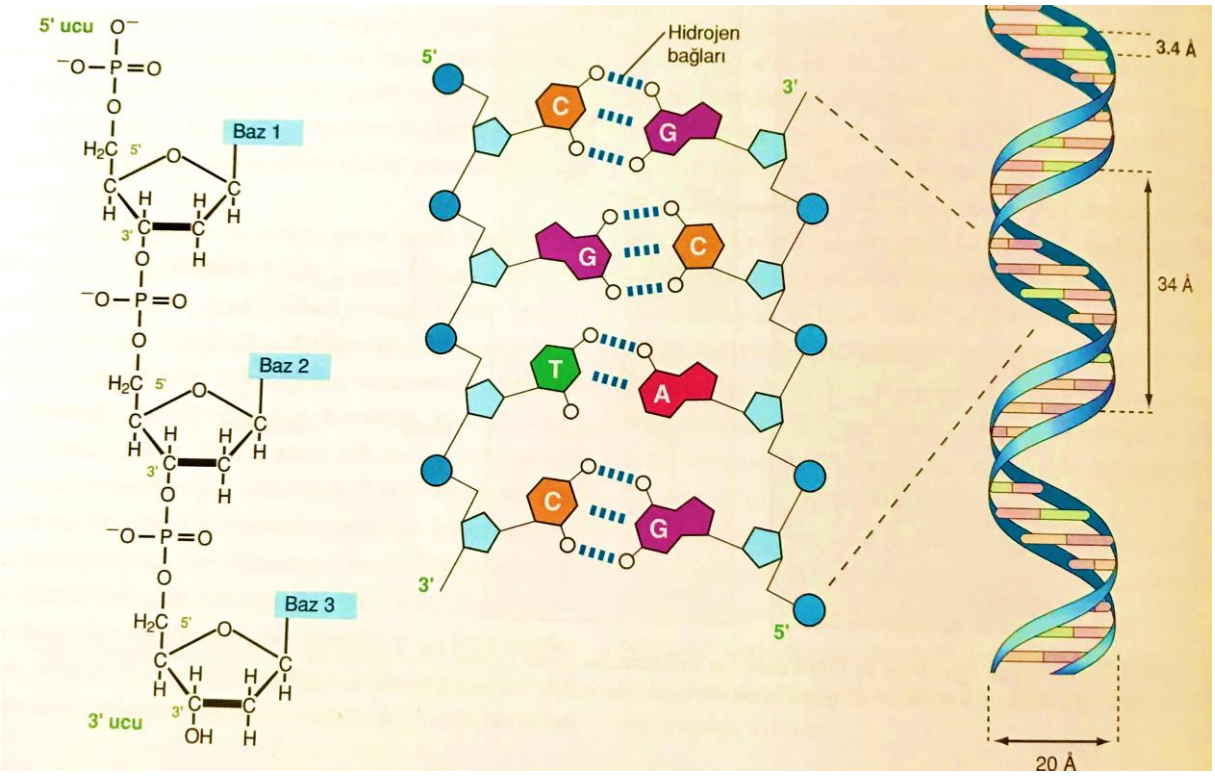
2.1.İNSAN GENOMU, GEN YAPISI VE İŞLEVİ

İnsan genomu, yapısında embriyogenez, gelişim, büyüme, metabolizma ve üreme gibi insanı işlevsel bir organizma yapan şeylerin bütün yönlerini belirlemek için gerekli genetik bilgiyi taşıyan oldukça büyük miktarda kimyasal DNA içerir. Vücuttaki tüm çekirdekli hücreler, yaklaşık 20.000 ila 50.000 arasında gen içeren, insan genomunun kendine ait bir kopyasını taşır. Genler, en basit tanımıyla, genetik bilginin işlevsel birimidir. Gen kelimesi ilk olarak 1908 yılında ortaya atılmış ve yaklaşık 150 yıl önce kalıtılabilir "birim karakterler" kavramının temel özelliklerinin botanikçi bilim insanı ve aynı zamanda rahip olan Gregor Mendel tarafından tanımlanmasından bu yana birçok farklı bağlamda kullanılmıştır. Mendel ve diğer erken dönem genetikçilere göre gen, organizma üzerinde görülür etkisine ve nesilden nesile istatistiksel olarak olarak belirli aktarımına bağlı tanımlanabilirken; tıbbi genetikçilere göre, klinik olarak gözlenebilen, karakteristik bir klinik bozukluğa sebep olan gözlemlenebilir bir varyant olarak kabul edilir. İnsan Genom Projesi insan genlerini tanımlamak adına sadece klinik yaklaşım ve aile çalışmalarındansa DNA dizi analizine dayalı, daha sistematik bir temel sağlamış oldu ancak 2003 yılında tamamlanan dizileme ürününe rağmen, bir dizinin bir genin varlığına veya kimliğine işaret eden özelliklerini ayırt etme kabiliyetimizin oldukça zayıf olduğu ortaya çıkmıştır. İnsan genom dizisinin ortaya çıkarılması ve varyasyonunun hastalık ve sağlık durumlarında insan biyolojisi ile ilişkilendirilmesi biyomedikal araştırma alanında halen zorlu bir iş olmaya devam etmektedir. İnsan genomunun nihai kataloğuna ulaşılması zor bir hedef olmasına karşın, genel olarak iki farklı gen tipi ortaya çıkmıştır: Ürünü protein olanlar ve ürünü işlevsel ribonükleik asit (RNA) olanlar.

Genler, genomda DNA içerisinde kodlanır ve her hücrenin çekirdeğinde kromozom adı verilen çubuk şeklindeki organeller olarak organize olur. Genin ve genetiğin sağlık ve hastalık üzerindeki etkisi oldukça derindir ve kökeni DNA içerisinde kodlanan bilgide saklıdır. Genler kromozom içerisinde doğrusal sıradadır ve her bir genin kesin bir bölgesi veya lokusu bulunur [1].

2.2.GEN İFADESİNİN TEMELLERİ VE SANTRAL DOGMA

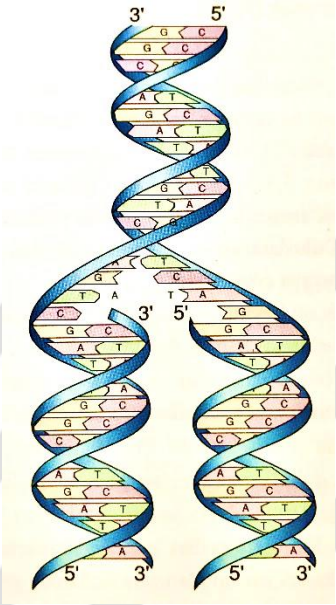
Polimerik nükleik asit makromolekül olan DNA, üç üniteden meydana gelmektedir: beş-karbon şekeri (deoksiriboz); azot içeren baz ve bir fosfat grubu. Bu azotlu bazlar pürinler ve primidinler olmak üzere iki çeşittir. DNA'da iki pürin bazı, adenin (A) ve guanin (G) ve iki primidin bazı, timin (T) ve sitozin (C) vardır. Her biri bir baz, bir yarım şeker ve bir fosfat grubundan oluşan nükleotidler, ardışık deoksiriboz üniteleri arasında kurulan 5' – 3' fosfodiester bağı ile uzun polinükleotid zinciri oluşturarak polimerize olurlar (Şekil 2.1).



Şekil 2.1: DNA'nın yapısı [1].

DNA'nın anatomik yapısı nesilden nesile tıpatıp aynı genetik bilginin aktarılabilmesine olanak sağlayan kimyasal bilgiyi taşır. 1953 yılında James Watson ve Francis Crick tarafından aydınlatılmış olan DNA'nın doğal yapısı çift sarmaldır [29]. Bu sarmal yapı, iki polinükleotid zinciri zıt taraflardan işleyen, baz çiftleri arasındaki hidrojen bağı ile bir arada tutulan, sağlak bir spiral merdivene benzemekte; bir zincirdeki T diğer zincirdeki A ile ve bir zincirdeki G diğer zincirdeki C ile eşleşmektedir. İnsan genomunda kodlanmış genetik bilginin özgül doğası, hem çekirdek hem de mitokondri içerisindeki her bir çift sarmal kromozomdaki C'ler, A'lar, G'ler ve T'lerden oluşan dizide yatar. DNA'nın iki dizisinin komplementer doğasından dolayı,

bir iplikteki nükleotid baz dizisinin bilinmesi otomatik olarak diğer ipliğin baz dizisinin de bilinmesini sağlar. DNA moleküllerinin çift iplikli yapısı, iki ipliğin ayrılmasını takiben orijinal kalıp iplik ile uyumlu iki yeni komplementer ipliğin sentezlenmesi ile birebir kopyalanmasına olanak sağlar (Şekil 2.2). Benzer şekilde, gerekli olduğunda, bu baz komplementerliği, hasarlı DNA moleküllerinin etkin ve doğru tamirine olanak sağlar [1].



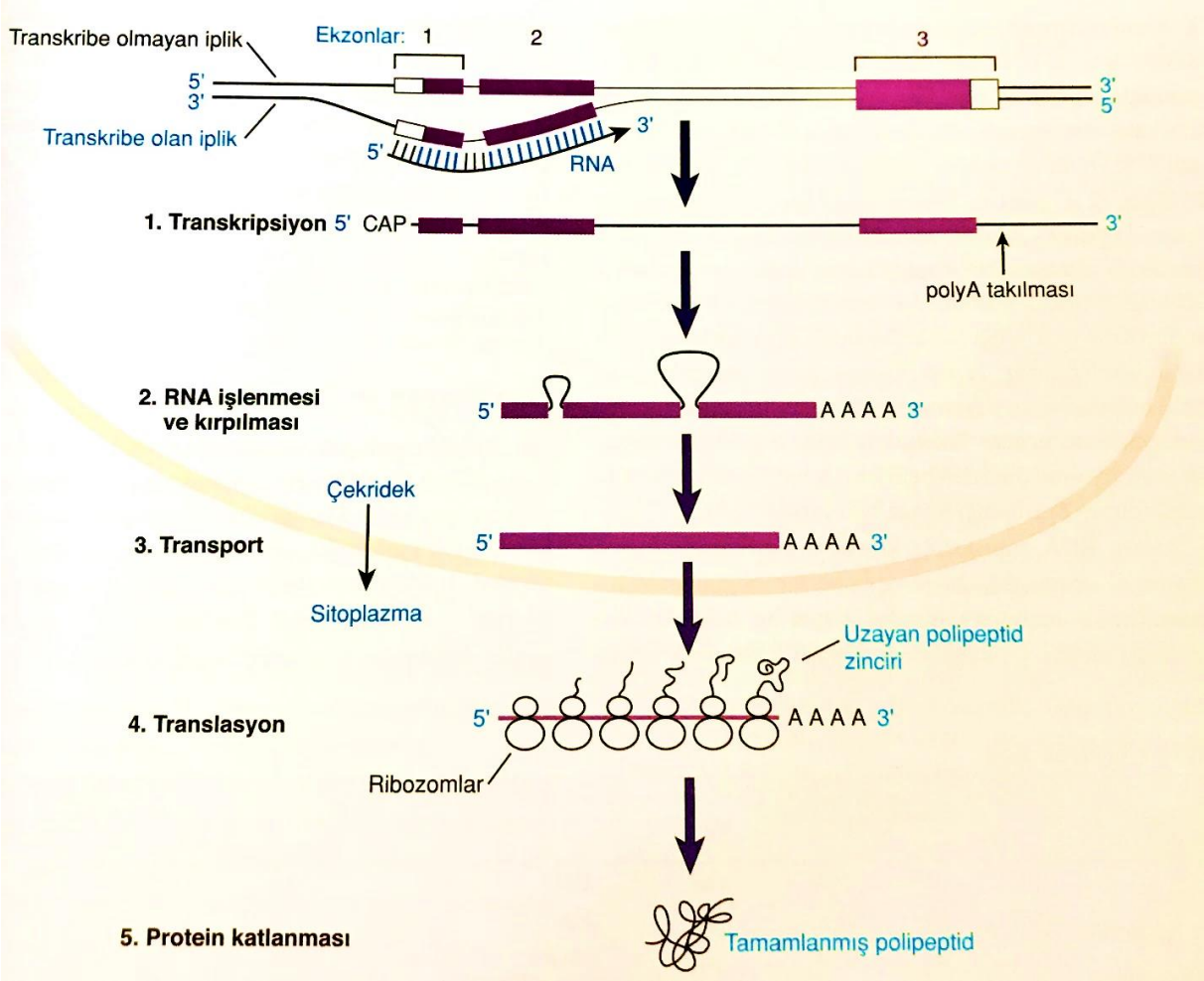
Şekil 2.2: Çift sarmal DNA'nın replikasyonu [1].

Genetik bilgi çekirdek içerisinde bulunan kromozomlardaki DNA'dadır. Ancak, genomda kodlanmış bilginin işlenmesi sitoplazmada gerçekleşen çeşitli hücresel işlevleri özgülleştirmek için kullanılmaktadır. Bu bölümlendirme, insan organizmasının ökaryot olduğu gerçeğini yansıtmakla birlikte, insan hücrelerinin sitoplazmadan çekirdek zarı ile ayrılmış olan ve genomu içeren bir çekirdeğe sahip olduğu anlamını taşımaktadır. Ökaryotik hücrelerde bu bölümlendirmeden dolayı, çekirdekten sitoplazmaya bilgi aktarımı rübonükleik asit (RNA) vasıtasıyla gerçekleştirilir. Bir diğer ifade ile, aktarılan bu iki ilişkili bilgi genlerin DNA kodu ile proteinlerin aminoasit kodudur. RNA'nın kimyasal yapısı deoksiriboz içeriği yerine riboz şekeri bulunması haricinde DNA'ya benzer; buna ek olarak, RNA'da primidin bazlarından timin yerine urasil (U) bulunur. DNA ile bir diğer farkı ise, bir çok organizmada RNA'nın tek zincirli bir molekül olarak yer almasıdır.

Genomik DNA, RNA dizisinin sentezini, RNA ise polipeptid dizisinin sentezini kontrol etmekte ve özgül proteinler DNA ve RNA'nın sentezi ve metabolizması içerisinde yer

almaktadır. Bu bilgi akışı moleküler biyolojinin santral dogması olarak adlandırılmaktadır. Genetik bilgi, genetik kod olarak, ardışık bazların bir araya gelip polipeptiddeki aminoasit dizisini belirleyecek olan genom DNA'sı içinde saklanmaktadır. İlk olarak, transkripsiyon adı verilen süreçte DNA kalıbından RNA sentezlenmektedir. RNA, mesajcı RNA (mRNA) adı verilen formda kodlanmış bilgiyi taşımakta, bu mRNA çekirdekten sitoplazmaya taşınmakta ve sentezlenecek proteinin aminoasit dizisine karar verilmek amacıyla deşifre edilmektedir (translasyon). Translasyon süreci protein sentezinde görev alan, mRNA dahil tüm ilişkili moleküllerin bağlanma bölgesindeki sitoplazmada bulunan bir organel olan ribozomlarda gerçekleşmektedir. Ribozomlar pek çok farklı yapısal proteinin ribozomal RNA (rRNA) adı verilen özgül RNA türleri ile etkileşime geçmesi ile oluşur. Ancak, translasyon üçüncü bir RNA türü olan her mRNA'nın baz dizisindeki kod ile o mRNA tarafından kodlanan proteinin amino asidi arasındaki moleküler ilişkiyi sağlayan taşıyıcı RNA (tRNA)'nın varlığını gerektirmektedir.

Tipik bir insan genini karakterize eden yapısal özellikleri oldukça fazladır. Genlerin çoğunluğu kodlayan diziler ile bir veya daha fazla kodlamayan dizilerden oluşur. Kodlayan dizilere ekzon, kodlamayan dizilere intron adı verilir. Ekzonlar proteinin amino asit dizisini belirleyecek gen bölgeleridir. İntronlar ise çekirdekte RNA'ya transkribe olan ancak sitoplazmadaki olgun mRNA'da bulunmayan (RNA kırpılması) gen bölgeleridir. RNA kırpılması primer RNA transkriptinin 5' ve 3' uçlarının modifikasyonundan sonra intron bölgelerin çıkarılıp ekzon bölgelerin birleştirilmesi sürecidir ve bu mekanizma intronların hem 5' hem 3' ucunda bulunan özgül dizilerce yönetilir. Primer RNA transkripti kimyasal bir "CAP" yapısının RNA'nın 5' ucuna eklenmesi ve kodlayan bilginin downstreamindeki 3' ucunun sonundaki özgül bir noktadan kırpılması ile oluşur. Bu kırpılmayı takiben RNA'nın 3' ucuna poliA kuyruğu takılır; bu kuyruk, sonuçta elde edilecek poliadenile edilmiş RNA transkriptinin 3' translasyon olmayan bölgesinde bulunan AAUAAA (veya bunun bir varyantı) dizisinin varlığı ile özgülleşmiştir. Transkripsiyon sonrası modifikasyonların tümü, çekirdek içerisinde gerçekleşir. Tamamen işlenmiş RNA artık mRNA olarak adlandırılır ve translasyonun gerçekleştiği yer olan sitoplazmaya taşınır (Şekil 2.3).



Şekil 2.3: DNA'dan RNA'ya ve proteine bilgi akışı [1].

Translasyon için anahtar olay, özgül amino asitleri mRNA üzerinde üçlü ardışık baz dizileri ile ilişkilendiren koddur. Her üç baz seti belli bir amino aside özgül olan kodonu oluşturur. Herhangi bir pozisyonda, dört olasılık (A,U,C veya G) vardır; bu durumda $4^3 = 64$ olası üçlü kombinasyon vardır. Bu 64 kodon genetik kodu oluşturmaktadır. Sadece 20 amino asit olduğundan, çoğu amino asit birden fazla kodonla özgülleşmiştir. Kodonların üçü (UGA, UAA veya UAG) dur kodonu olarak adlandırılır, çünkü mRNA'nın translasyonunu o noktada sonlandırır. Bir tane de başlangıç kodonu (AUG) bulunmaktadır ve mRNA'nın okuma çerçevesini belirler; her bir kodon proteinin amino asit dizisini oluşturur.

2.3.KANSERDE GENETİK

Klinik tıptaki en yaygın ve ciddi hastalıklardan biri de kanserdir. Giderek bir kitleye veya tümöre (neoplazm) dönüşecek şekilde kontrolsüz hücre büyümesi sürecinin (neoplazi) daha

virulan formu olan kanserin, köken aldığı bölgeyi çevreleyen komşu dokuları işgal edebilme ve daha uzak bölgelere yayılabilme (metastaz) yeteneği vardır. Kanser genetik bir hastalık olup, mutasyonları kansere neden olan genlere sürücü genler denir. Sürücü genler iki ayrı kategoriye ayrılır: aktifleşen onkogenler ve tümör baskılayıcı genler (TBG). Onkogenler hücre çoğalmasını uyararak veya apoptozisi (programlı hücre ölümü) önleyerek malign (kötü huylu) dönüşümü kolaylaştırırlar (Örn. N-MYC geni akciğer kanserinin gelişiminde rol oynamaktadır. Aktivasyon mekanizması ise DNA amplifikasyonudur.). TBG'ler hücresel işlevlerin birçok yönüyle ilgili proteinleri kodlarlar; doğru kromozom sayısının ve yapısının korunması, DNA onarımı, hücre döngüsünün düzenlenmesi, hücre çoğalması bunlardan bazılarıdır. TGB'lerde mutasyon yoluyla oluşan işlev kaybı, ya doğrudan hücre büyümesi üzerindeki düzenleyici kontrolü ortadan kaldırır, ya da dolaylı olarak artmış mutasyon hızı veya anormal gen ifadesi yoluyla böyle kayıplara yol açar (Örn. p53 geni kolon ve meme kanseri, PTEN geni prostat kanseri, DCC geni kolon kanseri ile ilişkilidir.). Kanser tek bir hastalık olmayıp malignitenin birçok biçimi ve derecesi ile gelebilir. Üç ana kanser sınıfı vardır:

- *Sarkomlar;* kemik, kas, bağ dokusu veya sinir sistemi gibi mezenkimal dokulardan köken alır.
- *Karsinomlar;* barsakları, bronşları veya meme kanallarını döşeyen epitelden köken alır.
- *Hematopoietik ve lenfoid malign neoplazmlar;* lösemi ve lenfoma gibi, kemik iliği, lenfatik sistem ve periferik kan kökenlidir.

Kanserler adlandırılırken, Tablo 2.1'de de verildiği şekilde ön ek alırlar [30].

Tablo 2.1: Kanser adlandırmalarında kullanılan ön ekler ve anlamları.

Ön Ek	Anlamı
adeno-	bez (*İng.* gland)
chrondro-	kıkırdak (*İng.* cartilage)
erythro-	kırmızı kan hücreleri (*İng.* red blood cell)
hemangio-	kan damarları (*İng.* blood vessels)
hepato-	karaciğer (*İng.* liver)
lipo-	yağ (*İng.* fat)
lympho-	lenfosit (*İng.* lymphocyte)
melano-	pigment hücresi (*İng.* pigment cell)
myelo-	kemik iliği (*İng.* bone marrow)
myo-	kas (*İng.* muscle)
osteo-	kemik (*İng.* bone)
retino-	retina (*İng.* retina)

Kanser riski altında bulunan bireylerin kanser gelişmeden önce belirlenmesi genetik araştırmaların önemli bir hedefidir.

2.4.MİKRODİZİ TEKNOLOJİSİ

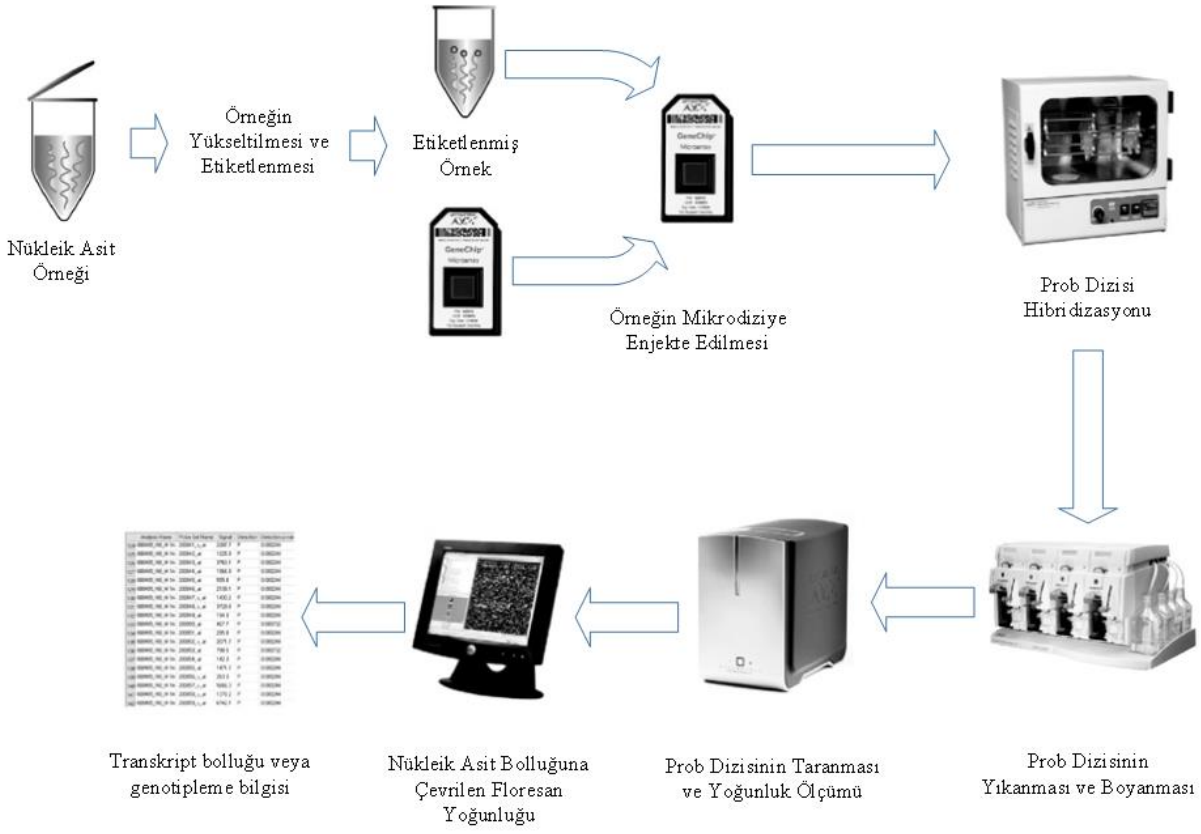
1990'lı yılllardan başlayarak, büyük teknolojik gelişmeler sayesinde atılım gerçekleştiren genomik devrim, bilim insanlarının virüsler, bakteriler, omurgasızlar dahil olmak üzere çeşitli organizmaların dizileri ile birlikte insan genomunun dizisini de tamamlamalarını sağlamıştır. Organizmaların genomik dizi bilgileri kullanılarak, yaşamın ve hastalık durumlarının daha iyi anlaşılması, daha iyi teşhisler koyulması, yeni terapötik müdahaleler sunacak işlevsel biyolojik mekanizmalara dönüştürülmesi, vb. görevleri yerine getirecek eşsiz niteliklere sahip olan mikrodiziler, bilimsel araştırmalarda önemli bir araç haline gelmiştir.Mikrodiziler, giderek artan genomik bilgiyi etkili bir şekilde ölçeklendiren güvenilir, hızlı ve uygun maliyetli bir yöntem sağlarlar.

Affymetrix GeneChip teknolojisi ile diğer mikrodizi tabanlı yöntemler arasında önemli farklılıklar vardır. GeneChip mikrodizilerinin en ayırt edici özelliği, üretiminin fotokimyasal sentezle yönlendirilmesidir. Bu üretim teknolojisi nedeniyle, bir dizi üzerinde kabaca küçük resim boyutunda bir milyondan fazla farklı prob sentezlenebilir. Bu sayılar, aynı hedef sekansı sorgulamak için birden fazla probun dahil edilmesine izin vererek, veri yorumlamasına istatistiksel titizlik sağlar. Yıllar boyunca Affymetrix GeneChip platformu, güvenilir ve sağlam bir sistem olduğunu kanıtlamış ve bilimsel topluluk tarafından birçok yeni keşif ve atılım yapılmasını sağlamıştır [31].

Affymetrix GeneChip mikrodizileri bir dizi teknoloji, tasarım ölçütleri ve kalite kontrol süreçlerinin birleşiminin sonucunda ortaya çıkan esnek bir platformdur. Daha açık bir ifade ile, dizilere ek olarak, tek bir platform olarak geliştirilen bu teknoloji standartlaştırılmış testlere ve reaktiflere, enstrümantasyona (akışkanlar sistemi, hibridizasyon fırını ve tarayıcı) ve veri analiz araçlarına dayanır.

Nükleik asit örneği elde edildiğinde, hedef amplifikasyonu (yükseltilmesi) ve etiketlenmesi işleminin ardından etiketli bir örnek elde edilir. Etiketlenmiş örnek daha sonra prob dizisine enjekte edilir ve hibridizasyon fırınında gece boyunca hibridleşmesine izin verilir. Prob dizisinin yıkanması ve boyanması, aynı anda dört prob dizisini işleyebilen akışkanlar istasyonunda gerçekleşir. Prob dizisi daha sonra Affymetrix GeneChip tarayıcıda taranmaya

hazır hale getirilmiştir. Burada her özelliğin floresan yoğunluğu okunur. Tarayıcının en son sürümü 0.51 um piksel çözünürlüğe sahiptir ve 2.5 um aralıklarla özellikleri tarayabilir. Tarayıcı 65.000'den fazla farklı floresan yoğunluğunu çözebilir. Tarama işlemi sırasında bir fotoçoğaltıcı tüp, floresans değerlerini toplar ve elektronik bir sinyale dönüştürür; bu daha sonra karşılık gelen sayısal değerlere dönüştürülür. Bu sayısal değerler, görüntü veri dosyasını (.dat dosyası) içeren piksel değerleri olarak depolanan floresan yoğunluklarını temsil eder. Veri çıktıları her transkript için bir yoğunluk ölçümü veya ayrıntılı dizi veya genotipleme bilgileri içermektedir.



Şekil 2.4: Affymetrix GeneChip mikrodizi ile bir deneyin akış şeması.

Gen ifade profili çıkarma çalışmaları, dokuların, doku tiplerinin ve hücresel yanıtların, değişen biyokimyasal yolaklara veya hastalığın moleküler ve biyolojik mekanizmalarına biyolojik kavrayış kazandırmak için değişen büyüme koşulları, kanser ve bulaşıcı süreçler gibi çeşitli uyaranlarla karşılaştırılması amacıyla gerçekleştirilmektedir. Bugüne kadar, tüm genom ekspresyon analizi, bilim insanlarının hastalığı tabakalandırmasına, hasta sonucunu tahmin etmesine, suşları değişen virülanslarla karşılaştırmasına, konakçı ve parazit arasındaki ilişkiyi incelemesine ve bazı hastalıkların etkilenen moleküler yollarını anlamalarına yardımcı

olmuştur. Kanser araştırması, mikrodizilerin tartışılmaz bir etkiye sahip olduğu klinik alanlardan biridir. Kanserli hücrelerin tüm genom ekspresyon profilleri, bilim insanlarının kanser alt tiplerini sınıflandırmasına, hastanın prognozunu tahmin etmesine, alternatif tedaviler arasında seçim yapmasına ve hatta yeni tümör sınıflarını tanımlamasına izin vermiştir.

2.5.GEN İFADESİ OMNİBUS

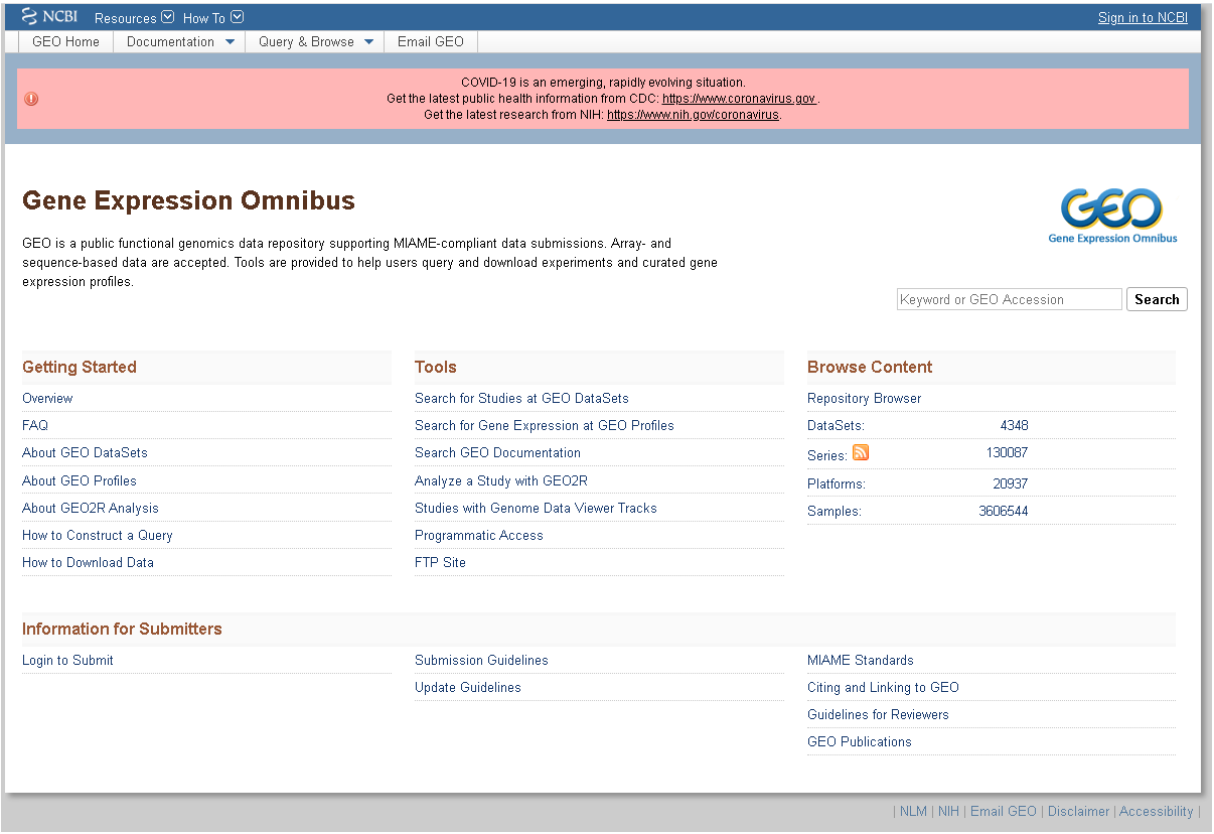
Ulusal Biyoteknoloji Bilgi Merkezi (*İng.* National Center for Biotechnology Information, NCBI)'ndeki Gen İfadesi Omnibus (*İng.* Gene Expression Omnibus, GEO) deposu, DNA mikrodizi teknolojisi tarafından üretilen gen ekspresyonu verilerini, yüksek verimli moleküler bolluk verilerini arşivler ve ücretsiz dağıtır. Veritabanı, işlenmemiş ve işlenmiş verilerin çeşitli tarzlarını, tamamen açıklamalı yüklemeleri destekleyen "Bir Mikrodizi Deneyi Hakkında Minimum Bilgi (*İng.* Minimum Information About a Microarray Experiment, MIAME)" – destekleyici altyapısı ile esnek bir tasarıma sahiptir. GEO şu anda en büyük ve tamamen halka açık gen ekpresyon kaynağı olup, 3.606.544 örnek, 20.937 platform, 130.087 seri ve 4.348 veri kümesini depolamaktadır. Bu verilerin kullanılabilirliğini en üst düzeye çıkarmak için, bu verilerin bireysel genler veya tüm çalışmalar düzeyinde etkili bir şekilde araştırılmasını, sorgulanmasını ve görselleştirilmesini sağlayan, kullanıcı dostu web tabanlı arayüzler ve uygulamalar geliştirilmiştir [32], [33], [34].

GEO veritabanında göndericileri tarafından sağlanan verinin depolanması MSSQL sunucu ilişkili veritabanında üç ana başlıkta gerçekleştirilir.

- Platform: Dizinin özet açıklamasını ve dizi şablonunu tanımlayan bir veri tablosunu içerir. Tablodaki her satır tek bir öğeye karşılık gelir ve gönderici tarafından sağlanan sekans açıklama (*İng.* sequence annotation) ve izleme bilgilerini içerir.
- Örnek: Biyolojik kaynağın ve maruz kaldığı deney protokollerinin bir açıklamasını ve ilgili platformdaki her bir eleman için hibridizasyon ölçümlerini içeren bir veri tablosunu içerir.
- Seri: Bir çalışmanın parçası olduğu düşünülen bir dizi ilgili örneği tanımlar ve genel çalışma amacını ve tasarımını açıklar.

Bu üç nesnenin her birine, kayıtları alıntılamak ve almak için kullanılabilecek bir erişim numarası atanır. Örnek veri tablolarına ve açıklayıcı bilgilere ek olarak, orijinal mikrodizi

tarama görüntüleri veya önceden işlenmiş niceleme (*İng.* quantification) verileri gibi ek dosyalar kabul edilir ve veritabanı bağlantıları içeren bir FTP sitesinde depolanır.



Şekil 2.5: GEO veri tabanının web arayüzü.

Veri tabanından veri kümelerine GEO erişim numaraları ile veya GEO kayıt numaraları (GEO DataSet Record) ile nasıl erişilebileceğine dair bağlantı adresleri Tablo 2.2'de verilmiştir.

Tablo 2.2: GEO veritabanından veri kümelerine erişim yöntemleri.

	Açıklama	**Erişim Bağlantı Adresi**
GEO Erişim Numaraları	Seri	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSExxx
	Platform	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPLxxx
	Örnek	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSMxxx
GEO Kayıt Numarası	Veri Kümesi	https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDSxxx

3. MALZEME VE YÖNTEM

3.1. ÇALIŞMADA KULLANILAN VERİ KÜMELERİ

Tez çalışması kapsamında, NCBI GEO [33] veritabanından GDS Browser hizmeti ile ücretsiz erişilebilen, aynı gen çipi platformunu kullanan 5 kanser türüne ait çalışmaların mikrodizi gen ifade veri kümeleri kullanılmıştır. Richardson ve ark. meme kanseri üzerine yaptıkları mikrodizi çalışmasında, X kromozom anormalliklerinin, kalıtsal ve sporadik bazal-benzeri karsinomanın patogenezine katkıda bulunduğunu göstermişlerdir. Alimonti ve ark. Pten geninin insan kanserinde sıklıkla kaybedilen ve kalıtsal kanser yatkınlık sendromlarında mutasyona uğrayan bir TBG olduğunu vurgulayarak meme kanseri üzerine yaptıkları mikrodizi çalışmasında, kanser gelişimi için TBG'lerin dozundaki ince azalmanın dokuya özgü bir şekilde tümör oluşumuna yatkın olduğu alternatif bir çalışma modeli sunmuşlardır (GEO erişim numarası GSE3744; [35], [36]). Sato ve ark. pulmoner rezeksiyon uygulanan hastalardan alınan 23 klinik küçük hücreli akciğer kanseri (*İng.* small cell lung cancer, SCLC) örneği ve akciğer dahil 42 normal doku örneğini mikrodizi teknolojisini kullanarak analiz etmişlerdir. SCLC, kötü prognozu olan bir akciğer kanseri alt tipidir. Araştırmacıların elde ettikleri sonuçlar, SCLC'nin altında yatan moleküler mekanizmalar hakkında fikir vermektedir (GEO erişim numarası GSE43346; [37]). Bowen ve ark. normal yumurtalık yüzey epitelinin ve yumurtalık kanseri epitel hücrelerinin mikrodizi teknolojisi kullanarak yaptıkları analiz sonrasında karşılaştırmasını yapmışlardır. Araştırmacıların elde ettikleri sonuçlar, yumurtalık adenokarsinomunun gelişiminde yumurtalık yüzey epitelinin rolü hakkında fikir vermektedir (GEO erişim numarası GSE14407; [38]). Arredouani ve ark. mikrodizi teknolojisini kullanarak iyi huylu ve kötü huylu prostat dokularının analizini yapmışlardır. Araştırmacıların bu çalışmalarından elde ettikleri sonuçlar prostat kanseri için bir gen ifade imzası hakkında fikir vermektedir (GEO erişim numarası GSE55945; [39]). Hiraoka ve ark. epitelyal neoplastik hücrelerin tümör gelişiminin 3 aşamasında ve ana pankreatik kanaldan normal hücrelerin analizini mikrodizi teknolojisini kullanarak gerçekleştirmişlerdir. Araştırmacılar anti-tümör bağışıklığının tümör ilerlemesi boyunca değiştiğini vurgulamışlardır. Elde ettikleri sonuçlar, kanser progresyonu sırasında etkili bağışıklık sürveyansının nasıl ve ne zaman aşıldığına dair fikir vermektedir (GEO erişim numarası GSE19650; [40]). 5 kanser türüne ait veri kümelerinin özellikleri Tablo 3.1'de özetlenmiştir.

Veri kümelerinin analize hazır getirilmesi için gen ifade verileri, gen sembolleri (bundan sonra sadece gen olarak belirtilecektir) ve örneklerin sınıf bilgisini içeren uygun bir veri matrisinin oluşturulması gerekmektedir. Bu işlem R programlama ile Tablo 1'deki kayıt numaraları kullanılarak ve RStudio yazılımında GEOQuery kütüphanesinin fonksiyonlarından yararlanılarak gerçekleştirilmiştir [41]. Ardından, veri kümeleri MS Excel dosya formatlarından .xlsx uzantısı ile üretilmiş ve MATLAB yazılımı ile içe aktarılarak analizin devamı için gerekli olan "veriyi içe aktarma (*İng.* data import) " basamağı tamamlanmıştır.

Tablo 3.1: Beş kanser türüne ait mikrodizi veri kümelerinin özellikleri.

Özellik	Meme Kanseri	Akciğer Kanseri	Over Kanseri	Prostat Kanseri	Pankreas Kanseri
Affymetrix Gen Çipi	Affymetrix Human Genome U133 Plus 2.0 Array				
Platform	GPL570				
Organizma	Homo sapiens				
Referans Serisi	GSE3744	GSE43346	GSE14407	GSE55945	GSE19650
Kayıt Numarası	GDS2250	GDS4794	GDS3592	GDS4824	GDS3836
Örnek Sayısı	47	65	24	21	22
Gen/Prob Sayısı	54675				
Sınıf Sayısı	4	2	2	2	4

3.2.VERİ ÖN İŞLEME YÖNTEMLERİ

3.2.1. Eksik Değer Kestirimi

Mikrodizi teknolojisi, yirmi yıllık geçmişine rağmen hala eksik değer (*İng.* missing value) içerebilmektedir [42], [43], [44], [45]. Fabrikasyon hataları, zayıf hibridizasyon, yetersiz çözünürlük veya çip (*İng.* gene chip) üzerindeki kirler eksik verilerin oluşum sebeplerinden bazılarıdır. Eksik değerlerle ilgili yapılan en basit işlem, eksik değer içeren probları gözlem verisinden çıkarmak olmuştur. Fakat bu yöntem (i) çok az sayıda eksik değer içeren prob sayısı söz konusu olduğunda veya (ii) tüm örneklerin çözümlemesi yapılırken, yorumlama süresince ciddi bir ön yüklemeye (*İng.* bias) sebep olmayacaksa yararlı olabilir [46], [47].

Eksik verilerin kontrolü ve düzeltilmesi, bu eksik değerlerin yerine kestirilen yeni değerlerin yerleştirilmesini amaçlayan bir dizi yöntemi içermektedir. Çoğu durumda, bir veri kümesinin nitelikleri birbirinden bağımsız değildir. Bu yüzden, nitelikler arasındaki ilişkilerin tanımlanması ile eksik değerler belirlenebilir. Eksik veri problemini çözmenin basit bir çözümü

mikrodizi deneylerini tekrarlamaktır. Ancak bu çözüm çok maliyetli ve verimsizdir. Bir başka çözüm de, bir veya daha fazla eksik değer içeren probları kaldırmaktır ki bu işlemin önemli bilgilerin de kaybolmasına sebep olacağı açıkça anlaşılmaktadır. Sonuç olarak, eksik değerlerin doğru şekilde düzeltilmesi için daha gelişmiş algoritmaların geliştirilmesi zorunludur.

Eksik değer düzeltme çalışmalarında ilk yaklaşımların eksik değerlerin yerine sıfır veya satır/sütun ortalama değerlerinin (aynı zamanda ortalama ve ortanca değer yöntemleri olarak da ifade edilmektedir.) yerleştirilmesi üzerine olduğu görülmüştür. Bu oldukça basit yöntemlerin, veri kümesinde özilinti ilişkisini hesaba katmamakla birlikte kestirim doğruluğunun azalmasına sebep oldukları da yine literatürdeki çalışmalarda belirtilmiştir [47], [48].

Eksik değer kestirimi için mikrodizi verilerinin çözümlenmesinde başarımı arttıran ve yüksek doğruluğa sahip olan ileri algoritmalara ihtiyaç duyulmaktadır. Son zamanlarda, eksik değerlerin kestirim başarımını belirgin farkla iyileştiren bu gelişmiş algoritmalara çeşitli türlerde biyolojik verileri içeren çalışmalarda rastlanmıştır. Ancak, önceden yapılan çalışmalarda da sunulduğu gibi, tüm veri kümesi türleri için iyi çalışan bir algoritma henüz geliştirilememiştir. Sonuç olarak, en iyi algoritma bulunamaz, fakat her veri kümesi için uygun değer (optimal) algoritmaları bulunabilir [45], [49], [50].

K-En Yakın Komşu Yöntemi

K-en yakın komşu (*İng.* K-nearest neighbor, KNN) algoritması uygulanırken öncelikle komşu sayısı k değeri belirlenir. Diğer belirli gen ifade verilerinden hedefteki eksik değere olan Öklid uzaklıkları hesaplanır. Hesaplanan bu uzaklıklar sıralanır ve en küçük uzaklığa bağlı olarak en yakın komşular bulunur. En yakın komşu kategorileri toplanır ve en uygun komşu kategorisi seçilir. n boyutlu uzayda Öklid uzaklığının hesaplanması Eşitlik 3.1 ile sağlanır.

$$u(k,l) = u(l,k) = \sqrt{(k_1 - l_1)^2 + (k_2 - l_2)^2 + \cdots + (k_n - l_n)^2} = \sqrt{\sum_{i=1}^{n} (k_i - l_i)^2} \qquad (3.1)$$

Burada $i = 1,2, \ldots n$ için $k = (k_1, k_2, \ldots, k_n)$ ve $l = (l_1, l_2, \ldots, l_n)$ kartezyen koordinatlarına sahip k ve l değerlerinin arasındaki uzaklığın hesaplanmasında Pisagor Teoremi esas alınır.

KNN yönteminin avantajı, artan eksik değer oranına rağmen başarımında daha az bozulma göstermesidir.

Bayesçi Temel Bileşen Analizi Yöntemi

Bayesçi Temel Bileşen Analizi (*İng.* Bayesian Principal Component Analysis, BPCA) yöntemi, temel bileşen analizi için yinelemeli iyileştirme algoritması olan beklenti –ençoklama (*İng.* Expectation – Maximization, EM) algoritması ile Bayesçi modelini birleştirir. Standart PCA'da eğitim kümesinden uzak, temel alt uzaya yakın olan veri aynı geri çatma hatasına (*İng.* reconstruction error) sahip olabilir. BPCA geri çatma hata değerlerine göre benzerlik (*İng.* likelihood) hesaplar. BPCA yönteminin avantajı, kendisinden önce geliştirilen yöntemlerden (özellikle de KNN tabanlı yöntemden) belirgin farkla daha iyi sonuçlar vermesidir. Sebebi ise, diğer yöntemlerin belirlenmesi oldukça zor olan model parametrelere ihtiyaç duyması; BPCA yönteminin ise parametrelere ihtiyaç duymamasıdır. BPCA yöntemi özelikle mikrodizi çalışmalarında eksik değer düzeltme işlemi için geliştirilmiştir [51].

Yerel En Küçük Kareler Yöntemi

Yerel en küçük kareler (*İng.* Local least squares, LLS) yöntemi, eksik değerli genin en yakın k-komşularının doğrusal bir kombinasyonuna dayanır. Değişkenler arasındaki uzaklık Pearson, Spearman veya Kendall özilinti sabitlerinin mutlak değeri olarak tanımlanır. LLS yöntemi ile BPCA yönteminin en büyük farkı, LLS yönteminin yerel benzerlik yapısına dayanan bir optimizasyon süreci olması; BPCA yönteminin ise temel bileşenlere dayanan bir optimizasyon süreci olmasıdır.

Ortalama ve Ortanca Değer Yöntemleri

Ortalama ve ortanca yöntemleri, eksik değerli genin tüm örneklerdeki değerlerinin (satır) ortalaması veya ortanca değerinin hesaplanmasına dayanır. Ortalama veya ortanca değer hesaplanarak eksik değer yerine yerleştirilir.

3.2.2. Veri Dönüşümü

Gen ifade verilerinin, normallik veya sabit varyans gibi istatistiksel özelliklerin istendiği durumlarda dönüştürülmesi gerekebilir. Dönüştürülmüş değer $y_{gn} = h(\omega_{gn})$ olsun. Burada h,

seçilmiş bir matematiksel fonksiyonu, ω_{gn} ise gen ifade verisini ifade etmektedir. Bu bölümde, monoton artan dönüşüm yöntemlerine yer verilmiştir [52], [53], [54].

Logaritmik Dönüşüm

Mikrodizi okumalarına uygulanan en yaygın dönüşüm, logaritmik dönüşümdür (Eşitlik 3.2). Logaritmanın tabanı 2, 10 veya doğal logaritmik sabit olan e olabilir.

$$y_{gn} = log_2\, \omega_{gn} \tag{3.2}$$

Box-Cox Dönüşüm Ailesi

Önceki iki dönüşüm, Box-Cox dönüşüm ailesinin iki üyesidir. Bu aile aşağıdaki gibi tanımlanır (Eşitlik 3.3).

$$y_{gn} = \frac{{\omega_{gn}}^{d} - 1}{d} \tag{3.3}$$

d parametresinin ½ değerinde dönüşüm karekök dönüşümüne yaklaşırken; d parametresinin 0' a yaklaşması dönüşümü logaritmik dönüşüme yaklaştırır. d parametresinin 1'e eşit olduğu durumda, orijinal yoğunluk verisini değiştirmek dışında hiçbir dönüşüm yapmayacaktır. Box-Cox dönüşüm ailesi, hangi d değerinin istenen istatistiksel özelliklere sahip dönüştürülmüş değerler verdiğini görmek için incelenebilen bir dizi dönüşüm sağlar (esneklik).

İlgin Dönüşüm

Mikrodizi yoğunluk ölçümlerinin ölçeğinin başlangıç noktası a gibi sabit bir miktarda kaydırmak gerekebilir veya istenebilir. Bu durumda dönüştürülmüş değer aşağıdaki gibi olacaktır (Eşitlik 3.4).

$$y_{gn} = \omega_{gn} + a \tag{3.4}$$

Burada, a öteleme parametresi olup, gerçek yoğunluk verisinden belirli bir miktar kaymayı ifade eder. Bu dönüşüm yöntemi, logaritmik dönüşüm yöntemi ile birleşerek yeni bir ifade kazanabilir (Eşitlik 3.5).

$$y_{gn} = log_2(\omega_{gn} + a) \tag{3.5}$$

Logaritmik dönüşüm pozitif değerlere uygulanabildiği için, mikrodizinlerde arka-plan düzeltme gerekliliği, negatif okuma değerleri içerme, vb. durumlar söz konusu olduğunda ilgin dönüşüm (*İng.* affine transformation) logaritmik dönüşüm ile birlikte kullanılarak bu işlemler gerçekleştirilebilmektedir. Bu durumda, tüm değerlerin pozitif olmasını sağlayacak yeterli büyüklükteki bir a öteleme değeri seçilmelidir.

3.2.3. Veri Normalizasyonu

Normalizasyon tekniklerinin kullanımı, oligonükleotid mikrodizilerinin veri kalitesini arttırmak ve biyolojik farklılıkları (diferansiyel ifadeyi) saptamak için kullanılan önem (*İng.* significance) testlerinin gücünü artırmak amacıyla mikrodizi gen ifade ölçümleri arasındaki değişimi azaltmayı hedeflemektedir [9].

Minimum – Maksimum Normalizasyonu

En yaygın kullanılan normalizasyon yöntemlerinden biri olan minimum – maksimum normalizasyonu yöntemi [55], doğrusal bir dönüşüm yöntemi olup, veri kümesinin en küçük elemanını istenilen bir A değerine, en büyük elemanını istenilen bir B değerine ve aradaki değerleri rank değerine orantılı olarak (A,B) aralığına dönüştürmektedir (Eşitlik 3.6).

$$y_{gn,i} = \frac{\omega_{gn,i} - \min(\omega_{gn})}{\max(\omega_{gn}) - \min(\omega_{gn})} \times (B - A) + A \tag{3.6}$$

Burada, *i*, veri dizisindeki her bir gen ifade verisini ifade etmektedir.

Tez kapsamındaki tüm çalışmalarımızda veri kümeleri (0,1) aralığına dönüştürülmüştür. Bu durumda A=0, B=1 ve eşitlik aşağıdaki gibi olmaktadır (Eşitlik 3.7).

$$y_{gn,i} = \frac{\omega_{gn,i} - \min(\omega_{gn})}{\max(\omega_{gn}) - \min(\omega_{gn})} \tag{3.7}$$

Z-değeri Normalizasyonu

Z-değeri normalizasyonu yöntemi [55], bir önceki yöntemin veri kümesindeki olası az sayıda çok büyük veya çok küçük değerleri (*İng.* outliers) dikkate almamasından kaynaklı ara

değerlerin belirli bir değer aralığında sıkışması problemini çözmek için kullanılabilecek bir yöntem olarak karşımıza çıkmaktadır. Bu problemi aşmak için veri kümesinin istatistiksel özelliklerinden olan ortalama ve varyans değerlerini kullanmaktadır. Bu yöntemin formülasyonu Eşitlik 3.8'de verildiği gibidir.

$$y_{gn,i} = \frac{\omega_{gn,i} - \mu_{\omega_{gn}}}{\sigma_{\omega_{gn}}} \tag{3.8}$$

Burada, $\mu_{\omega_{gn}}$, gen ifade verilerinin ortalaması, $\sigma_{\omega_{gn}}$, gen ifade verilerinin standart sapmasıdır.

Minimum – maksimum normalizasyonu yöntemi, veri kümesinin pozitif bir aralıkta kalmasını garantileyebilirken, z-değeri normalizasyonu yönteminde dönüştürülmüş değerler negatif ve pozitif değerler almaktadır. Bu yöntemde, düşünüldüğü gibi, ortalamadan küçük yoğunluk değerleri negatif değerler elde edilmesinin kaynağıdır.

Kantil Normalizasyon

Kantil normalizasyon yöntemi iki ya da daha fazla farklı özellikteki veri kümesinin benzer dağılıma sahip olmasını sağlamak üzere kullanılan bir istatistiksel normalizasyon yöntemidir [56]. Bu yöntemin algoritması aşağıda maddeler halinde açıklanmıştır:

- İki boyutlu veri kümesinin satırları öznitelikleri (*İng.* attributes), sütunları ise örnekleri (*İng.* samples) göstermiş olsun. Her bir örnek için sütun bazında özniteliklerin küçükten büyüğe sıralandığı durumdaki sıra numarası kaydedilir.
- Ardından, her bir örnek için sütun bazında öznitelikler küçükten büyüğe sıralanır.
- Her bir özniteliğin ortalaması alınır (satır ortalaması).
- Küçükten büyüğe sıralanmış olarak ve öznitelik sayısınca elde edilen bu ortalama değerler, her bir örnek için özniteliklerin sıra numaralarına göre gerçek değerlerin yerine yerleştirilir.

Bu normalizasyon yöntemi ile örneklerin ortalaması ve standart sapması gibi istatistiksel özellikleri aynı yapılarak, her örneğin aynı dağılıma sahip olması sağlanır.

3.3. İSTATİSTİKSEL ÇÖZÜMLEME YÖNTEMLERİ

3.3.1. Kat Değişimi

Kat değişimi (*İng.* fold change) ifadesi mikrodizin çalışmalarında K-kat değişim (*İng.* K-fold change) olarak biyolojik anlamlılığı ortaya çıkarmada kullanılmaktadır. Tabii ki, biyolojik anlamlılığın bulunmasının ontoloji çalışmaları ve yolak çözümleme yöntemleri ile sağlandığı ve kat değişimi parametresinin bir ölçüt olarak kullanıldığı dikkate alınmalıdır. Karşılaştırılacak iki örneğin öznitelikleri arasındaki mutlak oran olarak ifade edilebilen kat değişimi oranının (*İng.* fold change ratio) matematiksel ifadesi Eşitlik 3.9'da verilmiştir [57].

$$k_i = \frac{\bar{x}_{A,i}}{\bar{x}_{B,i}} \tag{3.9}$$

Burada, $\bar{x}_{A,i}$ ve $\bar{x}_{B,i}$ değerleri, A ve B örneklerinin i. nci özniteliğine ait ortalama değerlerini göstermektedir.

İleride değinilecek olan farksal olarak ifade edilen genlerin tespit edilmesinde $k_i > K$ veya $k_i < 1/K$ ölçütü dikkate alınmaktadır. Bu ölçütlere uyan genlerin önemli ölçüde değişim gösterdiği varsayılmaktadır.

Bir başka çalışmada [58], kat değişimi farkları (*İng.* fold change difference) tanımı da kullanılmıştır (Eşitlik 3.10).

$$k_i = \bar{x}_{A,i} - \bar{x}_{B,i} \tag{3.10}$$

Kat değişimi oranının mikrodizin çalışmalarında yaygın kullanılan formunun, logaritma-2 tabanında ve değişim miktarının 2^k şeklinde kullanılarak ifade edildiği bilinmektedir [59], [60]. Eşitlik 3.9'daki ve 3.10'daki ifadeler bu tanıma göre yeniden düzenlendiğinde Eşitlik 3.11 elde edilecektir.

$$\tilde{k}_i = \log_2 k_i = \log_2 \left(\frac{\bar{x}_{A,i}}{\bar{x}_{B,i}} \right) = \log_2 \bar{x}_{A,i} - \log_2 \bar{x}_{B,i} \tag{3.11}$$

3.3.2. Welch t – testi

Karşılaştırılan (normal dağılımlı ve bağımsız) iki popülasyonun varyansları eşit olmadığı durumda ortalamalarının farkları arasında önem derecesini ortaya çıkaran hipotez testi Eşitlik 3.12'de verilen Welch t-testidir [61].

$$t_i = \frac{\bar{x}_{A,i} - \bar{x}_{B,i}}{\sqrt{\frac{s_{A,i}^2}{n_A} + \frac{s_{B,i}^2}{n_B}}} \tag{3.12}$$

Burada, t_i, i. nci gene ait t-değeri, $\bar{x}_{A,i}$ ve $\bar{x}_{B,i}$, A ve B örneklerinin gen ifade değerlerinin ortalama değerleri, $s_{A,i}^2$ ve $s_{B,i}^2$, A ve B örneklerinin varyans değerleri, n_A ve n_B, A ve B örneklerinin sayısıdır. t-değeri farksal olarak ifade edilen genlerin tespit edilmesinde tercih edilen önemli bir parametrik test istatistiği değeridir. Öyle ki, incelenen iki popülasyonda t_i parametresinin (mutlak değerinin) elde edilen büyük değerleri, ilgili i. nci genin istatistiksel olarak farksal olarak ifade edildiğine güçlü bir kanıt olarak araştırmacılar tarafından kabul edilmektedir. Anlamlılık düzeyinin seçilmesi ve p-değerlerinin hesaplanması (uygun serbestlik derecesine karşılık Student-t dağılımı tablosundan bulunarak) sonucunda farksal olarak ifade edilen genlerin tespiti yapılmaktadır [62].

3.3.3. Tek Yönlü Varyans Analizi

Tek faktörlü ve ikiden çok örneklemi olan normal dağılımlı problemlerde varyans analizi (ANOVA) yaklaşımı kullanılmaktadır. ANOVA sürecinde verideki değişkenliğin iki kaynağı olduğu varsayılmaktadır: (1) Grup (ya da örneklem) içi değişkenlik, (2) Gruplar arası değişkenlik. Bunun yanında, yığın ortalamalarının eşitliğini test etmek için kullanılan tek yönlü ANOVA'da varsayımlar şu şekilde ifade edilebilir: Karşılaştırma yapılacak grup sayısı en az üç ve üzeri sayıdadır, kullanılacak veri setinde her bir örnek diğerinden bağımsız ölçümler içermektedir, sürekli değerlerden oluşmaktadır ve grup içi dağılım normaldir. En az iki ortalamanın eşit olmadığı alternatifine karşı, k sayıda yığın ortalamasının eşit olduğunu ifade eden yokluk (H_0) ve alternatif (H_1) hipotezleri aşağıdaki gibi yazılabilir:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k,$$

H_1: En az iki ortalama eşit değildir.

Tek yönlü ANOVA'da, toplam kareler toplamı (SS_T), işlem kareler toplamı (SS_A) ve hata kareler toplamı (SS_E) olmak üzere üç önemli değişkenlik ölçüsü tanımlanır (Eşitlik 3.13, 3.14 ve 3.15).

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{3.13}$$

$$SS_A = n\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 \tag{3.14}$$

$$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 \tag{3.15}$$

Burada, y_{ij}, i. nci işlemdeki j. nci gözlemi, $\bar{y}_{..}$, nk sayıdaki tüm gözlemlerin ortalamasını, $\bar{y}_{i.}$, i. nci işleme ilişkin örneklemdeki tüm gözlemlerin ortalamasını göstermektedir.

Yukarıda bahsedilen iki değişkenlik kaynağı, örneklem varyanslarını kullanarak önemli bir oran üretir ve oranlar F dağılımı (varyans oran dağılımı olarak da adlandırılır) ile bağlantılı olarak kullanılır. Yığın ortalamaların eşitliğinin testi için kullanılan F-oranı (s_1^2/s^2), H_0 doğru iken, $k-1$ ve $k(n-1)$ serbestlik dereceli (*İng.* degrees of freedom) F-dağılımına sahip rastgele değişken F'in değeridir. Bir varyans analizi problemindeki hesaplamalar Tablo 3.2'de görüldüğü gibi genellikle tablolaştırılarak özetlenir [63].

Tablo 3.2: Tek yönlü ANOVA hesaplama sonuçları.

Değişimin Kaynağı	**Kareler Toplamı**	**Serbestlik Derecesi**	**Kareler Ortalaması**	**Hesaplanan *f***
İşlemler	SS_A	$k-1$	$s_1^2 = \frac{SS_A}{k-1}$	$\frac{s_1^2}{s^2}$
Toplam	SS_E	$k(n-1)$	$s^2 = \frac{SS_E}{k(n-1)}$	
Ortalama	SS_T	$kn-1$		

Bunun ardından hipotezler belirlenen α anlamlılık düzeyinde p-değeri yaklaşımı ile reddedilir veya kabul edilir. Yokluk hipotezinin reddedilmesi için, – sonuçları hazır veren istatistik paket

programları kullanılmayacaksa – F-dağılımına ilişkin tablolardan yararlanılarak, hesaplanan f-değerinin, belirlenen α anlamlılık düzeyindeki kritik değerden büyük olması şartı aranır.

Farksal olarak gen ifade analizlerinde ANOVA kullanılarak anlamlı ölçüde ifade edilen genlerin tespit edilmesi sağlanmıştır [5], [64], [65]. Sahte pozitif seçimlerin önüne geçmek için çoklu karşılaştırma testleri ile gerçek pozitif olan genler ortaya çıkarılmıştır.

3.3.4. Wilcoxon Sıra Sayıları – Toplam Testi

Gen ifade verilerinin genellikle normal dağılımlı olmaması araştırmacıları parametrik olmayan test istatistiklerinin kullanımına yönlendirmiştir [6], [7], [8], [66]. Parametrik t-testin parametrik olmayan alternatifi konumundaki Wilcoxon Sıra Sayıları-Toplam testi (*İng.* Wilcoxon Rank-Sum Test) tez çalışmaları kapsamında yaygın olarak kullanılmıştır. Dağılımdan bağımsız yöntemler olarak da adlandırılan parametrik olmayan testler sıralama ölçeğini dikkate aldığından, örneklem tarafından sağlanan bilginin tamamını kullanmazlar. Bu dezavantaj olarak görülse de, normal dağılımdan çok uzaklaşan bir dağılıma sahip örneklem üzerinde parametrik olmayan testler daha etkilidir. Bunun yanı sıra, normal dağılım gösteren örneklem üzerinde parametrik olmayan testler, parametrik testler kadar etkili olabilmektedir.

Normal olmayan iki sürekli dağılımın ortalamalarının eşitliği test edileceği zaman ve örneklemler bağımsız olduğunda (yani gözlemler eşleştirilmemiş ise) Wilcoxon sıra sayıları-toplam testi (WRS) veya Wilcoxon iki örneklem testi iki örneklem t-testine uygun bir alternatiftir. WRS testinin uygulanışı şu şekildedir: $\tilde{\mu}_1 = \tilde{\mu}_2$ şeklindeki H_0 yokluk hipotezi uygun bir alternatife karşı test edilmektedir. Önce, her bir yığından bir rasgele örneklem seçilir. Bu durumda, küçük örneklemdeki gözlemlerin sayısı n_1 ve büyük örneklemdeki gözlemlerin sayısı n_2 olsun. Bu örneklemlerin çapı eşit olduğunda n_1 ve n_2 rasgele biçimde atanabilir. Birleştirilmiş örneğin $n_1 + n_2$ sayıdaki gözlemi her bir gözlem için artan sırada ve $1,2,\ldots,n_1 + n_2$ sıra sayıları atanarak düzenlenir. Eşitlik durumunda (aynı değerli gözlemler), gözlemleri farklı oldukları durumda alabilecekleri sıra sayılarının ortalamasıyla değiştirilir. Küçük örneklemdeki n_1 gözleme karşılık gelen sıra sayılarının toplamı ω_1 ile gösterilsin. Benzer şekilde ω_2 de büyük örneğe karşılık gelen n_2 sıra sayısının toplamını temsil etsin. $\omega_1 + \omega_2$ toplamı sadece iki örneklemdeki gözlem sayısına bağlıdır ve hiçbir şekilde deneyin sonuçlarından etkilenmez (Eşitlik 3.16).

$$\omega_1 + \omega_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \tag{3.16}$$

olur ve bu da 1,2,..., $n_1 + n_2$ tamsayılarının aritmetik toplamıdır. ω_1 belirlendikten sonra ω_2 Eşitlik 3.17 kullanılarak bulunabilir.

$$\omega_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - \omega_1 \tag{3.17}$$

WRS testinde ω_1 küçük ve ω_2 büyük ise, $\tilde{\mu}_1 = \tilde{\mu}_2$ yokluk hipotezi $\tilde{\mu}_1 < \tilde{\mu}_2$ alternatifi lehinde reddedilir. Benzer şekilde, ω_1 büyük ve ω_2 küçük ise, $\tilde{\mu}_1 > \tilde{\mu}_2$ alternatifi kabul edilebilir. İki yanlı test için, eğer ω_1 küçük, ω_2 büyük veya ω_1 büyük, ω_2 küçük ise H_0, H_1 lehinde reddedilir. Diğer bir ifade ile, ω_1 yeterince küçük ise $\tilde{\mu}_1 < \tilde{\mu}_2$ alternatifi, ω_2 yeterince küçük ise $\tilde{\mu}_1 > \tilde{\mu}_2$ alternatifi, ω_1 ve ω_2 den küçük olanı yeterince küçükse $\tilde{\mu}_1 \neq \tilde{\mu}_2$ alternatifi kabul edilir. Gerçek uygulamada, anlamlı ilişkinin kararı genellikle U_1 ve U_2 istatistikleriyle (Mann-Whitney U testi) ilişkilendirilmiş $u_1 = \omega_1 - \frac{n_1(n_1+1)}{2}$ veya $u_2 = \omega_2 - \frac{n_2(n_2+1)}{2}$ değerlerine veya U_1 ve U_2 nin küçüğü olan U istatistiğinin u değerine dayalı olarak verilir. U_1 ve U_2 nin her ikisi de simetrik örneklem dağılımına sahip olduğundan ve $u_1 + u_2 = n_1 n_2$ olacak şekilde 0 dan $n_1 n_2$ ye kadar olan aralıkta değerler aldıklarından, bu istatistikler kritik değer tablolarının oluşturulmasını kolaylaştırır. u_1 ve u_2 nin formüllerinden ω_1 küçük ise u_1 in küçük olacağı ve ω_2 küçük ise u_2 nin küçük olacağı görülmektedir. Dolayısıyla, seçilen anlamlılık düzeyinde uygun istatistik U_1, U_2 veya U değerleri kritik değer tablolarında verilen örneklem sayılarına karşılık gelen kritik değere eşit veya daha küçük ise yokluk hipotezi reddedilmektedir. Yokluk hipotezine karşı oluşturulabilecek farklı alternatif hipotezler Tablo 3.3'te özetlenmiştir.

Tablo 3.3: Sıra Sayıları-Toplam testi için farklı yöntemler.

H_0	H_1	Hesaplama
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	u

Tez çalışmaları kapsamında, çerçeve boyutunun 5, 9, ve 15 olarak yapılandırılması durumunda çeşitli anlamlılık düzeyleri için WRS testinde dikkate alınan kritik değerler Tablo 3.4'te verilmiştir.

Tablo 3.4: Çerçeve boyutunun değişimi ile WRS testindeki kritik değerin değişimi.

Çerçeve Boyutu (L_F)	Örneklem Sayısı (n_1,n_2)	Anlamlılık Düzeyleri için Kritik Değerler			
		(Tek Yönlü, $\alpha = 0.01$)	(Tek Yönlü, $\alpha = 0.05$)	(Çift Yönlü, $\alpha = 0.01$)	(Çift Yönlü, $\alpha = 0.05$)
5	(5,5)	1	4	0	2
9	(9,9)	14	21	11	17
15	(15,15)	56	72	51	64

3.3.5. Kruskal Wallis – H Testi

Kruskal Wallis–*H* testi, sıra sayıları toplam testinin $k > 2$ örneklem durumu için genelleştirilmiş biçimidir. 1952 yılında W.H. Kruskal ve W.A. Wallis tarafından önerilen bu test, örneklemlerin normal yığınlardan seçilmiş olma varsayımından kaçınılmak istendiğinde, tek-faktör varyans analizindeki ortalamaların eşitliğini test etmek için kullanılan parametrik olmayan bir yöntemdir.

i. nci örneklemdeki gözlemlerin sayısı $n_i (i = 1,2, \dots, k)$ olsun. İlk olarak, k örneği birleştirilir, $n = n_1 + n_2 + \cdots + n_k$ gözlemi artan sırada düzenlenir ve her bir gözlem için $1,2, \dots, n$ sıra sayılarından uygun olanı verilir. Eşit değerler durumunda (aynı gözlemler), gözlemler farklı olmuş olsaydı alacağı sıra sayılarının ortalamasını bu gözlemlere sıra sayısı olarak atamak biçimindeki geleneksel yöntem takip edilir. i. nci örneklemdeki n_i gözleme karşılık gelen sıra sayılarının toplamı r_i rastgele değişkeni ile gösterilsin. "k bağımsız örneklemleri aynı yığından gelmektedir" biçimindeki yokluk hipotezi, H_0' ı test etmek için, bu test, Eşitlik 3.18'de verilen h değerinin hesaplanması ve Ki-kare dağılımına göre seçilen anlamlılık düzeyindeki kritik değer ile kıyaslanması şeklinde kullanılır.

$$h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{{r_i}^2}{n_i} - 3(n+1) \tag{3.18}$$

Burada, r_i, i. nci öneklemin sıra sayılarının toplamı, n, tüm örneklemlerin toplam sayısı, n_i, i. nci öneklemin sayısıdır. Sonuç olarak, eğer h, serbestlik derecesi $v = k - 1$ olmak üzere kritik bölge $h > \chi_\alpha^2$ ye düşerse, α anlamlılık düzeyinde H_0 reddedilir, aksi halde H_0 reddedilmez.

3.4.ÇOKLU KARŞILAŞTIRMA TESTLERİ

Hipotez testleri, gruplar arasında anlamlı bir fark olup olmadığını istatistiksel olarak belirleyebilirken; bu testlerin uygulanması ile yokluk hipotezinin reddedilmesi durumunda, hangi gruplar arasında istatistiksel olarak anlamlı bir fark olup olmadığını çoklu karşılaştırma testleri belirleyebilmektedir. Bununla birlikte, çoklu karşılaştırma testleri, özelliklerin veya özniteliklerin çokluğundan kaynaklanabilen hatalı sonuç çıkarmaları düzenli bir şekilde kontrol edebilen istatistiksel işlemlerdir. Her hipotezin testi için Tip I hata (yokluk hipotezinin gerçekte doğru iken reddedilmesi) ve Tip II hata (yanlış olduğu halde yokluk hipotezinin ret edilmemesi) söz konusudur (Tablo 3.5). Tip I hatanın ortaya çıkma olasılığı, aynı zamanda anlamlılık düzeyi (α) olarak adlandırılmaktadır.

Tablo 3.5: İstatistiksel bir hipotezin test edilmesinde olası durumlar.

	H_0 **doğrudur**	H_0 **yanlıştır**
H_0 **reddedilmez**	Doğru karar	Tip II hata
H_0 **reddedilir**	Tip I hata	Doğru karar

Bir araştırmada, ilk olarak α anlamlılık düzeyinin belirlenmesi gerekmektedir. İkinci olarak, analiz edilecek hata oranının belirlenmesi gerekmektedir. Tez çalışmaları kapsamında, deneysel (veya ailesel) hata oranının (*İng.* Family-wise error rate, FWER) kontrol edilmesine yönelik yöntemlerden biri olan Holland – Copenhaver yöntemi tercih edilmiştir. FWER değerinin kontrol edilmesi ile Tip I hatanın ortaya çıkma olasılığı, belirlenen α anlamlılık düzeyine göre ayarlanmaktadır. FWER değerinin kontrolü ile ilgili işlemlerin temel avantajı, karşılaştırma sayısının artması ile paralel olarak FWER değerinin artmamasıdır [67], [68]. Tablo 3.6'da FWER değerinin kontrolü için kullanılan bazı yöntemlere ait özet bilgiler verilmiştir.

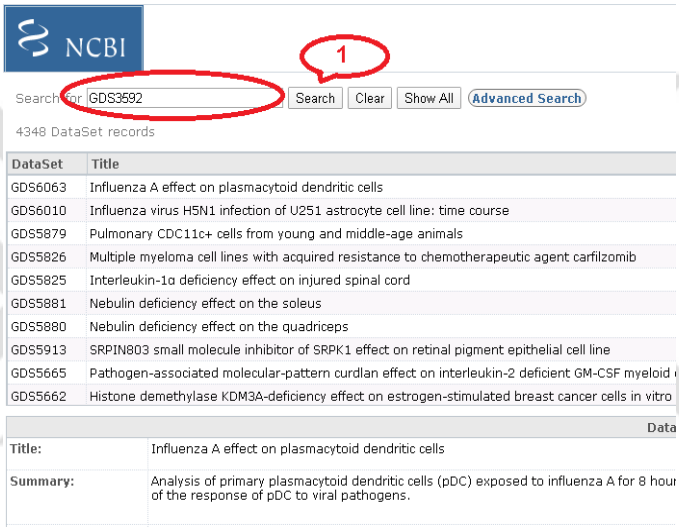
Tablo 3.6: FWER değerinin kontrolünde kullanılan yöntemler.

Yöntem	Eşitsizlik	Açıklama
Orijinal Bonferroni	$p_{(i)} < \alpha'$	Tek aşamalı bir yöntemdir. Yeni anlamlılık düzeyi, $\alpha' = \alpha/m$ ifadesi bulunur. Burada, m, karşılaştırılan hipotez sayısıdır.
Holm [69]	$p_{(m-1)} \geq \alpha/(m-1)$	Azalan aşamalı bir yöntemdir. Öncelikle $H_{(1)}, H_{(2)}, \dots, H_{(m)}$ hipotezlerine karşılık gelen ve p değerlerini gösteren $p_{(1)}, p_{(2)}, \dots, p_{(m)}$ değerleri sırasıyla büyükten küçüğe doğru sıralanır. Ardından, bu yöntemin eşitsizliği dikkate alınarak hipotezlerin geçerliliği değerlendirilir.
Holland – Copenhaver [70]	$p_{(i)} > 1 - (1-\alpha)^{1/(m-i+1)}$	Araştırmacılar, Sidak [71] tarafından geliştirilen, m karşılaştırılacak hipotez sayısını göstermek üzere, $\alpha/m = 1 - (1-\alpha)^{1/m}$ eşitliğini kullanarak her bir hipotez testi için bir ölçüt hazırlamışlardır. Öncelikle $H_{(1)}, H_{(2)}, \dots, H_{(m)}$ hipotezlerine karşılık gelen ve p değerlerini gösteren $p_{(1)}, p_{(2)}, \dots, p_{(m)}$ değerleri öncelikle büyükten küçüğe doğru sıralanır. Ardından, bu yöntemin eşitsizliği dikkate alınarak hipotezlerin geçerliliği değerlendirilir.

3.5.GEO2R PLATFORMU İLE FARKSAL OLARAK İFADE EDİLEN GENLERİN TESPİTİ

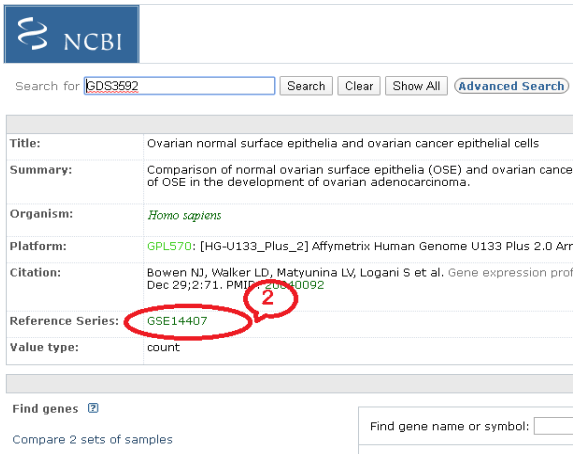
NCBI veri tabanının hizmetlerinden biri olan GEO2R platformu ile kolaylıkla farksal olarak gen ifade analizi işlemi gerçekleştirilmektedir. En anlamlı farklılık gösteren 250 gen web arayüzünde görüntülenmekte olup, kullanıcı isteğine bağlı olarak anlamlı farklılık gösteren tüm genlerin sıralanmış (p-değerlerinin artan sıralı olacak şekilde) formatının indirilmesi imkanını vermektedir. Bu işlemler, tüm veri kümelerinin sınıflarının ikili grupları baz alınarak farksal olarak ifade edilen genler GEO2R platformu ile bulunmuştur. GEO2R platformu ile bir veri kümesinin anlamlı farklılık gösteren ilk 250 geninin nasıl bulunduğu maddeler halinde anlatılarak şekillerle desteklenmiştir (Şekil 3.1 – 3.9).

a) NCBI veri tabanının GDS Browser hizmeti (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser) ile çalışmamızda kullandığımız veri kümelerinden biri olan over kanserinin kayıt numarası (GDS3592) Şekil 3.1'de gösterildiği gibi ilgili kutucuğa yazılıp "Search" butonu ile arama yaptırılmıştır. Kayıt numarasını bilmeyen kullanıcılar anahtar kelimelerle aradığı veri kümesinin ismini de yazarak arama yaptırabilirler (*Örnek:* Ovarian cancer). Birden fazla çıkabilecek sonuç içerisinden kullanıcının kendisi filtre uygulayarak (yani, organizma türünü, gen çipi platformunu, vs. belirleyerek) daha keskin seçim yapabilmektedir.



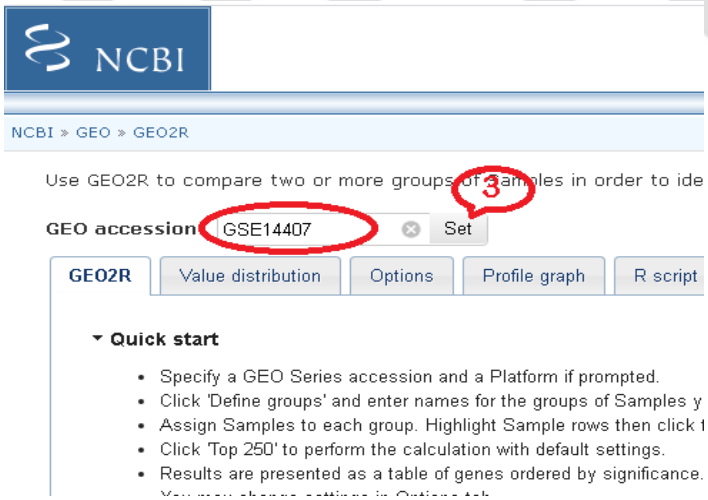
Şekil 3.1: İlgilenilen veri kümesinin kayıt numarasının girilmesi.

b) Açılan ekranda, ilgilenilen veri kümesinin "Title", "Summary", "Organism", "Platform", "Reference Series", "Sample Count", vs. gibi özellikleri hakkında bilgi edinilebilinmektedir. Burada referans seri numarasına ihtiyaç olacağı için bu değer kaydedilmelidir. BU seri numarasına tıklandığında da açılan ekrandan "Analyze with GEO2R" butonu yoluyla ilerleyen adımlardaki işlemler devam ettirilebilir.



Şekil 3.2: Referans seri numarasının görüntülenmesi.

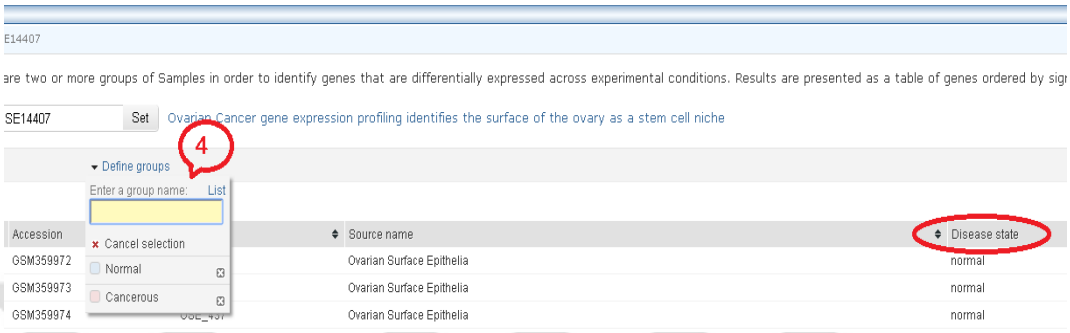
c) NCBI veritabanının GEO2R platformu kullanılarak bağlantı adresi verilen (https://www.ncbi.nlm.nih.gov/geo/geo2r/) web arayüzünden farksal olarak ifade edilen ilk 250 genin bulunması hedefleniyorsa, bağlantı adresinden açılan ekranda kullanıcının karşısına Şekil 3.3'teki gibi bir ekran gelecektir. GEO erişim kutucuğuna 2. adımda kaydedilen referans seri numarası girilip "Set" butonuna basılması gerekmektedir.



Şekil 3.3: GEO2R platfromunda referans seri numarasının girilmesi.

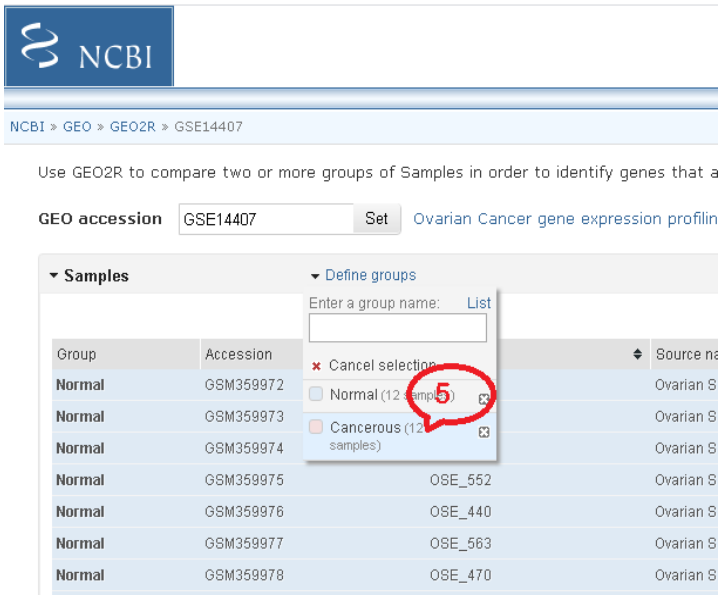
d) Bu platform farksal olarak gen ifade analizini R paket yazılımlarından biri olan *limma* kütüphanesini kullanarak gerçekleştirmektedir. Bu kütüphanenin yapısı incelendiğinde, karşılaştırılacak grupların tanımlanması gerektiği sonucu ortaya çıkmaktadır. Bu sebeple Şekil 3.4'teki ekran görüntüsünde de görülebieceği gibi "Define groups" butonu vasıtasıyla kullanıcı tanımlı grup isimleri verilmelidir. Bunun için öncelikle "Disease state" sütunundan sınıf bilgilerinin dikkate alınması gereklidir. İki sınıflı bir küme olan

bu veri kümesinde normal sağlıklı bireyler için “Normal”, kanseri olan bireyler için “Cancerous” tanımlaması yapılabilir (veya kısaca normal bireyler için “n”, kanseri olan bireyler için “t” denilebilir).



Şekil 3.4: Grupların tanımlanması.

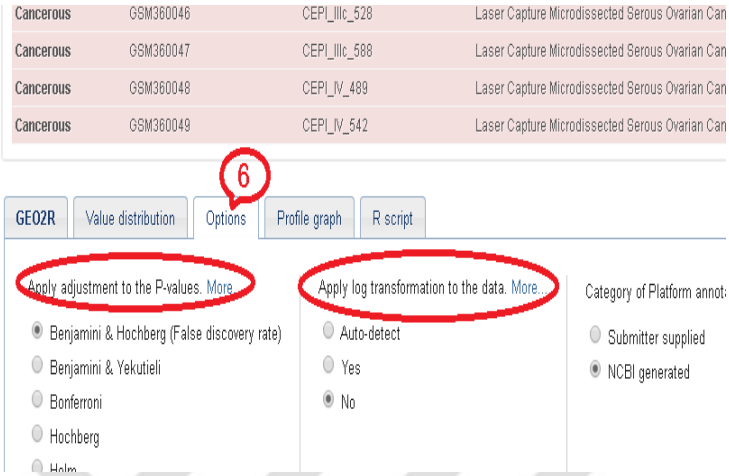
e) Şekil 3.5’te belirlenen bu gruplara örneklerin (bireylerin) atamasının nasıl yapıldığı gösterilmiştir. Analizin yapılması için her iki gruba en az birer örneğin atanması gerekmektedir. Tüm örnekler sınıfına göre seçilerek (yani, tüm normal bireyleri normal grubuna, diğerlerini cancerous grubuna atayarak) analiz devam ettirilir.



Şekil 3.5: Gruplara örneklerin atanması.

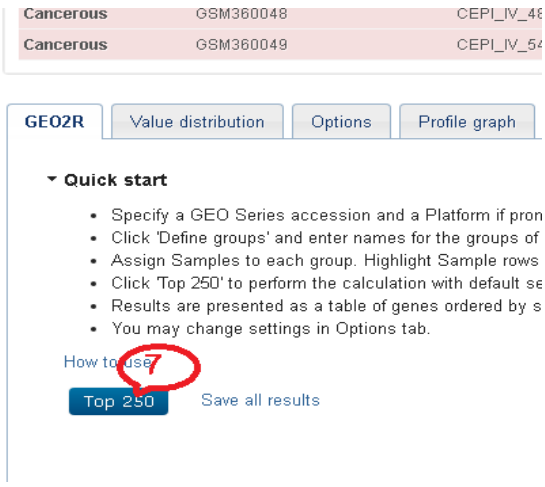
f) Şekil 3.6’da kullanıcının belirleyebileceği ayarlar sekmesi görülmektedir. Elde edilen p-değerlerine FDR düzeltmesi yöntemlerinden hangisinin uygulanacağı; veri kümesinin elemanlarına log transformasyonunun uygulanıp uygulanmayacağı, vs. gibi

ayarlamaları kullanıcı yapabilir. Tez kapsamındaki yaklaşım ile analizi yapılan veri kümesinin logaritmik dönüşümü yapılmadığı için burada da "No" seçeneği seçilerek veri kümesine bu dönüşümün uygulanmaması sağlanmıştır.



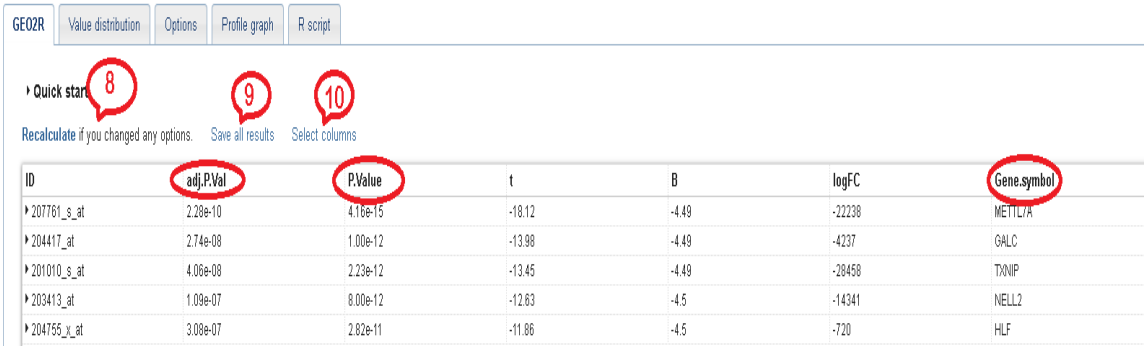
Şekil 3.6: Ayarlar (isteğe bağlı).

g) Gruplara örnek atamalarının yanında ayarların da yapılmasıyla "GEO2R" sekmesi altında bulunan "Top250" butonuna basılacaktır.



Şekil 3.7: Farksal olarak ifade edilen ilk 250 genin hesaplanması.

h) Bu adımda ilk 250 anlamlı farklılık gösteren genler listelenmiştir.



Şekil 3.8: Farksal olarak ifade edilen ilk 250 genin listelenmesi

i) "Select columns" butonu ile görüntülenmesi istenilen diğer parametrelerin de görüntülenmesi sağlanmaktadır. "P-value" ile "Adjusted.PValue" sütunları incelenirse artan sırada genlerin listelendiği görülecektir. (h-j) adımları için Şekil 3.8 incelenebilir.

j) Tüm sonuçlar istenirse (yani sadece ilk 250 gen değil genlerin tamamını içeren ve artan p-değerlerine göre sıralanmış olarak) "Save all results" butonu ile kaydedilebilir. Bu şekilde bir kayıt dosyası .txt uzantılıdır.

3.6.SINIFLANDIRMA YÖNTEMLERİ

3.6.1. KNN

Parametrik olmayan bir sınıflama yöntemi olarak kullanılan KNN algoritması, basit olmakla birlikte çok etkin sonuçlar verebilen bir yöntemdir [72]. Bu yöntem, sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak, yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılmaktadır. Örnek kümedeki gözlemlerin her birinin, sınıflandırılacak yeni gözlem değerine olan uzaklıklarının hesaplanması (Öklid uzaklık formülü kullanılabilir) ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır. Seçilen k sayıdaki gözlemin kategorileri (sınıfları) arasında en çok tekrar eden sınıf belirlenir ve bu sınıf yeni gözlemin kategorisi veya sınıfı olarak kabul edilir.

3.6.2. NB

NB (*İng.* Naive Bayes Algorithm) sınıflandırıcısının temeli Bayes teoremine ve koşullu olasılık kavramına dayanmaktadır. Daha açık ifade etmek için, X sınıflandırılacak yeni gözlem verisi olsun ($X = \{x_1, x_2, \dots, x_n\}$). Ele alınan sınıf niteliğinde ise c sınıf olduğu varsayılsın ($C = \{C_1, C_2, \dots, C_c\}$). X niteliğinin seçimine bağlı olarak C_i sınıflarının olasılığını elde etmek için Eşitlik 3.23'teki bağıntı kullanılabilir.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{3.23}$$

Hesaplamadaki işlem yükünü azaltmak amacıyla $P(X|C_i)$ olasılığı Eşitlik 3.24'teki bağıntı [72] ile basitleştirilirse (karşılaştırılacak gözlem verisi X'in değerlerinin birbirinden bağımsız olduğu kabul edilerek),

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{3.24}$$

(3.23) nolu eşitlikte $P(C_i|X)$ tanımında yer alan paydalar birbirine eşit olduğundan sadece pay değerlerini kıyaslamak yeterli olacaktır. Bu değerler içinden en büyük olanı seçilerek bilinmeyen örneğin bu sınıfa ait olduğu belirlenmiş olur.

3.6.3. SVM

SVM, sınıflandırmayı doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla yerine getirir ve veriyi birbirinden ayırmak için en uygun fonksiyonun tahmin edilmesi esasına dayanır. Temelde iki sınıfa ait verinin en uygun hiper düzlem ile tanımlanması sağlanır ve eğitim verisi olarak kullanılmakta olan karar parametrelerini tahmin eder. Veri setini optimal karar doğrusu ile ayırır ve sınır eksenlerinin iki sınıf için de çizgiye en yakın mesafede olmasına dikkat edilir. Bu eksenlere en yakın mesafelere destek bölgesi adı verilmektedir. Böylelikle veri kümesi SVM ile uygun sınıflara ayrılabilmektedir [73].

3.6.4. RF

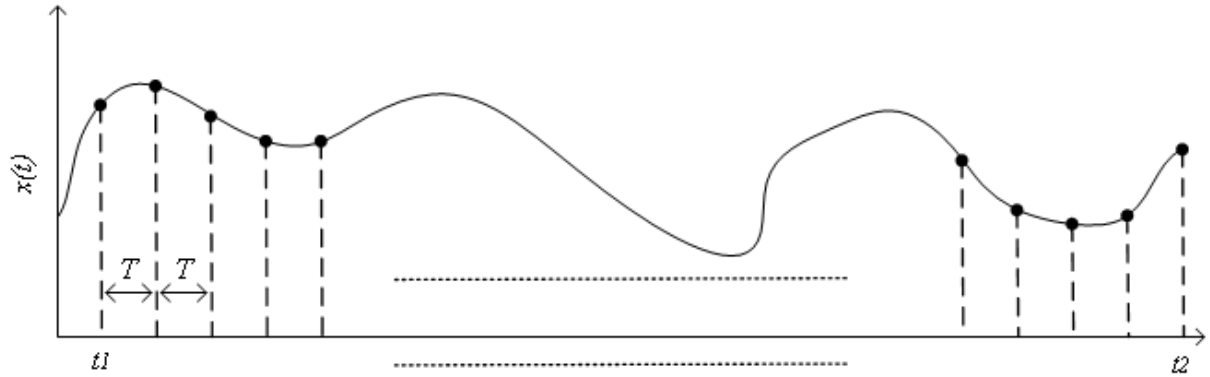
RF yönteminde, veri kümesinden bootstrap yöntemiyle örnekler seçilir ve bu örneklere dayalı olarak sınıflandırma ağaçları oluşturulur. Bu sınıflandırma ağaçları kullanılarak her bir ağaç için gözlemin sınıfı kestirilir. Sınıflandırmalar arasında en çok tekrar eden sınıf değeri seçilir [74].

4. GEN İFADE VERİLERİNİN MODELLENMESİNDE YENİ YAKLAŞIM: SYMPES

Büyük boyutlu mikrodizi gen ifade verilerinin gen çipi platformlarına bağlı olarak değişmekle birlikte çok fazla sayıda prob içerdiği bilinmektedir. Bu durumda, veriyi sıkıştırarak mümkün olduğunca az bilgi ile ifade edilmesini sağlamak ve veri kalitesini de aynı oranda korumak en temel amacımızdır. Dönüşüm yöntemleri içinde, yapılan bir çok çalışmada başarımı kanıtlanmış olan Karhunen Loeve dönüşümünün (KLD) kullanımının uygun olacağı öngörülmüştür. Bu öngörümüzün dayanakları, SYMPES yöntemi ile biyomedikal alanındaki bir çok uygulamada işaretlerin (konuşma, tıbbi görüntüleme cihazlarından elde edilen görüntü, biyopotansiyel işaretler, vs.) doğasını en iyi biçimde yansıtan ve tanımlayan ve temel tanım dizileri olarak adlandırılan özgün fonksiyonların bulunmuş olması, bu fonksiyonların gen ifade verilerinin bir dizi şeklindeki yapısı için de üretilebilir olması ve arka plandaki dönüşüm yöntemlerinin uyarlanabilir olmasıdır.

4.1. GEN İFADE VERİ KÜMELERİ İÇİN ÖNCEDEN TANIMLI ENERJİ VE BİLGİ BLOKLARININ BULUNMASI

Bu çalışmada, gen ifade verilerinin SYMPES yöntemi ile modellenmesine ilişkin yeni bir yaklaşım sunulmaktadır. Gen ifade verilerinin modellenmesi işlemine geçmeden önce SYMPES yönteminin sürekli ve ayrık zaman işaretini modelleme yaklaşımını incelemekte yarar vardır. Modellenmesi istenen tek boyutlu sürekli zaman işareti $x(t) = f(t)$ olsun.



Şekil 4.1: Sürekli işaretin örneklenerek ayrık işarete dönüştürülmesi.

SYMPES modelleme yönteminde temel fikir, $f(t)$ işaretinin $e(t)$ bilgisini ve $s(t)$ enerjisini ayırmaktır.$f(t)$ işareti belirli zaman aralığında $(T = t_2 - t_1)$ iki ayrı sinyalin $(e\,(t), s\,(t))$ çarpımı olarak Eşitlik 4.1'deki gibi modellenebilir.

$$f(t) = C * e(t) * s(t); \quad t_1 \leq t \leq t_2 \,; T = t_2 - t_1 \tag{4.1}$$

Eşitlik 4.1'de, $s(t)$, orijinal işaretin taşıyıcı kısmını temsil eder ve Eşitlik 4.2'de gösterildiği gibi T aralığı için hatayı (δ) minimum yapacak en küçük ortalama kareler (*İng.* least mean squares, LMS) anlamında $f(t)$'ye en uygun ortonormal fonksiyonlar arasından seçilir.

$$\delta = \frac{1}{T}\int_{t_1}^{t_2} [f - Cs(t)]^2 dt \tag{4.2}$$

Dolayısıyla, $s(t)$, $f(t)$'nin imzası veya enerji kısmı olarak adlandırılır.

Reel sabit olan C, $s(t)$ imza (veya enerji) fonksiyonunun şiddeti veya genliği olarak ifade edilir. Dolayısıyla, C sabiti, $f(t)$'nin kazancı veya eş anlamda ölçekleme katsayısı olarak adlandırılır (Eşitlik 4.3).

$$C = \int_{t_1}^{t_2} f(t)s(t)dt \,; \quad t_1 \leq t \leq t_2 \,; T = t_2 - t_1 \tag{4.3}$$

$s(t)$ ve C bileşenleri belirlendikten sonra $e(t)$ fonksiyonu Eşitlik 4.4'teki gibi oluşturulur.

$$e(t) = \frac{f(t)}{Cs(t)}; \quad t_1 \leq t \leq t_2 \,; T = t_2 - t_1 \tag{4.4}$$

Eşitlik 4.4'ten görüleceği üzere, $e(t)$, enerji fonksiyonu ve kazancı olmaksızın $f(t)$'yi temsil eder. Bu nedenle, $f(t)$'ye bağlı olduğu için, zarf veya bilgi fonksiyonu olarak adlandırılır.

Ayrık zaman domeninde $f(t)$ işareti T zaman aralığında bir zaman serisi ($f_i = \{f(t_i); i = 1,2, \dots, n\} = \{f_1, f_2, \dots, f_n\}$) veya bir vektör formunda Eşitlik 4.5'teki gibi tanımlanabilir.

$$[f]^T = [f_1, f_2, \dots, f_n] \tag{4.5}$$

$[f]$ sütun vektörü, n-boyutlu ortogonal vektör uzayına genişletilebilir (Eşitlik 4.6).

$$[f] = \sum_{i=1}^{n} c_i [V_i] \tag{4.6}$$

(4.6) nolu eşitlikte verilen serinin ilk teriminden sonraki terimleri kesilebilir. Böylece Eşitlik (4.7)'de verilen ifadeye yaklaşılır.

$$[f] = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} \cong c_1[V_1] = c_1 \begin{bmatrix} v_1^{(1)} \\ \vdots \\ v_n^{(1)} \end{bmatrix} \tag{4.7}$$

V_i ortonormal vektörleri Eşitlik (4.8)'de tanımlanan öz değer probleminden elde edilir.

$$[R][V_i] = \lambda_i[V_i] \tag{4.8}$$

Burada, $[R]$, $\{[f][f]^T\}$'nin beklenen değerinin özilinti matrisidir ve λ_i, $[V_i]$ vektörünün öz değeridir. Diğer bir ifade ile,

$$[R] = E\{[f][f]^T\} \tag{4.9}$$

En büyük öz değer λ_{max} ve bunun öz vektörü $[s] = [V_i]$ olsun. Bu durumda (4.10) nolu eşitlik elde edilir.

$$[R][s] = \lambda_{max}[s] \tag{4.10}$$

Sonra, $[f]$, LMS anlamında Eşitlik 4.11'deki gibi yaklaştırılabilir.

$$[f] = \begin{bmatrix} f_1 \\ \dots \\ f_n \end{bmatrix} \cong C[s] = C \begin{bmatrix} s_1 \\ \dots \\ s_n \end{bmatrix} \tag{4.11}$$

Burada, c, reel bir sabit olup Eşitlik 4.12'den elde edilir.

$$C = [f]^T[s] \tag{4.12}$$

Bu durumda, zarf (veya bilgi) vektörünün girişleri ($[e]^T = [e_1 \dots e_n]$) (4.13) nolu eşitlikten elde edilir.

$$e_i = \frac{f_i}{Cs_i} \tag{4.13}$$

Sonuç olarak, bir-boyutlu SYMPES Eşitlik 4.14'teki gibi ifade edilebilir.

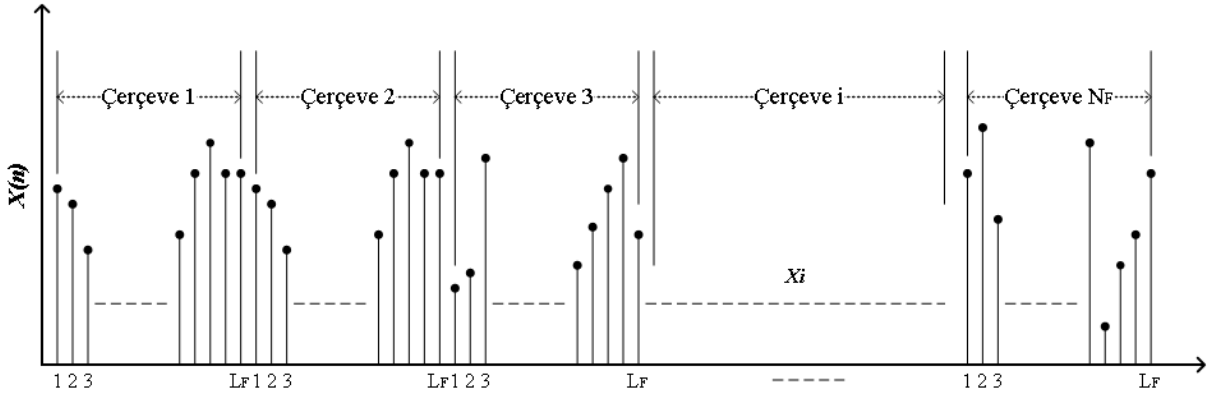
$$[f] = C[e_d][s] \tag{4.14}$$

Burada,

$$[e_d] = \begin{bmatrix} e_1 & 0 & \cdots & 0 \\ 0 & e_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & e_n \end{bmatrix} = diag[e] \tag{4.15}$$

Eşitlik 4.14'te verilen ifade, Eşitlik 4.1'de verilen ifadenin ayrık versiyonudur.

Affymetrix gen çipleri kişilerden temin edilen ve kişilere ilişkin gen ihtiva eden sıvı örnekler üzerinde çalışır. Her gen, adına hibridizasyon dediğimiz bir prosese tabi tutularak parlaklık düzeyine göre cihaz tarafından tanımlanır. Başka bir deyişle, cihaz tarafından art arda dizilen genlerin yaklaşık 55.000 gene ilişkin parlaklık düzeyleri bir çıktı olarak cihaz kullanıcısına sunulur. Buna göre, kullanıcı cihaz tarafından dizilimi sabitlenen genleri teker teker $N = 1,2,\ldots,54675$'e kadar örneklenmiş olduğunu kabul eder ve her gene ilişkin parlaklık düzeyi kayıt altına alınır. Böylesi bir yaklaşımda, gen dizisindeki genler n ile ifade edilirken, bunlara karşılık gelen parlaklık düzeyleri $X(n)$ ile ifade edilir. Böylece, $X(n) = f(n)$ gibi ayrık zaman domeninde bir fonksiyon olarak ifade edilmiş olur. Böylece SYMPES modelleme yöntemi doğrudan $X(n) = f(n)$ üzerinde uygulanır. SYMPES yöntemi uygulanmadan önce fonksiyonun minimum ve maksimumu alınarak, gen ifade verileri 0 ila 1 aralığında normalize edilir. Enerji ve bilgi bloklarının oluşturulması aşamasında, ilk olarak, gen ifade yoğunluklarını içeren veri kümesi dizi bazında çerçevelere (veya bloklara) ayrılmaktadır (Şekil 4.2).



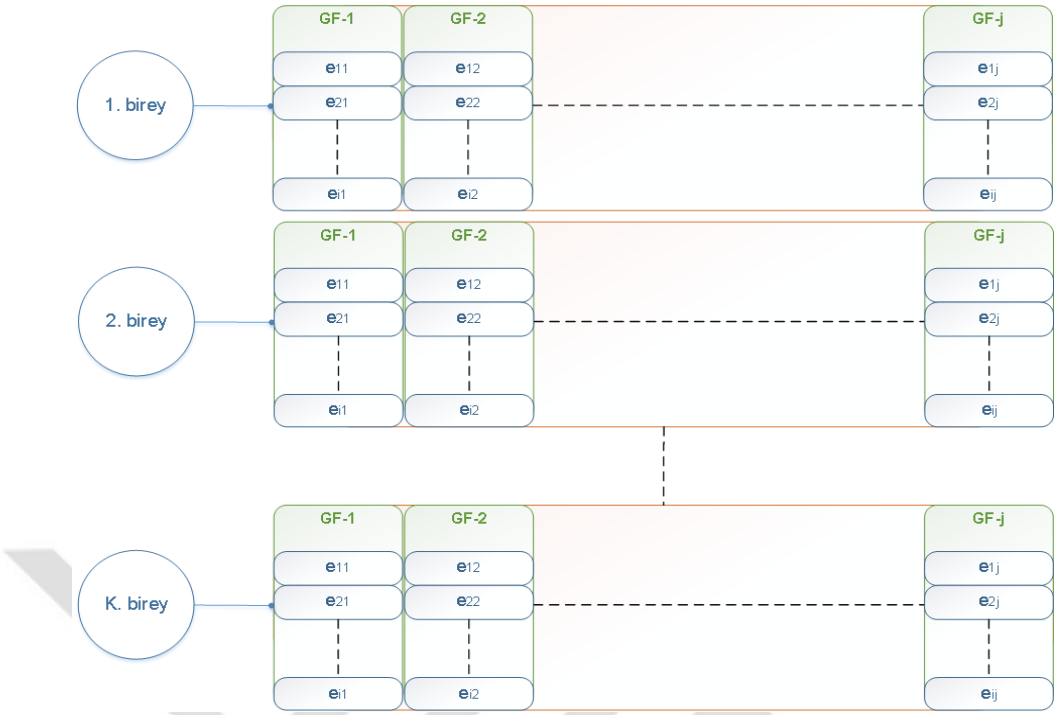
Şekil 4.2: Bir diziye ait gen ifade verisinin çerçevelere bölünmesi.

Bu veri kümesi aslında $N \times K$ boyutunda, ön işlemden geçtikten sonraki değerlerinin 0 ila 1 arasında gerçek değerlerden oluştuğu bir veri matrisidir. N ve K değerleri, sırasıyla prob ve birey sayısını gösteren satır ve sütunlardır. Her bir bireye ait gen ifade verisinin $N \times 1$ boyutunda satır vektörü olduğu göz önünde bulundurularak, matematiksel model için giriş işareti olan gen ifade verisi öncelikle Şekil 4.3'te gösterildiği gibi $GF_J^{(k)}$ $(j = 1,2, \dots, N_F; k = 1,2, \dots, K)$ olarak adlandırılan bloklara (gen çerçeveleri, GF) ayrılmaktadır. Her bir GF, $L_F \times 1$ boyutunda bir satır vektörüdür. Burada, L_F çerçeve uzunluğu olmak üzere, N_F çerçeve sayısı, $N_F = N/L_F$ olarak belirlenmektedir. Şekilden de görüleceği üzere, herhangi bir bireye ait gen ifade verisi işaretinin GF bazında elemanları ${e_{i,j}}^{(k)}$, $(i = 1,2, \dots, L_F;\ j = 1,2, \dots, N_F; k = 1,2, \dots, K)$ olarak ifade edilebilir. Her bir GF çerçeve matrisi ($GF_J^{(k)}$), alt çerçeve vektörleri cinsinden (4.16) nolu eşitlikteki gibi ifade edilebilir.

$$GF_J^{(k)} = \begin{bmatrix} {e_{(j-1)L_F+1}}^{(k)} \\ {e_{(j-1)L_F+2}}^{(k)} \\ \vdots \\ {e_{(j-1)L_F+L_F}}^{(k)} \end{bmatrix} \quad , j = 1,2, \dots, N_F; k = 1,2, \dots, K \tag{4.16}$$

Tüm veri kümesinin bloklarına ayrılması işlemi sonunda elde edilen gen ifade matrisi (4.17) nolu eşitlikteki gibi yazılabilir.

$$GF^{(k)} = \begin{bmatrix} GF_1^{(k)} & GF_2^{(k)} & \cdots & GF_{N_F}^{(k)} \end{bmatrix} \tag{4.17}$$



Şekil 4.3: Tüm veri kümesine ait gen ifade verisinin bloklarına ayrılması ile vektörel formda yeniden düzenlenmesinin gösterimi.

Yukarıda da açıklandığı gibi, önerilen yöntemdeki matematiksel model ölçekleme sabiti, işaretin enerji ve bilgi bloklarının çarpımından oluşmaktadır. Çalışmada, uzunluğu L_F olan, herhangi j. nci GF'nin yaklaşık olarak değeri (4.18) nolu eşitlikteki gibi ifade edilebilir.

$$GF_j \cong G_j E_{IE} S_{IS} \qquad ,(j = 1,2, \dots, N_F) \tag{4.18}$$

Burada G_j, gerçek sabit olup, GF_j'nin kazanç faktörü, diğer bir ifade ile ölçekleme sabitidir. $IE \in \{1,2, \dots, N_{IE}\}$, önceden tanımlı bilgi bloğunun (*İng.* Predefined Envelope Block, PEB) indis sayısı ve $IS \in \{1,2, \dots, N_{IS}\}$ önceden tanımlı enerji bloğunun (*İng.* Predefined Signature Block, PSB) indis sayısıdır. N_{IE} ve N_{IS} sırasıyla PEB ve PSB'nin toplam indis sayıları olup, IE, IS, N_{IE} ve N_{IS} değerleri birer tamsayıdır.

Vektör formundaki PSB, gen ifadesinin ve temel olarak enerji karakteristiğini içeren GF'nin geniş anlamda incelenmesi ile elde edilen yoğunluk bilgisinden faydalanılarak oluşturulmuştur. Bu durum matematiksel olarak (4.19) nolu eşitlikteki gibi belirtilebilir.

$$[S_{IS}]^T = [S_{IS,1} \quad S_{IS,2} \quad \cdots \quad S_{IS,L_F-1} \quad S_{IS,L_F}] \tag{4.19}$$

$PSB = G_j S_{IS}$ niceliği LMS anlamında neredeyse GF_j'nin en yüksek enerjisini taşımaktadır. Bu çarpımda, G_j sabiti olarak ifade edilen katkı GF_j'nin yoğunluk seviyesinin ölçeklenmesini sağlayan ölçüttür.

$$G_j = [PSB]^T S_{IS} \tag{4.20}$$

PEB, $(L_F \times L_F)$ köşegen matrisi olacak şekilde vektörel olarak (4.21) nolu eşitlikteki gibi ifade edilmektedir.

$$E_{IE} = diag[E_{IE,1} \quad E_{IE,2} \quad \cdots \quad E_{IE,L_F-1} \quad E_{IE,L_F}] \tag{4.21}$$

E_{IE} , PEB'nin üretiminde gen ifadesi blok verisini gözden geçirerek ayırt edici özelliklerini yansıtmaktadır.

Ortonormal vektörler, $\{\varphi_{jk}\}$, tarafından oluşturulan vektör uzayında her GF dağınık şekilde olabilir. Gerçel ortonormal vektörler transpozu alınmış dönüşüm matrisinin, (Φ_j^T), sütunları olsun. Bu durumda,

$$\Phi_j^T = [\varphi_{j1} \quad \varphi_{j2} \quad \cdots \quad \varphi_{j(L_F-1)} \quad \varphi_{jL_F}] \tag{4.22}$$

olduğunda, açıktır ki,

$$GF_j = \Phi_j^T G_j \tag{4.23}$$

Burada, $G_j = [g_1 \quad g_2 \quad \cdots \quad g_{L_F-1} \quad g_{L_F}]$ olacaktır. $\Phi_j^T = \Phi_j^{-1}$ özelliğinden sırasıyla $GF_j \Phi_j = \Phi_j \Phi_j^{-1} G_j$ eşitlikleri elde edilebilir. Sonrasında, GF_j bu ortonormal vektörlerin ağırlıklı toplamları olarak yazılabilir.

$$GF_j = \sum_{k=1}^{L_F} g_k \varphi_{jk}\,, \quad k = 1,2,\dots,L_F \tag{4.24}$$

(4.24) nolu eşitlikten, GF' lerin katsayıları oluşturulabilir.

$$g_k = GF_j \varphi_{jk}^T \qquad ,k = 1,2,\dots,L_F \tag{4.25}$$

$GF_{jt} = \sum_{k=1}^{t} g_k \varphi_{jk}$, GF_j' nin $1 \le t \le L_F$ olacak şekilde kesilmiş biçimi olsun. Dikkate alınmalıdır ki eğer, $t = L_F$ ise $GF_{jt} = GF_j$ olacaktır. Bu durumda yaklaşık hata (4.26) nolu eşitlikteki gibi olacaktır.

$$\varepsilon_t = GF_j - GF_{jt} = \sum_{k=t+1}^{L_F} g_k \varphi_{jk} \tag{4.26}$$

Bu denklemde, φ_{jk} hatanın beklenen değerini φ_{jk}' ya bağlı olarak, LMS anlamında en aza indirecek şekilde belirlenir.Yukarıda da belirtildiği gibi LMS süreci takip eden özvektör probleminin sonucudur [75], [76]. Sonuç olarak, φ_{jk}' lar GF_j' nin özilinti matrisi, R_j' nin özvektörleri olarak hesaplanır. Ortonormallik durumunu kullanarak, LMS hatası,

$$\varepsilon_t \varepsilon_t^T = \sum_{k=t+1}^{L_F} g_k^2 \tag{4.27}$$

$$J_t = E[\varepsilon_t \varepsilon_t^T] = \sum_{k=t+1}^{L_F} E[g_k^2] \tag{4.28}$$

olarak verilir.

J_t, toplam kareler hatasının $\varepsilon_t \varepsilon_t^T$ beklenen değerini belirtsin. O halde,

$$E[c_k^2] = E\left[\varphi_{jk}^T \left(GF_j^T GF_j\right)\varphi_{jk}\right] = \varphi_{jk}^T R_j \varphi_{jk} \tag{4.29}$$

Burada, $R_j = E[(GF_j^T GF_j)]$, GF_j' nin özilinti matrisi olarak tanımlanmıştır. En uygun dönüşümün olması amacıyla istenen, verilen t için, J_t' yi en aza indiren ortonormallik kısıtlamasına uygun olacak φ_{jk}' yı bulmaktır. En aza indirgemeyi Lagrange çarpanı λ_k kullanarak sağlanır.

$$J_t = \sum_{k=t+1}^{L_{GE}} [\varphi_{jk}^T R_j \varphi_{jk} - \lambda_k(\varphi_{jk}^T \varphi_{jk} - 1)] \tag{4.30}$$

(4.30) nolu eşitliğin gradyanı alınarak (4.31), (4.32) ve (4.33) nolu eşitlikler yazılabilir.

$$\frac{\partial J_t}{\partial \varphi_{jk}} = \frac{\partial}{\partial \varphi_{jk}} \sum_{k=t+1}^{L_F} [\varphi_{jk}^T R_j \varphi_{jk} - \lambda_k(\varphi_{jk}^T \varphi_{jk} - 1)] \tag{4.31}$$

$$2R_j \varphi_{jk} - 2\lambda_k \varphi_{jk} = 0 \tag{4.32}$$

$$R_j \varphi_{jk} = \lambda_k \varphi_{jk} \tag{4.33}$$

R_j özilinti matrisidir. R_j gerçel, diyagonal elemanlarının simetrik olmasına bağlı olarak yarı kesin, pozitif ve Toeplitz matristir.

$$R_j = \begin{bmatrix} r_j(1) & r_j(2) & r_j(3) & \cdots & r_j(L_F) \\ r_j(2) & r_j(1) & r_j(2) & \cdots & r_j(L_F - 1) \\ r_j(3) & r_j(2) & r_j(1) & \cdots & r_j(L_F - 2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_j(L_F) & r_j(L_F - 1) & r_j(L_F - 2) & \cdots & r_j(1) \end{bmatrix} \tag{4.34}$$

$$r_j(d+1) = \frac{1}{L_F} \sum_{q=[(j-1)L_F+1]}^{[jL_F-d]} x_q x_{q+d} \quad , d = 0,1, \dots, (L_F - 1) \tag{4.35}$$

λ_{jk} ve φ_{jk} gözden geçirilen problemin özdeğer ve özvektörleridir. R_j' nin özdeğerleri de gerçektir, ayrıktır (değerleri birbirine eşit değildir) ve pozitiftir. Daha da fazlası φ_{jk}' nın özdeğerlerinin tamamı ortonormaldir. KLD yöntemine göre, özilinti matrisi, R_j' nin özdeğerleri olan $\lambda_{1j}, \lambda_{2j}, \dots, \lambda_{L_F j}$ azalan sırada ($\lambda_{1j} \geq \lambda_{2j} \geq \cdots \geq \lambda_{L_F j}$) sıralanmakta ve bu özdeğerlere karşılık gelen özvektörler belirlenmektedir.

$$GF_j^T GF_j = \sum_{k=1}^{L_F} g_{jk}^2 = \sum_{k=1}^{L_F} \lambda_{jk} \tag{4.36}$$

(4.36) nolu eşitlik, GF_j' nin en yüksek enerjisine sahip ilk p tane temel bileşenin alınmasıyla aşağıda gösterildiği gibi kesilebilir.

$$GF_j \cong \sum_{k=1}^{p} g_k \varphi_{jk} \tag{4.37}$$

$p = 1$ iken Eşitlik 4.37'nin basit hali (4.38) nolu eşitlik biçiminde oluşacaktır. Bu durumda özvektör φ_{jk} enerji vektörü adını alacaktır. Kısaca, enerji vektörünün, LMS anlamında yaklaşık en yüksek enerjiye sahip, GF_j' ye ait tüm blokların, enerji vektörü olduğu söylenebilmektedir.

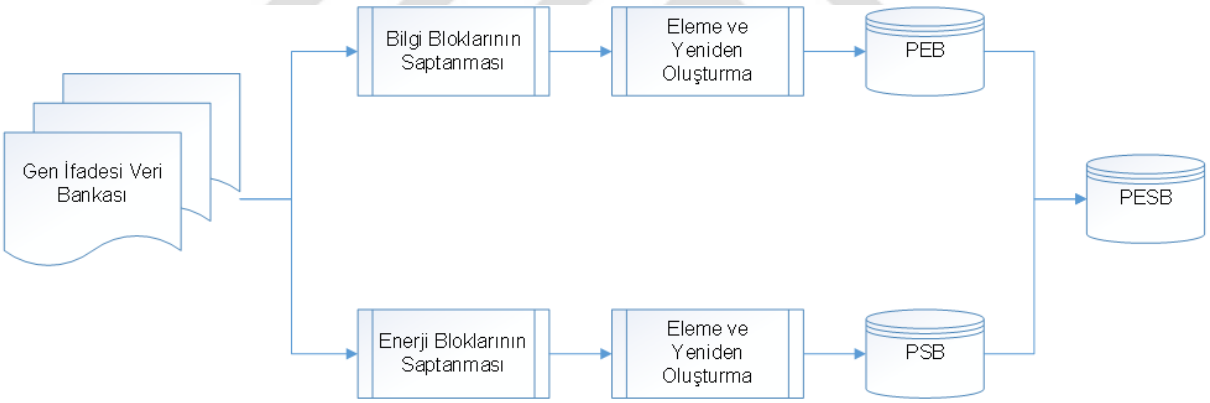
$$GF_j \cong g_1 \varphi_{j1} \tag{4.38}$$

Bu durumda, L_F, neredeyse tüm enerjinin (4.38) nolu eşitlikte gösterildiği gibi ilk terimlerde toplayacak kalanlarının ihmal edileceği bir katsayı olarak ayarlanabilir. Bu nedenle enerji vektörü olarak adlandırılan φ_{j1}, orijinal işaretin gözden geçirilmesiyle oluşan GF'nin en değerli bilgisini taşımaktadır.

Bu çalışma kapsamında, sayısı binlerle ifade edilebilecek düzeyde enerji ve bilgi blokları oluşturulmuş; bu blokların tekrarlayan benzerlikler sergilediği görülmüştür. Bu durumda, benzeyen bilgi ve enerji blokları, önceden tanımlı bilgi ve enerji blokları şeklinde oluşturulmuştur. Eleme işlemi için, Pearson özilinti katsayısı (*İng.* Pearson Correlation Coefficient, PCC) kullanılmıştır (Eşitlik 4.39). PCC'ye ait eşitlik aşağıdaki gibidir:

$$\rho_{XY} = \frac{\frac{\sum_{i=1}^{L} x_i y_i - [\sum_{i=1}^{L} x_i \sum_{i=1}^{L} y_i]}{L}}{\sqrt{\left[\frac{\sum_{i=1}^{L} x_i^2 - (\sum_{i=1}^{L} x_i)^2}{L}\right]\left[\frac{\sum_{i=1}^{L} y_i^2 - (\sum_{i=1}^{L} y_i)^2}{L}\right]}} \tag{4.39}$$

Karşılaştırılan iki dizi $X = [x_1 \quad x_2 \quad \cdots \quad x_L]$ ve $Y = [y_1 \quad y_2 \quad \cdots \quad y_L]$ olsun. Eğer $0.9 \leq \rho_{XY} \leq 1$ ise, bu iki dizinin neredeyse özdeş olduğu varsayılır. Böylelikle, benzer enerji ve bilgi blokları bu doğrultuda elenmiş olacaktır.



Şekil 4.4: PESB oluşturma basamakları akış şeması.

Bu bölüme başlarken de ifade edildiği gibi, önceden tanımlı bilgi ve enerji blokları olarak adlandırılan PEB ve PSB'nin oluşturulması süreci akış şeması biçiminde Şekil 4.4'te gösterilmiştir. Tez kapsamında çalışılan iki önemli konudan biri olan farksal olarak gen ifade analizi sürecinde, PEB vektörlerinin yalnız başına ilgili analiz için yüksek başarımlı sonuçlar vermesinin yanında işlemsel yükün azaltılması anlamında da katkısının olduğu gözlemlenmiş, tüm veri kümeleri için farksal olarak ifade edilen genlerin tespitinde bu vektörler kullanılmıştır.

Modellenecek yeni bir gen ifade verisinin (PESB'nin oluşumu için kullanılmamış) enerji ve bilgi bloklarının yapılandırılması süreci "kodlama" (*İng.* encoding) ve "kod çözme" (*İng.* decoding) adı verilen iki süreçten oluşmaktadır. Tez kapsamındaki ikinci önemli çalışma

konusu olan gen ifade verilerinin sınıflandırılması sürecinde, sınıflandırılacak gen ifadesi verisi için en iyi enerji ve bilgi blokları ile ölçekleme katsayısı bileşenleri hesaplanmaktadır. Bu bileşenlerin PESB'ye gönderilmesi sonrasında yeni bir gen ifade verisi en benzer yeni bileşenlerle çerçeve çerçeve oluşturulmaktadır. En az iki sınıflı olan mikrodizi verilerinde her bir sınıf için iki algoritma da kullanılarak önceden tanımlı bilgi ve enerji blokları oluşturulmaktadır. Bu blokların her biri için yukarıda bahsedilen en benzer bileşenlerle oluşturulan yeni gen ifade verileri modellenecek orijinal gen ifade verisi ile Öklid mesafesi hesaplanarak komşuluk derecesi hesaplanmaktadır. Yakın komşunun sınıfına dahil edilmek üzere süreç diğer modellenecek tüm veriler için aynı şekilde yürütülmektedir. Kodlama ve kod çözme algoritmalarının aşamaları izleyen bölümlerde; bu algoritmalara ait akış şemaları ise sırasıyla Şekil 4.5 ve Şekil 4.6'da verilmiştir.

4.1.1. SYMPES Kodlama Algoritmasının Aşamaları

Girişler:

Kodlanacak gen ifade verisi ana çerçeve vektörü, GF.

Çerçeve uzunluğu, L_F.

Sayısal Adımlar:

Adım 1: Toplam çerçeve sayısının hesaplanması, $N_F = N/L_F$.

Adım 2: Ana çerçeve vektörünün çerçeve vektörlerine bölünmesi, GF_j.

Adım 3: Her bir GF_j için R_j özilinti matrisinin hesaplanması.

Adım 4: Her bir R_j için, özdeğerlerinin (λ_{jk}) ve karşılık gelen özvektörlerinin (φ_{jk}) hesaplanması.

Adım 5a: Büyükten küçüğe doğru sıralanmış özdeğerlerden en büyük özdeğere λ_{jr}, $\lambda_{jr} = \max\{\lambda_{j1}, \lambda_{j2}, \dots, \lambda_{jL_F}\}$ karşı gelen özvektörün (S_{jr}) belirlenmesi ve belirlenen özvektörün çerçeve indisinin S_{j1} olarak değiştirilmesi.

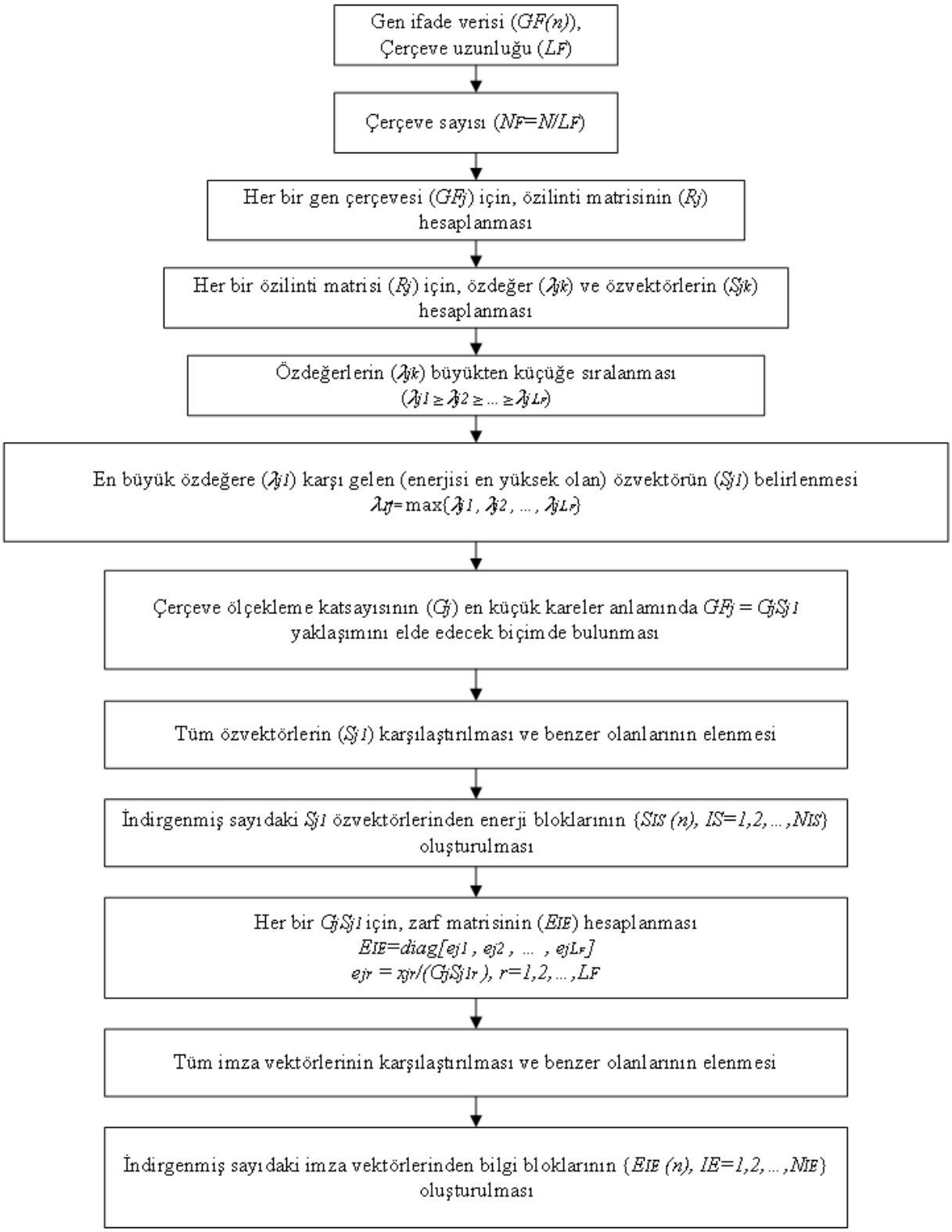
Adım 5b: Çerçeve ölçekleme katsayısının (G_j) en küçük kareler anlamında $GF_j \cong G_j S_{j1}$ yaklaşımını elde edecek biçimde bulunması.

Adım 6: Adım 5'in tüm çerçeveler için ($j = 1,2, \dots, N_F$) gerçeklenmesi. (Not: Bu adımın sonunda, her bir çerçeve için en yüksek enerjili özvektörler elde edilmiş olunmaktadır.)

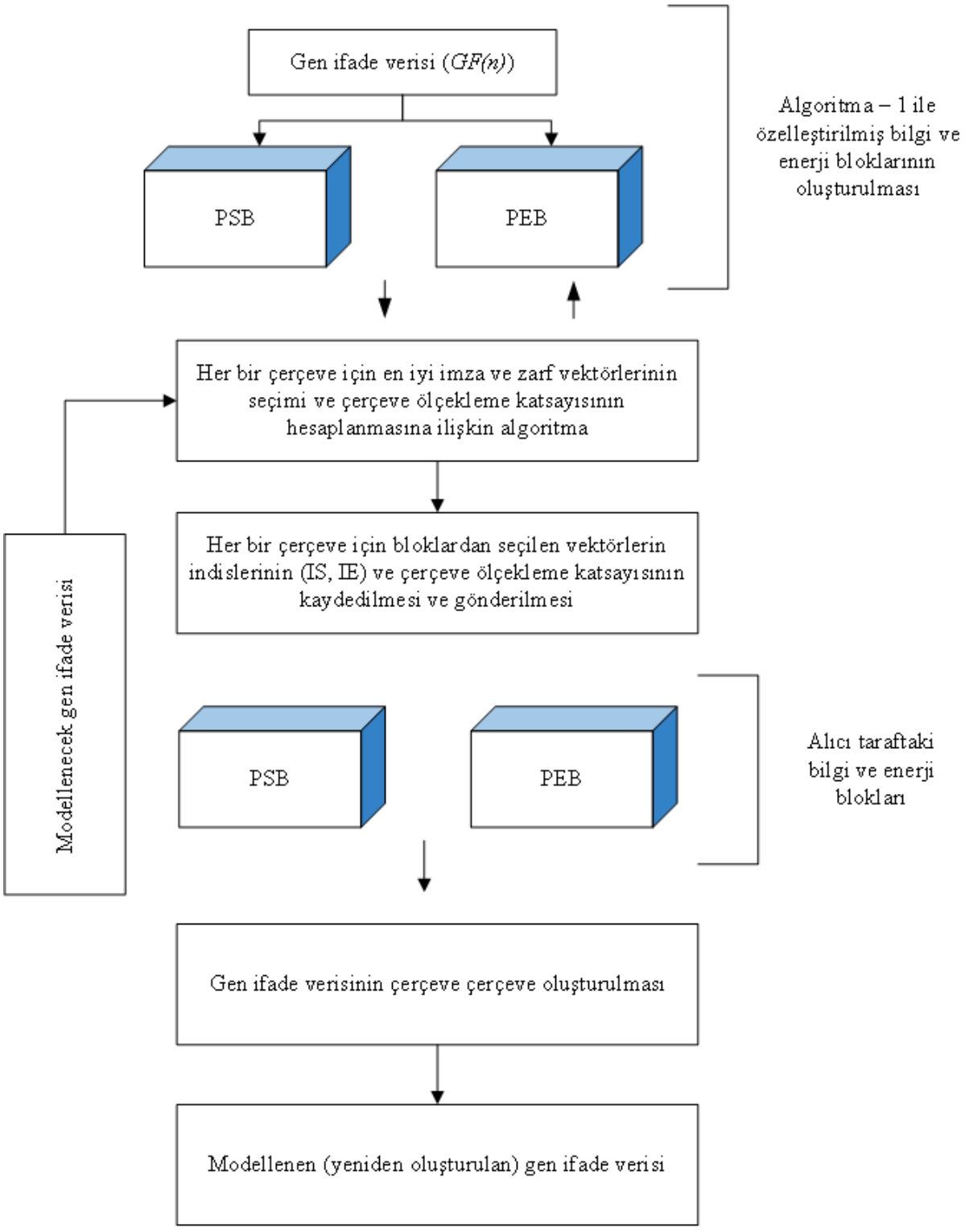
Adım 7: Adım 6'da elde edilmiş tüm özvektörlerin hızlı ve verimli bir algoritma ile karşılaştırılması ve benzer özvektörlerin elenmesi. (Eleme işlemi için Pearson özilinti sabiti kullanılmıştır.) İndirgenmiş sayıdaki S_{j1} özvektörlerinden önceden tanımlı enerji bloklarının $\{S_{IS}(n); IS = 1,2, \dots, N_{IS}\}$ oluşturulması. Burada N_{IS}, enerji bloklarındaki özgün enerji vektörlerinin toplam sayısını göstermektedir.

Adım 8: Her bir çerçeve veya Adım 5b'de hesaplanan her bir $G_j S_{j1}$ için köşegenleştirilmiş bilgi (veya zarf) matrisinin köşegen elemanlarının $e_{jr} = \frac{x_{jr}}{G_j S_{j1r}}; r = 1,2, \dots, L_F$ eşitliği yardımıyla hesaplanması.

Adım 9: Adım 7 ile benzer biçimde, elde edilen tüm bilgi vektörlerinin hızlı ve verimli bir algoritma ile karşılaştırılması ve benzer bilgi vektörlerin elenmesi. İndirgenmiş sayıdaki özgün dizilerden bilgi bloklarının $\{E_{IE}(n); IE = 1,2, \dots, N_{IE}\}$ oluşturulması. Burada N_{IE}, bilgi bloklarındaki özgün bilgi vektörlerinin toplam sayısını göstermektedir. Her bir bireye ait bilgi ve enerji blokları yukarıdaki adımlar tekrarlanarak üretilir.



Şekil 4.5: SYMPES kodlama süreci akış şeması.



Şekil 4.6: SYMPES kod çözme süreci akış şeması.

4.1.2. SYMPES Kod Çözme Algoritmasının Aşamaları

Girişler:

Modellenecek gen ifade verisi ana çerçeve vektörü, $GF(n)$.

Çerçeve uzunluğu, L_F.

Kodlama algoritmasının kullanılması ile oluşturulmuş önceden tanımlı enerji (PSB) ve bilgi (PEB) blokları.

Sayısal Adımlar:

Adım 1: Modellenecek gen ifade verisinin ($GF(n)$) kodlama algoritmasında olduğu gibi çerçeve vektörlerine bölünmesi.

Adım 2a: Her bir çerçeve için, özgün ve yaklaştırılmış çerçeve vektörleri arasındaki uzaklığı veya hatayı ($\delta_{\widetilde{IS}} = \left\|GF_j - G_{\widetilde{IS}}S_{\widetilde{IS}}\right\|^2, \widetilde{IS} = 1,2,\dots,IS,\dots,N_{IS}$) en aza indiren uygun bir imza vektörünün önceden tanımlı enerji bloklarından çekilmesi. (Bu adım S_{IS}'nin IS indeksini verir. Bu durumda, $\delta_{IS} = min\left\{\left\|GF_j - G_{\widetilde{IS}}S_{\widetilde{IS}}\right\|^2\right\} = \left\|GF_j - G_{IS}S_{IS}\right\|^2$ olur.)

Adım 2b: S_{IS}'e karşılık gelen IS indis numarasının kaydedilmesi. (Bu durumda, $GF_j \approx G_{IS}S_{IS}$ olur.)

Adım 3a: Hatayı en aza indirecek uygun bir zarf dizisi (veya köşegen bilgi matrisi) E_{IE}'nin önceden tanımlı bilgi bloklarından çekilmesi. (Bu adım E_{IE}'nin IE indeksini verir. Bu durumda, $\delta_{IE} = min\left\{\left\|GF_j - G_{IS}E_{\widetilde{IE}}S_{IS}\right\|^2\right\} = \left\|GF_j - G_{IS}E_{IE}S_{IS}\right\|^2$ olur.)

Adım 3b: E_{IE}'e karşılık gelen IE indis numarasının kaydedilmesi. Bu adımın sonunda, uygun vektörlerin seçilmesiyle en iyi enerji ve bilgi dizileri bulunmuş olmaktadır. Bu durumda, $GF_j \approx G_{IS}E_{IE}S_{IS}$ ifadesi, j. nci GF'nin E_{IE} ve S_{IS} örüntüleri ile en iyi şekilde tanımlanabileceğini göstermektedir.)

Adım 4: Elde edilen sabitlenmiş E_{IE} ve S_{IS} vektörlerini kullanarak, bir bütünsel en küçük (global minimum) hata değeri bulacak biçimde yeni çerçeve katsayılarının ($G_j = (E_{IE}S_{IS})^T GF_j/(E_{IE}S_{IS})^T E_{IE}S_{IS}$) hesaplanması. G_j çerçeve katsayılarıGF_j ile $G_{IS}E_{IE}S_{IS}$

vektörleri arasındaki uzaklığı LMS anlamında en aza indirir. Bu durumda, global minimum hata $\delta_{global} = \left\| GF_j - G_j E_{IE} S_{IS} \right\|^2$ eşitliği ile ifade edilir.

Adım 5: Yukarıdaki tüm adımlar, modellenecek gen ifade verisinin her bir çerçevesinin oluşturulması için sırayla yinelenir.

4.2.ÖNCEDEN TANIMLI BİLGİ BLOKLARINDAN EVRENSEL ORTALAMA VE ORTANCA BİLGİ KÜTÜPHANELERİNİN OLUŞTURULMASI

Önceden tanımlı bilgi bloklarının kullanımına karar verilmesinin ardından, her bir sınıf veya grup için oluşturulan evrensel bilgi kütüphanesinin ortalama, varyans ve ortanca değerlerinden oluşan $L_F \times N_F$ boyutlu matrisler oluşturulmuştur. c. sınıfa ait evrensel ortalama bilgi kütüphanesinin oluşturulmasında kullanılan Eşitlik 4.40'da verilmiştir.

$$\mu_{e_{i,j,c}}^{(k)} = \frac{1}{N_{s,c}} \sum_{j=1}^{N_F} \sum_{i=1}^{L_F} \sum_{k=1}^{N_{s,c}} e_{i,j,c}^{(k)} \tag{4.40}$$

(4.40) nolu eşitlikte, c farklı grupları (H,T1,T2,T3,vs.) temsil etmekte olup, $N_{s,c}$ parametresi c. gruptaki birey sayısını ifade etmektedir. Yine bu eşitlikte verilen ifade ile oluşturulan kütüphanenin matris formu Eşitlik 4.41'deki gibidir.

$$\mu_{UE,c} = \begin{bmatrix} \mu_{e_{1,1,c}} & \cdots & \mu_{e_{1,N_F,c}} \\ \vdots & \ddots & \vdots \\ \mu_{e_{L_F,1,c}} & \cdots & \mu_{e_{L_F,N_F,c}} \end{bmatrix} \tag{4.41}$$

$\mu_{UE,c}$ matrisine çalışmalarımızda iki grup içeren veri kümelerinde meanUE_H ve meanUE_T1; ikiden fazla grup (kanserin türleri) içeren veri kümelerinde artan sırada meanUE_T2, meanUE_T3, vb. şekilde adlandırmalar yapılmıştır. Tüm farklı gruplar için gruba özgü evrensel ortalama bilgi kütüphanesi böylelikle oluşturulmuştur.

c. sınıfa ait evrensel varyans bilgi kütüphanesi Eşitlik 4.42'de verilmiştir.

$$\sigma^{2}{}_{e_{i,j,c}}^{(k)} = \frac{1}{N_{s,c}-1}\sum_{j=1}^{N_F}\sum_{i=1}^{L_F}\sum_{k=1}^{N_{s,c}}(\mu_{e_{i,j,c}}^{(k)} - e_{i,j,c}^{(k)})^2 \tag{4.42}$$

(4.42) nolu eşitlikte verilen ifadenin matris formu Eşitlik 4.43'teki gibidir.

$$\sigma_{UE,c}^2 = \begin{bmatrix} \sigma_{e_{1,1}}^2 & \cdots & \sigma_{e_{1,N_F}}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{e_{L_F,1}}^2 & \cdots & \sigma_{e_{L_F,N_F}}^2 \end{bmatrix} \tag{4.43}$$

c. sınıfa ait evrensel ortanca bilgi kütüphanesi Eşitlik 4.44'te verilmiştir.

$$\tilde{\mu}_{e_{i,j,c}}^{(k)} = median\left(\sum_{j=1}^{N_F}\sum_{i=1}^{L_F}\sum_{k=1}^{N_{s,c}} e_{i,j,c}^{(k)}\right) \tag{4.44}$$

(4.44) nolu eşitlikte verilen ifadenin matris formu Eşitlik 4.45'teki gibidir.

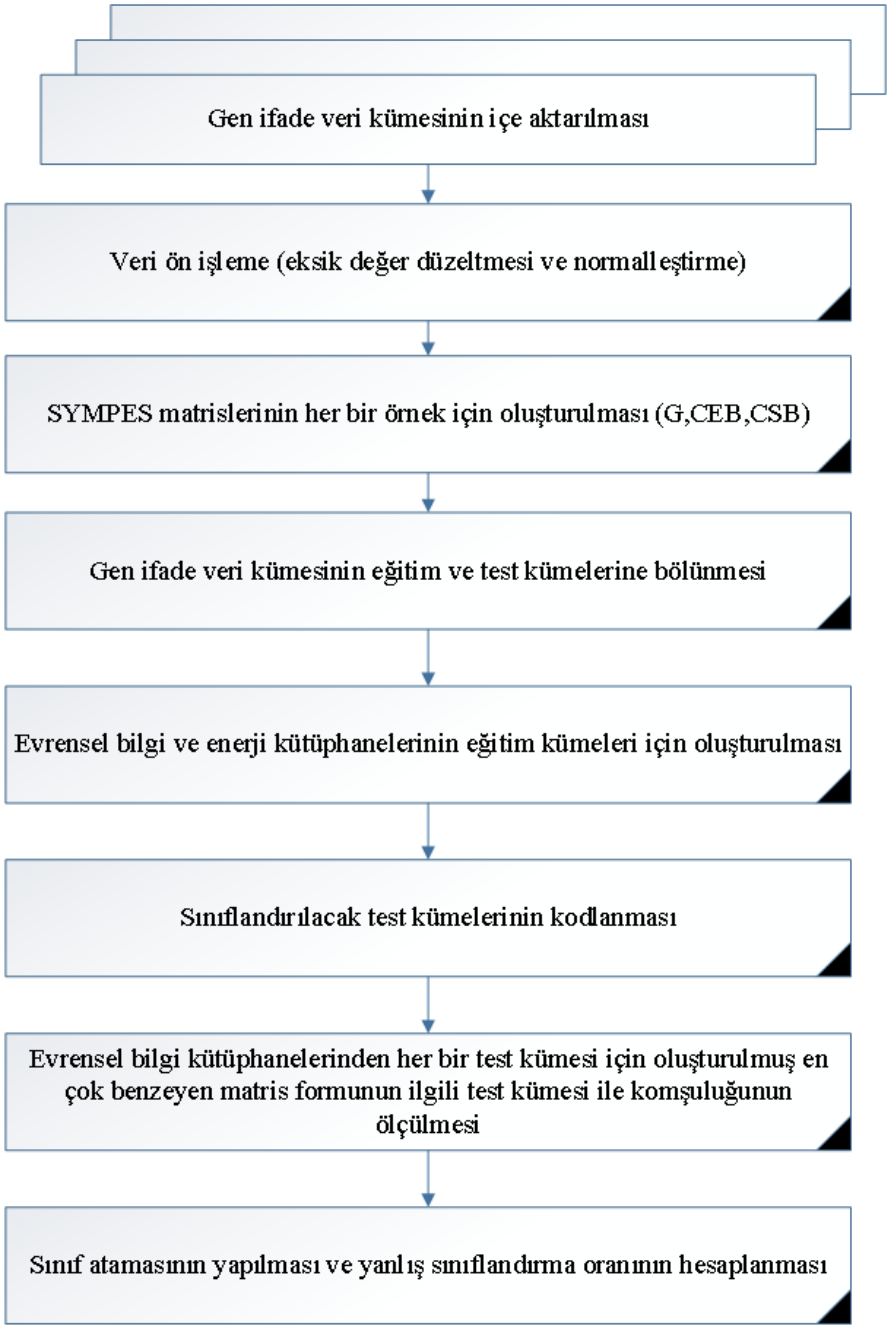
$$\tilde{\mu}_{UE,c} = \begin{bmatrix} \tilde{\mu}_{e_{1,1,c}} & \cdots & \tilde{\mu}_{e_{1,N_F,c}} \\ \vdots & \ddots & \vdots \\ \tilde{\mu}_{e_{L_F,1,c}} & \cdots & \tilde{\mu}_{e_{L_F,N_F,c}} \end{bmatrix} \tag{4.45}$$

$\tilde{\mu}_{UE,c}$ matrisine çalışmalarımızda iki grup içeren veri kümelerinde h_med ve t1_med; ikiden fazla grup içeren veri kümelerinde artan sırada t2_med, t3_med, vb. şekilde adlandırmalar yapılmıştır. Tüm farklı gruplar için gruba özgü evrensel ortanca bilgi kütüphanesi oluşturulmuştur.

4.3.SYMPES KÜTÜPHANELERİ KULLANILARAK YENİ BİR SINIFLANDIRMA YAKLAŞIMI

SYMPES kütüphanelerini ve komşuluk derecesinin ölçütlerinden en yaygın kullanılan Öklid mesafesini kullanarak gen ifade verilerinin sınıflandırılmasına yeni bir yaklaşım geliştirilmiştir. Şekil 4.7'de akış şeması olarak verilen diyagramda özetçe görülen işlem basamakları biraz daha açık bir dille maddeler halinde aşağıdaki gibi açıklanabilir:

a. Gen ifade verileri analizin yapılacağı paket yazılım vasıtasıyla içe aktarılır (yüklenir). Veri kümelerinin nasıl indirileceği Bölüm 3.1'de açıklanmıştı.
b. SYMPES ve diğer tekniklerin hesaplamalarının hatalı olmaması için öncelikle veri kümesinin eksik değer kontrolü yapılır. Ardından minimum – maksimum normalizasyonu yöntemi ile veri normalize edilir.
c. Veri kümesinin her bir örneği için SYMPES yöntemi ile önceden tanımlı bilgi ve enerji blokları oluşturulur (PEB ve PSB, sırasıyla *evrensel bilgi blokları* ve *evrensel enerji blokları*). Kazanç bileşeni (C) bu yaklaşımımızda kullanılmayacaktır.
d. Gen ifade veri kümesinden örnekleri rasgele seçilen, belirli yüzdelik dilimlerde örnek içerek şekilde eğitim ve test kümeleri oluşturulur. Bu eğitim ve test kümeleri %12.5 ve %87.5 örnek içerecek şekilde yapılandırılmıştır. Şöyle ki, 1. çalışmada %12.5 örnekli eğitim kümesi ve %87.5 örnekli test kümesi; 2. çalışmada %87.5 eğitim kümesi ve %12.5 test kümesi oluşturulmuştur (2 veya 3 örneklemi olan gruplarda, eğitim ve test kümelerinin oranları %40, %50, %70 şeklinde yapılandırılmıştır.).
e. Evrensel bilgi ve enerji blokları bir önceki maddede belirtilen eğitim kümeleri için oluşturulmuş ve test edilecek kümelerin kodlanması için kullanılmıştır.
f. Kodlama sürecinde, test kümelerinin her biri için özel, ona çok benzeyen eleman değerlerinin olduğu indis değerleri evrensel bilgi ve enerji kütüphaneleri kullanılarak kaydedilir.
g. Evrensel bilgi kütüphanelerinin bu indis değerlerine karşılık gelen değerlerinden oluşturulmuş, test edilecek küme ile aynı boyutlu olan her bir matrisin test kümesine olan uzaklığı Öklid mesafe ölçütü kullanılarak hesaplanır.
h. Her bir test kümesinin en yakın komşusuna sınıf ataması yapılır.
i. Yöntemin tüm test kümelerinin atandığı sınıfları ile gerçekte olan sınıfları kıyaslanarak (oranlanarak) yanlış sınıflandırma oranı (*İng.* misclassification ratio) hesaplanır.



Şekil 4.7: SYMPES ile sınıflandırma işlemi için akış şeması.

5. BULGULAR

5.1. RASGELE SEÇİLEN BİR DİZİNİN SYMPES YÖNTEMİ İLE DEĞERLENDİRİLMESİ

Bu çalışmada, Affymetrix dizileme cihazından elde edilen verilerin tamamı SYMPES yöntemi ile değerlendirmeye tabi tutulur. Bu değerlendirme yöntemi aşağıdaki algoritma ile yapılır:

a) Herhangi bir kişiden elde edilen gen dizisi, kullanılan gen çipi platformunun (GPL570) sağladığı toplamda 54.675 gen içermektedir. Bu gen dizisi öncelikle $L_F = \{5,9,15,25\}$ uzunluğunda çerçevelere ayrılır.
b) Seçilen çerçeve uzunluğundaki dizinin tamamı için, her bir çerçevenin SYMPES modeli kurulur ve buna ilişkin bilgi dizilerinin tamamı bir matriste saklanır.
c) Modellenen dizinin daha önce oluşturulan evrensel bilgi kütüphaneleri ile çerçeve bazında teker teker karşılaştırması yapılır.
d) Bilgi dizilerinin karşılaştırılması değişik yöntemlerle olabilir. Bu yöntemlerden biri olan SWRS, bilgi dizilerinin çerçeveleri bazında, Wilcoxon sıra sayıları-toplam testini $\alpha = 0.5$ anlamlılık düzeyinde kullanarak, kıyaslaması yapılan grupların anlamlı çerçevelerini ayrıştırmaktadır. Bu yönteme ek olarak, ikinci bir yöntem olan $SWRS_{cor}$, SWRS yönteminin bulabileceği sahte pozitif gen çerçevelerini elemek için, FWER düzeltme yöntemlerinden birini kullanarak, yeni anlamlı çerçevelerinin bulunmasını sağlar. Üçüncü bir yöntem olan SWT, parametrik olmayan t-testini (Welch t-test) $\alpha = 0.5$ anlamlılık düzeyinde kullanarak, kıyaslaması yapılan grupların anlamlı çerçevelerini ayrıştırmaktadır. Bu yönteme ek olarak, dördüncü yöntemde (SWT_{cor}), SWT yönteminin bulabileceği sahte pozitif gen çerçevelerini elemek için FWER düzeltmesi kullanılarak yeni anlamlı gen çerçeveleri elde edilir. Beşinci yöntemde ($SF_{min\text{-}max}$), grupların bilgi dizilerinin çerçevelerinde aynı indekste kayıtlı olan gen ifadelerinin bilgi değerlerinin oranı alınarak bağıl değişim gözlemlenir. Bu yöntemde maksimum ve minimum oranların elde edildiği indeks değerleri kaydedilir. Altıncı ve son yöntemde ($SPCCF_{min\text{-}max}$), Pearson özilinti katsayısı ($|\rho| \leq 0.1$) ölçütünü kullanarak grupların çerçevelerinde özilinti değerleri bu ölçüte uyan gen çerçevelerini dikkate alır ve bu çerçevelerin beşinci yöntemde olduğu gibi kat değişimi oranlarını

hesaplayarak her çerçevenin minimum ve maksimum oranlarına karşılık gelen indeks değerlerini kaydeder.

e) Eğer cihazdan gelen zarf dizileri ile hastalara ilişkin zarf dizileri arasında hiçbir uyuşma tespit edilmezse, cihazdan gelen zarf dizilerinin sağlıklı olduğu varsayılır. Tam tersine, eğer bu dizilerle hastalıklı dizilerin örtüşmesi söz konusu ise cihazdan gelen dizilerde hastalık belirtisinin olduğu kabul edilir.

f) Cihazdan gelen ve hastalık olduğu görülen dizilerde hangi genlerin etkin rol oynadığının (farksal olarak ifade edilen) tespiti yapılır. Farksal olarak ifade edilen genlerle hastalıklı genlerin karşılaştırması yapılarak önemli genler ortaya çıkarılmış olur.

Şunu belirtmeliyiz ki, farksal olarak ifade edilen genlerin ortaya çıkarılmasında, SYMPES yöntemi ile her bir grup için üretilen evrensel bilgi kütüphanelerinin Pearson özilinti katsayısı kullanılarak benzerlik ölçümünün yapılması sonucunda en yüksek başarımın elde edildiği görülmüştür. Benzerlik ölçütü olarak $\nu \geq 0.9$ kabul edilmiştir. Bir başka ifade ile, benzersizlik ölçütü olarak $\eta = 1 - \nu$ ifadesinden elde edileceği üzere, $\eta \leq 0.1$ eşitsizliği kabul edilmiştir.

5.2.ÖNEMLİ GEN ÇERÇEVELERİNİN VE FARKSAL OLARAK İFADE EDİLEN GENLERİN BULUNMASI

SYMPES yöntemi ile önemli gen çerçevelerinin bulunması sürecinde parametrik (Welch t-testi) ve parametrik olmayan (Wilcoxon sıra sayıları toplam testi) istatistiksel testler kullanılmıştır. Anlamlılık düzeyi $\alpha = 0.5, 0.1, 0.05, 0.01$ değerleri için yapılan denemelerden sonra uygun değer $\alpha = 0.5$ olarak belirlenmiştir. Çerçeve boyutları 5, 9, 15 ve 25 olarak yapılandırılarak evrensel bilgi kütüphaneleri oluşturulmuştur. FWER düzeltmesi yapmadan önceki ve yaptıktan sonraki durumların her ikisi için de bulunan önemli gen çerçeveleri belirtilmiştir. FWER düzeltmesi için Holland-Copenhaver yöntemi tercih edilmiştir. Düzeltme yapılırken, yeni anlamlılık düzeyi $\alpha = 0.5$ olarak belirlenmiştir. Bu sayede gerçek pozitif sonuçların elde edilmesi sağlanmıştır. Veri kümelerinin grupları ikili olarak ayrılarak istatistiksel testler uygulanmıştır. Gruplar normal sağlıklı ile kanser türleri arasında oluşturulmuştur. Buradan şu anlam çıkmaktadır: 2 gruplu bir veri kümesinde 1 tane ikili grup oluşurken; 3 gruplu bir veri kümesinde 2; 4 gruplu bir veri kümesinde 3 tane ikili grup oluşur. İkili gruplara ayrılarak istatistiksel testlerin uygulanmasının ardından, her bir çerçeve için maksimum bilgiyi taşıyan indise karşılık gelen genlerden (veya prob adlarından) oluşan bir gen (veya prob) listesi oluşturulmuştur. Oluşturulan bu gen (veya prob) listesinin, GEO2R

platformu ile $p < 0.05$ düzeyinde anlamlı bulunan gen veya prob listesi ile ne kadar örtüştüğü incelenmiş, elde edilen sonuçlar tablolaştırılarak sunulmuştur. Yeni yaklaşımla üretilen evrensel ortalama ve ortanca bilgi kütüphanelerinin farklı çerçeve boyutları için önemli bulunan GF sayısı ile gen sayıları ilerleyen bölümlerde ilgili veri kümeleri için sunulmuştur.

5.2.1. Meme Kanseri Veri Kümesinin Sonuçları

SYMPES yöntemi ile dört sınıflı meme kanserinin her bir sınıfı için oluşturulan evrensel ortalama ve ortanca bilgi kütüphanelerinde, çerçeve boyutlarına göre değişen GF sayısınca genlerden oluşan bir gen listesi gen çipinin sağladığı büyük gen listesi kullanılarak oluşturulmuştur.

Tablo 5.1: SYMPES yönteminin değişen çerçeve boyutlarına göre meme kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak).

Gruplar	Çerçeve Boyutu (L_F)	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$	$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$
H[3]	5	5180	5203	5180 (47.37)	5203 (47.58)
	9	2979	2982	2979 (49.04)	2982 (49.09)
	15	1830	1846	1830 (50.21)	1846 (50.64)
	25	1120	1123	**1120 (51.21)**	**1123 (51.35)**
T1[4]	5	5062	5063	5062 (46.29)	5063 (46.30)
	9	2874	2889	2874 (47.31)	2889 (47.56)
	15	1775	1758	1775 (48.70)	1758 (48.23)
	25	1072	1067	**1072 (49.02)**	**1067 (48.79)**
T2[5]	5	5066	5066	5066 (46.33)	5066 (46.33)
	9	2901	2901	2901 (47.75)	2901 (47.75)
	15	1768	1768	1768 (48.50)	1768 (48.50)
	25	1071	1071	**1071 (48.97)**	**1071 (48.97)**
T3[6]	5	5090	5083	5090 (46.55)	5083 (46.48)
	9	2922	2941	2922 (48.10)	2941 (48.41)
	15	1778	1774	1778 (48.78)	1774 (48.67)
	25	1079	1076	**1079 (49.34)**	**1076 (49.20)**

[3] H: Normal sağlıklı birey
[4] T1: 1. tür kanseri olan birey
[5] T2: 2. tür kanseri olan birey
[6] T3: 3. tür kanseri olan birey

Her bir çerçevenin en büyük bilgi taşıyan indislerine karşılık gelen genlerinden oluşturulan bu yeni gen listesinin referans gen listesi ile ortak genlerinden yeni bir küme oluşturulmuştur. Bu kesişim kümesindeki genlerin sayısı ve yüzdelik karşılığı Tablo 5.1'de verilmiştir. Çerçeve boyutunun artması ile iki gen listesinin kesişim kümesinin içerdiği önemli gen sayısı da artmıştır.

Tablo 5.2: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 5\,, \alpha = 0.5,\ |\rho| \leq 0.1$).

Gruplar	**Yöntemler**	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	94	102	127 (27.02)	147 (28.82)
	$SWRS_cor$	2	2	2 (20)	3 (30)
	SWT	--	3	0 (0)	5 (33.33)
	SWT_cor	--	1	0 (0)	0 (0)
	$SF_{min-max}$	6518	6417	6518 (59.61)	6417 (58.68)
	$SPCCF_{min-max}$	514	622	**514 (62.99)**	**622 (65.47)**
H ~ T2	$SWRS$	121	149	186 (30.74)	235 (31.54)
	$SWRS_cor$	4	4	7 (35)	6 (30)
	SWT	--	2	0 (0)	2 (20)
	SWT_cor	--	1	0 (0)	1 (20)
	$SF_{min-max}$	7641	7591	**7641 (69.88)**	**7591 (69.42)**
	$SPCCF_{min-max}$	565	586	565 (69.24)	586 (61.68)
H ~ T3	$SWRS$	169	185	318 (37.63)	354 (38.27)
	$SWRS_cor$	3	4	5 (33.33)	5 (25)
	SWT	--	4	0 (0)	9 (45)
	SWT_cor	--	1	0 (0)	2 (40)
	$SF_{min-max}$	8473	8437	**8473 (77.49)**	**8437 (77.16)**
	$SPCCF_{min-max}$	500	628	500 (61.27)	628 (66.11)

Tablo 5.2'de geliştirilen 6 yeni yöntemin her biri için yeni birer gen listesi oluşturulmuş, referans gen listesi ile kesişim kümelerindeki gen sayısı ve yüzdelik karşılığı verilmiştir. $SPCCF_{min-max}$ yöntemi H ~ T1 grubunda evrensel ortalama ve ortanca bilgi kütüphaneleri için sırasıyla %62.99 ve %65.47 ile en iyi sonucu verirken, diğer iki grup (H ~ T2 ve H ~ T3) için $SF_{min-max}$ yöntemi en iyi sonuçları vermiştir. Her iki evrensel kütüphane için sonuçlar

birbirine yakın çıksa da ortalama değerlerden oluşan kütüphanenin daha çok bilgi sağlayıcı özelliğe sahip olduğu görülmüştür.

Çerçeve boyutunun 5'ten 9'a arttırılmasıyla kesişen kümedeki gen sayılarının yüzdece daha çok arttığı, bu doğrultuda bilgi sağlama açısından daha etkin sonuçlar verdiği gözlemlenmiştir (Tablo 5.3). $SF_{min-max}$ yönteminin genel olarak tüm gruplarda en iyi sonucu veren yöntem olduğu karşımıza çıkmıştır.

Tablo 5.3: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 9\,, \alpha = 0.5, |\rho| \leq 0.1$).

Gruplar	**Yöntemler**	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	70	80	173 (27.46)	197 (27.36)
	$SWRS_cor$	3	2	8 (29.63)	6 (33.33)
	SWT	--	2	0 (0)	6 (33.33)
	SWT_cor	--	1	0 (0)	2 (22.22)
	$SF_{min-max}$	4336	4196	**4336 (71.37)**	4196 (69.07)
	$SPCCF_{min-max}$	328	365	328 (71.30)	**365 (73)**
H ~ T2	$SWRS$	98	120	269 (30.50)	338 (31.30)
	$SWRS_cor$	3	3	11 (40.74)	10 (37.04)
	SWT	--	1	0 (0)	1 (5.56)
	SWT_cor	--	1	0 (0)	1 (5.56)
	$SF_{min-max}$	4857	4835	**4857 (79.95)**	**4835 (79.59)**
	$SPCCF_{min-max}$	338	379	338 (73.48)	379 (75.80)
H ~ T3	$SWRS$	133	131	398 (33.25)	394 (33.42)
	$SWRS_cor$	3	3	11 (40.74)	11 (40.74)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	5331	5261	**5331 (87.75)**	**5261 (86.60)**
	$SPCCF_{min-max}$	345	355	345 (75)	355 (71)

Çerçeve boyutunun 9'dan 15'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%96'lara varan), yine genel olarak $SF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.4'te sunulmuştur.

Tablo 5.4: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 15\,, \alpha = 0.5,\ |\rho| \leq 0.1$).

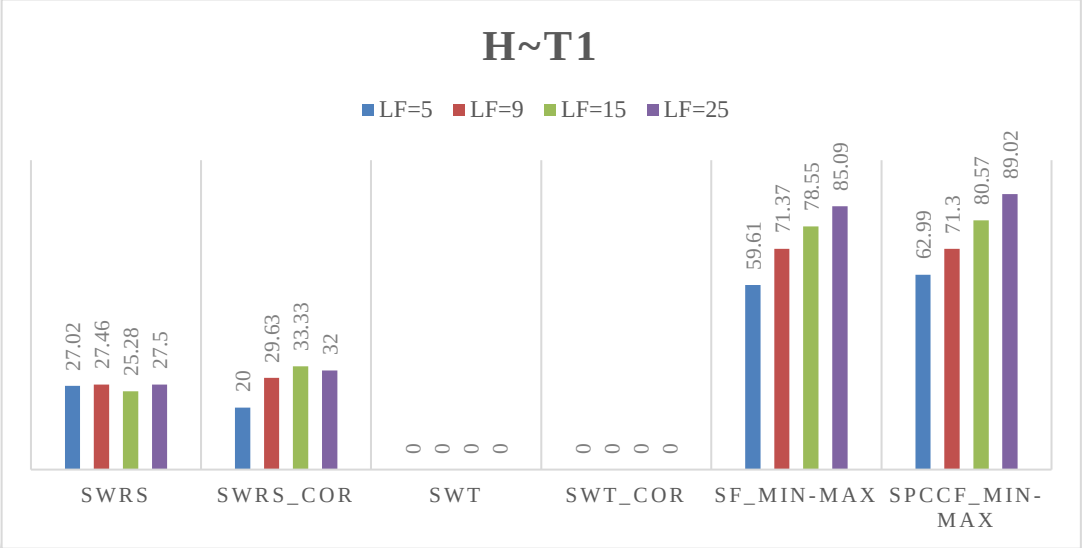
Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	24	25	91 (25.28)	97 (25.87)
	$SWRS_cor$	2	2	10 (33.33)	9 (30)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	2863	2958	2863 (78.55)	2958 (81.15)
	$SPCCF_{min-max}$	199	267	**199 (80.57)**	**267 (83.18)**
H ~ T2	$SWRS$	38	58	160 (28.07)	236 (27.13)
	$SWRS_cor$	1	1	1 (6.67)	3 (20)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	3128	3077	**3128 (85.82)**	**3077 (84.42)**
	$SPCCF_{min-max}$	190	262	190 (76.92)	262 (81.62)
H ~ T3	$SWRS$	63	50	282 (29.84)	235 (31.33)
	$SWRS_cor$	2	3	5 (16.67)	9 (30)
	SWT	--	1	0 (0)	7 (46.67)
	SWT_cor	--	1	0 (0)	7 (46.67)
	$SF_{min-max}$	3483	3456	**3483 (95.56)**	**3456 (94.81)**
	$SPCCF_{min-max}$	203	262	203 (82.19)	262 (81.62)

Çerçeve boyutunun 15'ten 25'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%98'lere varan), yine genel olarak $SF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.5'te sunulmuştur.

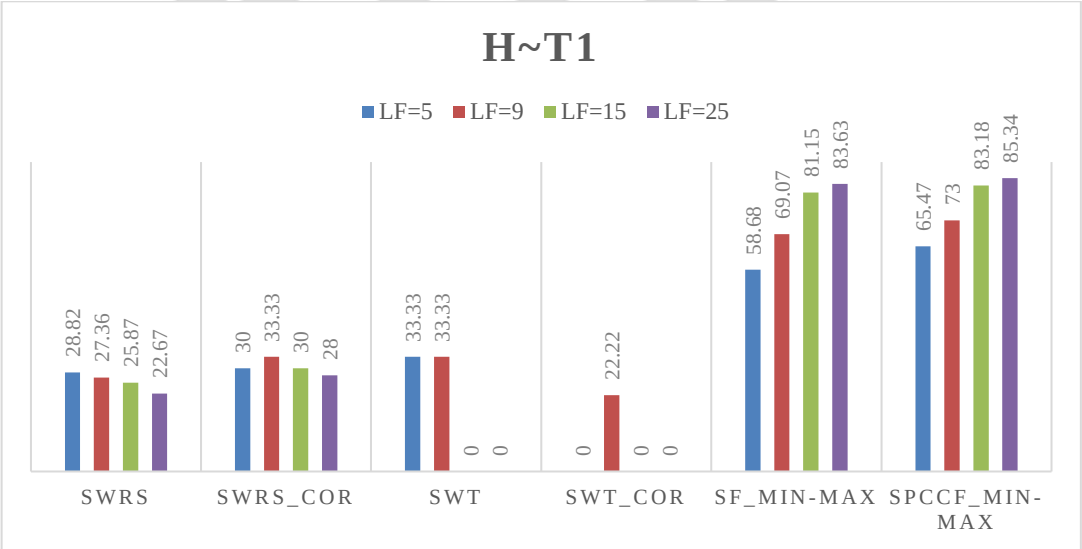
Tablo 5.5: Meme kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 25\,, \alpha = 0.5, |\rho| \leq 0.1$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	8	9	55 (27.50)	51 (22.67)
	$SWRS_cor$	1	1	8 (32)	7 (28)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	1861	1829	1861 (85.09)	1829 (83.63)
	$SPCCF_{min-max}$	154	163	**154 (89.02)**	**163 (85.34)**
H ~ T2	$SWRS$	25	50	189 (30.24)	353 (28.24)
	$SWRS_cor$	2	2	11 (44)	12 (48)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	1964	1936	**1964 (89.80)**	**1936 (88.52)**
	$SPCCF_{min-max}$	146	151	146 (84.39)	151 (79.06)
H ~ T3	$SWRS$	48	27	362 (30.17)	192 (28.44)
	$SWRS_cor$	2	2	12 (48)	11 (44)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	2136	2118	**2136 (97.67)**	**2118 (96.84)**
	$SPCCF_{min-max}$	148	155	148 (85.55)	155 (81.15)

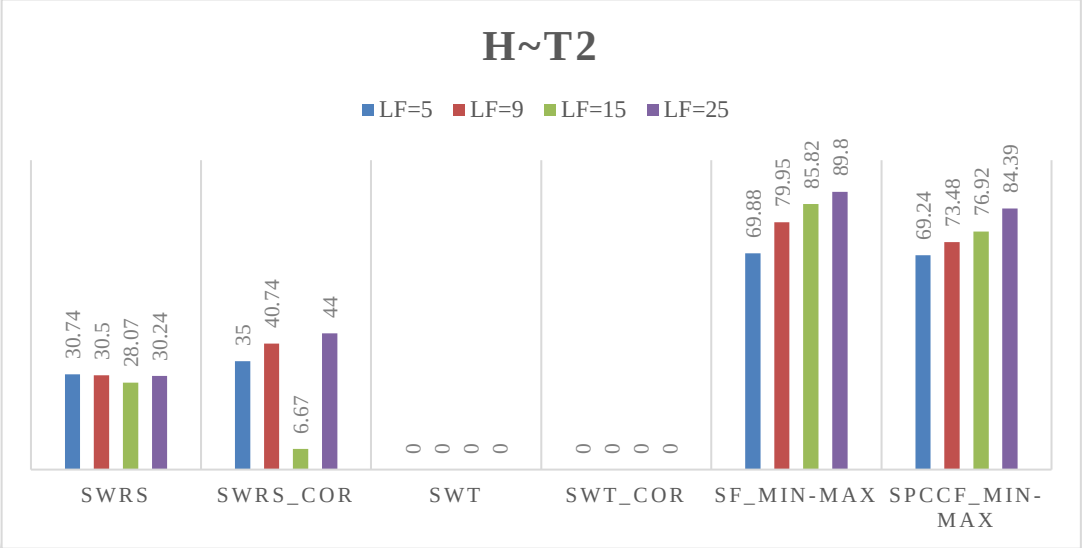
Tablo 5.2 – 5.5'te detaylı olarak verilen bulguların yüzdelik değerleri, karşılaştırılan gruplar bazında (H ~ T1, H ~ T2 ve H ~ T3) evrensel ortalama ve ortanca bilgi kütüphaneleri için ayrı ayrı çubuk grafiklerle Şekil 5.1 – Şekil 5.6'da verilmiştir.



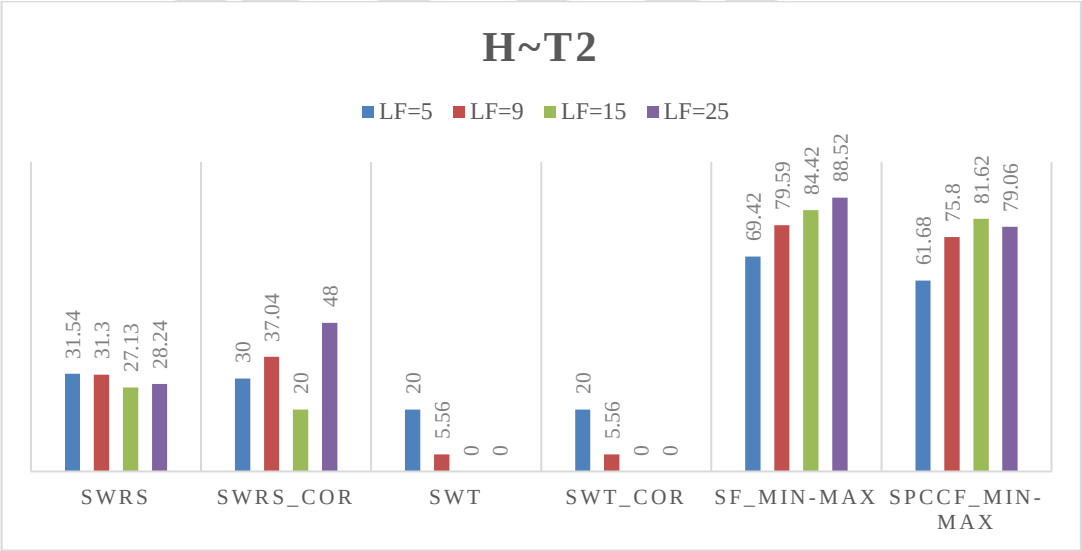
Şekil 5.1: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)



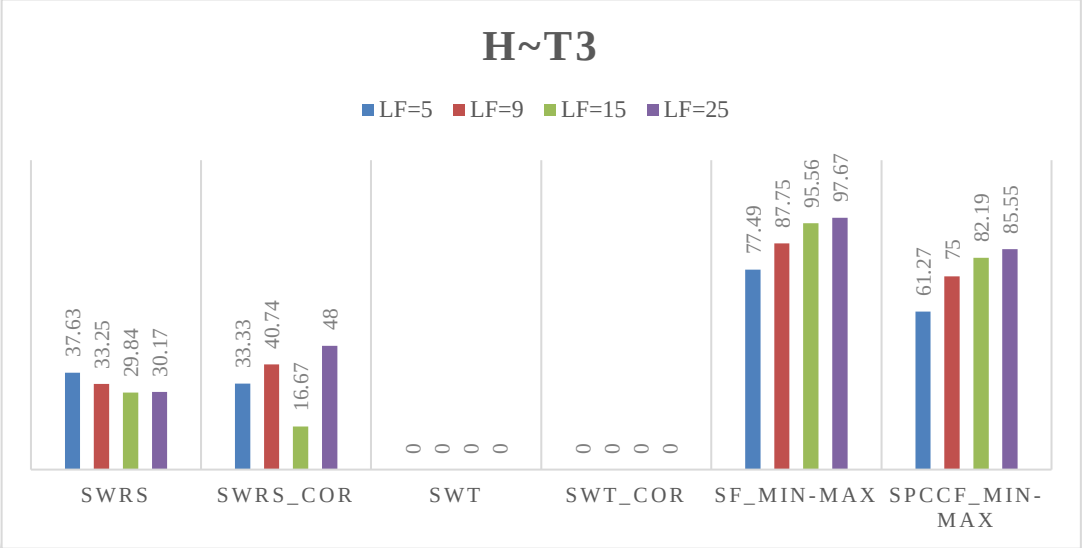
Şekil 5.2: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)



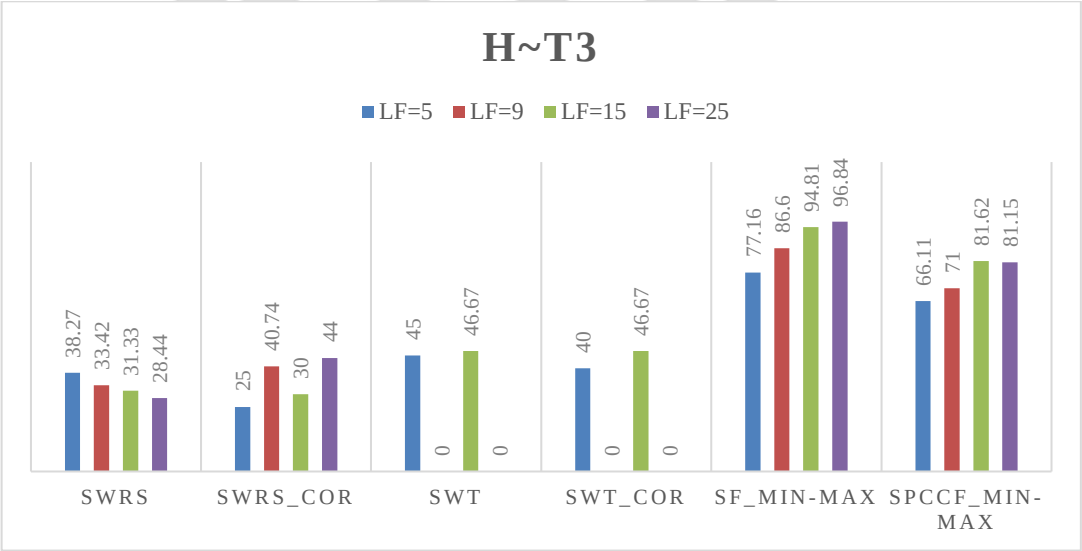
Şekil 5.3: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için)



Şekil 5.4: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için)



Şekil 5.5: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için)

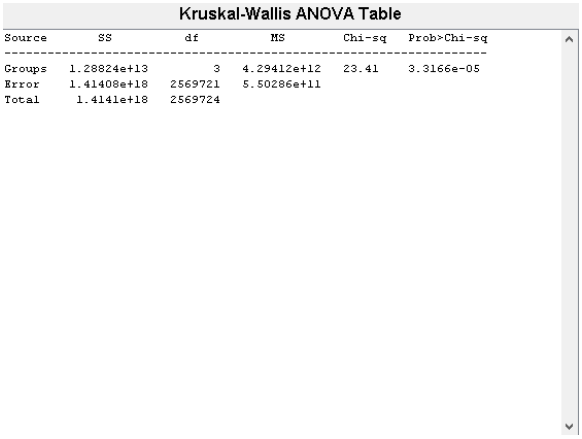


Şekil 5.6: Meme kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için)

Çerçeve boyutlarının değişimi ile meme kanseri veri kümesinin grupları arasında anlamlı farklılık olup olmadığı Kruskal – Wallis H testi ile gözlemlenmiştir. Sonuçlar ANOVA tablosu, kutu – bıyık grafiği (veri dağılımı hakkında genel bilgi veren grafiksel bir gösterim formu) ve seçilen her bir grubun diğerleri ile arasındaki anlamlı farklılığı gösteren grafikler şeklinde verilmiştir. Çerçeve boyutunun 5 olarak yapılandırılması ile bu testin uygulanması sonucunda elde edilen tüm sonuçlar Şekil 5.7'de verilmiştir. Aynı şekilde, çerçeve boyutunun 9 olarak

yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.8'de; çerçeve boyutunun 15 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.9'da ve son olarak çerçeve boyutunun 25 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.10'da verilmiştir.

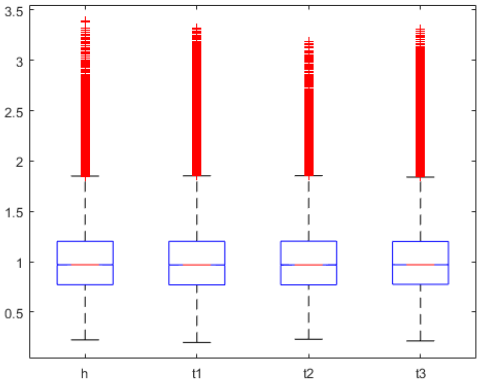
Özetle, tüm çerçeve boyutlarında gruplar arasında önemli seviyede anlamlı farklılık olduğu ($p \ll 0.05$), aykırı değerlerin dağılımının kutu – bıyık grafikleri ile gözlemlenebildiği, T2 grubunun diğer gruplar ile arasında anlamlı farklılığı bulunmadığı (bu durumun örneklem sayısının da diğerlerinden çok az olmasından kaynaklandığı öngörülmektedir), çerçeve boyutlarının artması ile anlamlı farklılığın daha çok arttığı (grupların birbirlerinden farklılaşması) gözlemlenmiştir.



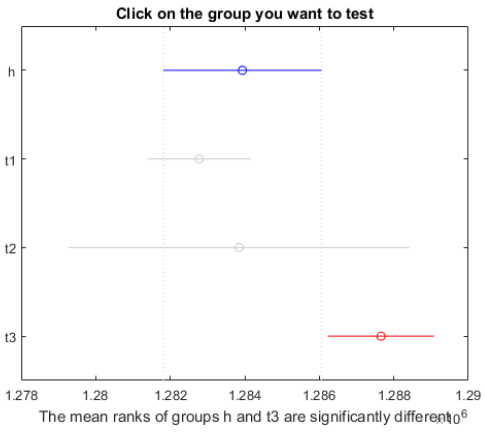
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.28824e+13	3	4.29412e+12	23.41	3.3166e-05
Error	1.41408e+18	2569721	5.50286e+11		
Total	1.4141e+18	2569724			

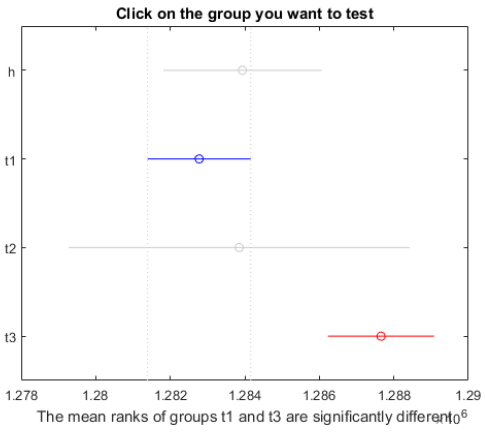
(a)



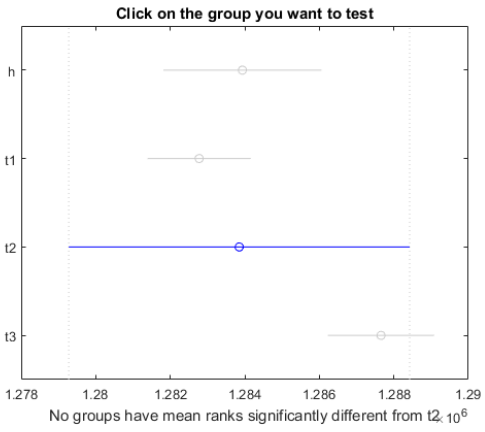
(b)



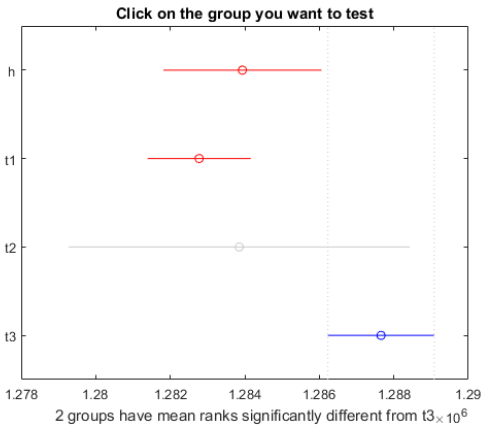
(c)



(d)

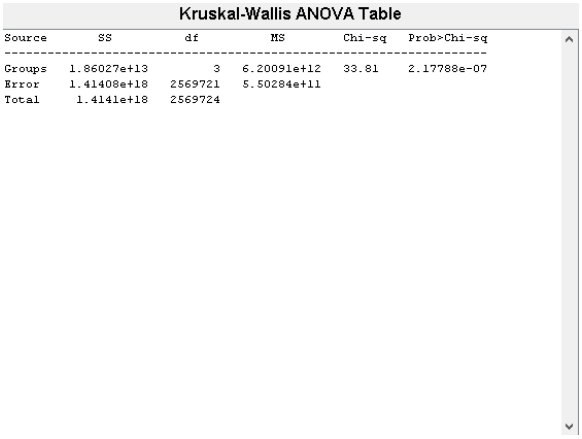


(e)

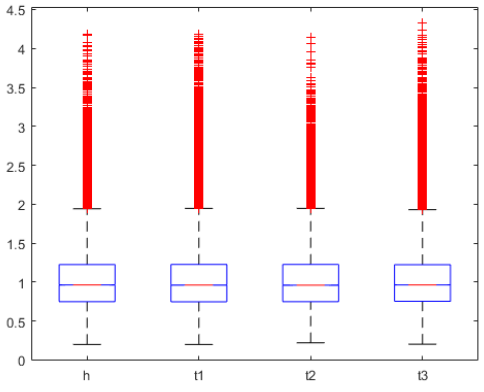


(f)

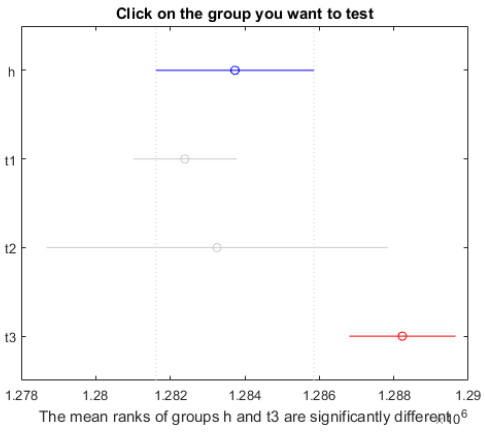
Şekil 5.7: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 5, \alpha = 0.05$).



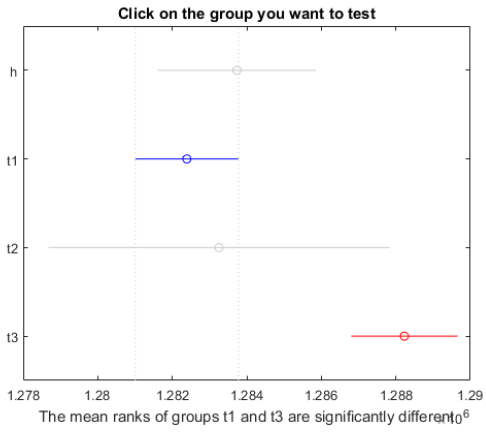
(a)



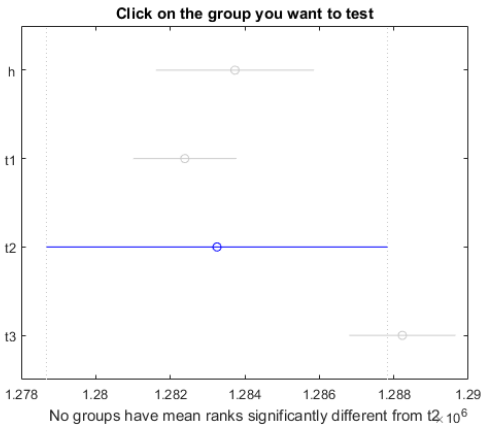
(b)



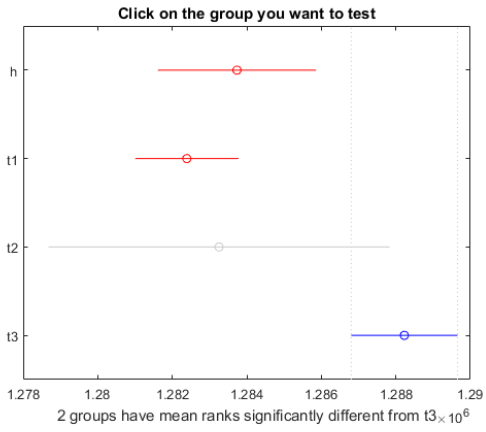
(c)



(d)

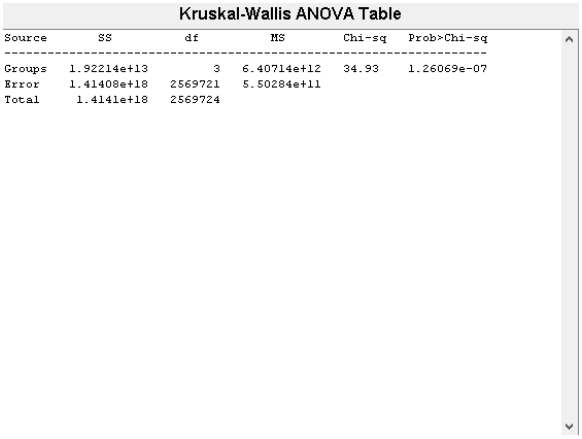


(e)



(f)

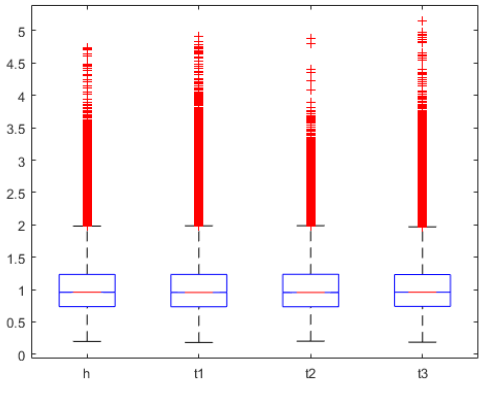
Şekil 5.8: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 9, \alpha = 0.05$).



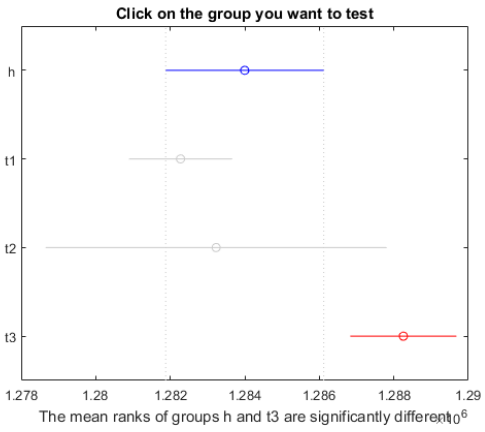
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.92214e+13	3	6.40714e+12	34.93	1.26069e-07
Error	1.41408e+18	2569721	5.50284e+11		
Total	1.4141e+18	2569724			

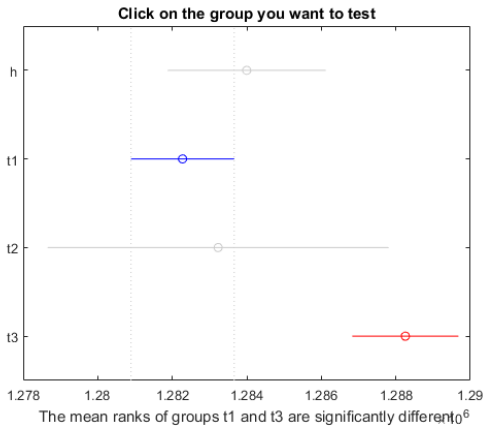
(a)



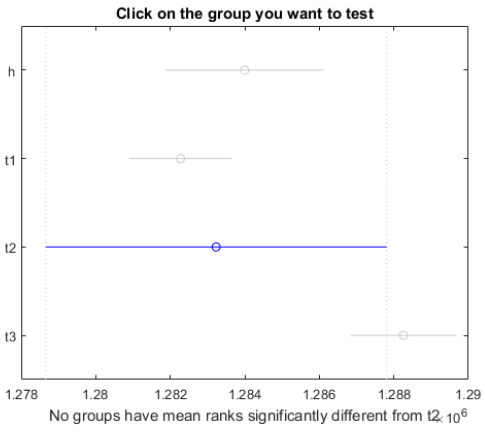
(b)



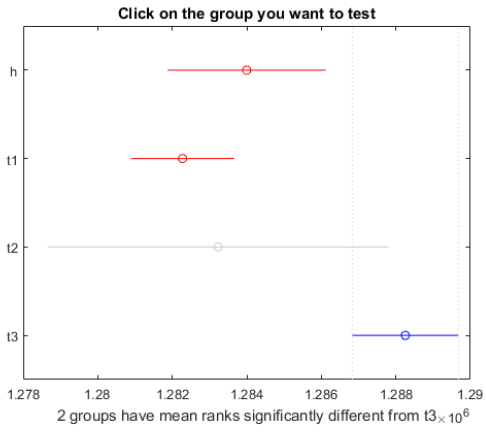
(c)



(d)

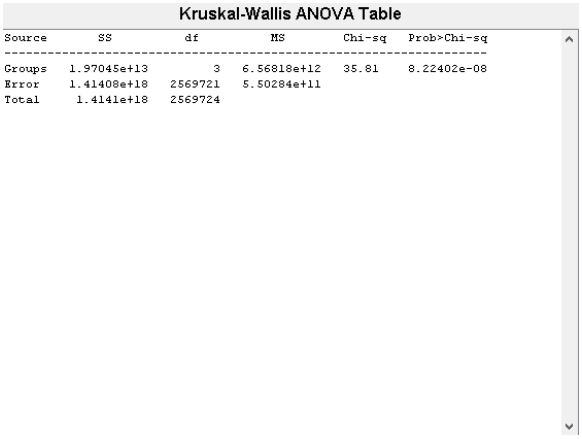


(e)

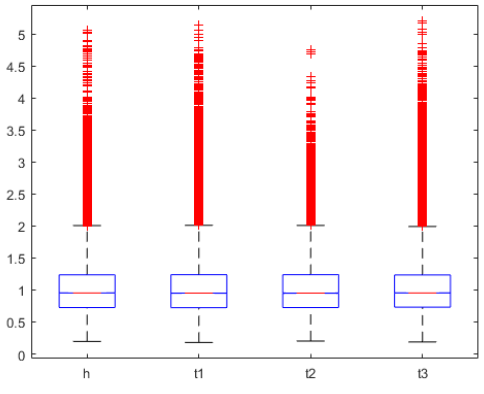


(f)

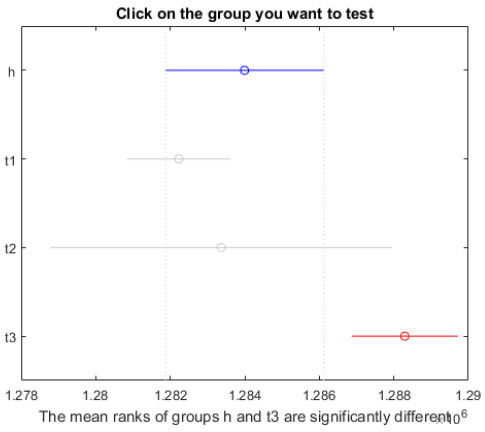
Şekil 5.9: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 15, \alpha = 0.05$).



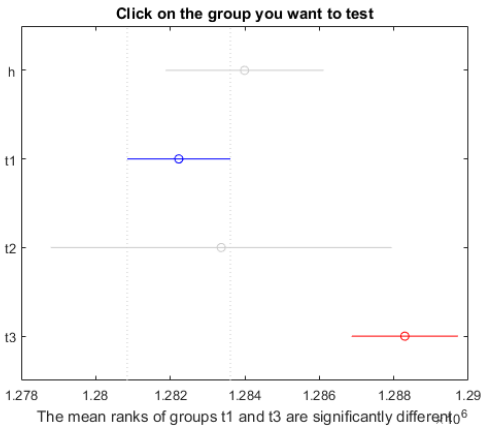
(a)



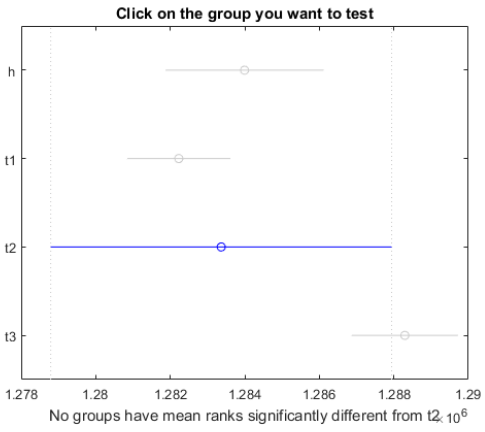
(b)



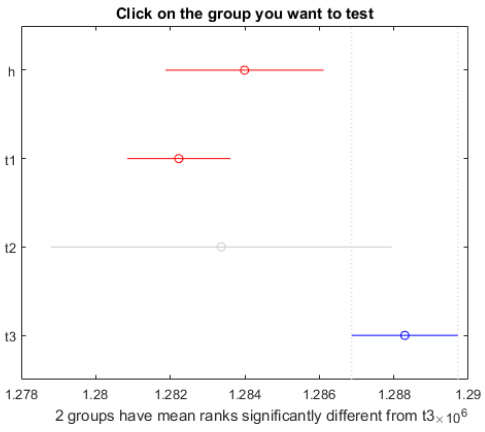
(c)



(d)



(e)



(f)

Şekil 5.10: Meme kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 25, \alpha = 0.05$).

5.2.2. Akciğer Kanseri Veri Kümesinin Sonuçları

SYMPES yöntemi ile iki sınıflı akciğer kanserinin her bir sınıfı için oluşturulan evrensel ortalama ve ortanca bilgi kütüphanelerinde, çerçeve boyutlarına göre değişen GF sayısınca genlerden oluşan bir gen listesi gen çipinin sağladığı büyük gen listesi kullanılarak oluşturulmuştur. İki sınıflı olan akciğer kanseri veri kümesinin çok sınıflı veri kümelerinden farklı olarak, Pearson korelasyon katsayısının ± 1 değerlerini içermesinden dolayı zıt ilişki durumu için geçerli olan $\rho < 0$ değerlerine karşılık gelen GF'ler anlamlı farklılık gösteren GF'ler olarak varsayılmıştır. $SPCCF_{min-max}$ yönteminde $\rho < 0$ değeri dikkate alınmıştır.

Tablo 5.6: SYMPES yönteminin değişen çerçeve boyutlarına göre akciğer kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak).

Gruplar	**Çerçeve Boyutu** (L_F)	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$	$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$
H	5	4251	4251	4251 (38.88)	4251 (38.88)
	9	2378	2423	2378 (39.14)	2423 (39.88)
	15	1396	1432	**1396 (40.91)**	1432 (39.29)
	25	845	880	845 (38.64)	**880 (40.24)**
T1	5	4239	4240	4239 (38.77)	4240 (38.77)
	9	2446	2443	2446 (40.26)	2443 (40.21)
	15	1491	1478	**1491 (40.91)**	**1478 (40.55)**
	25	868	880	868 (39.69)	880 (40.24)

Her bir çerçevenin en büyük bilgi taşıyan indislerine karşılık gelen genlerinden oluşturulan bu yeni gen listesinin referans gen listesi ile kesişen genlerinden yeni bir küme oluşturulmuştur.

Bu kesişim kümesindeki genlerin sayısı ve yüzdelik karşılığı Tablo 5.6'da verilmiştir. Çerçeve boyutunun artması ile iki gen listesinin kesişim kümesinin içerdiği önemli gen sayısı da artmıştır.

Tablo 5.7'de geliştirilen 6 yeni yöntemin her biri için yeni birer gen listesi oluşturulmuş, referans gen listesi ile kesişim kümelerindeki gen sayısı ve yüzdelik karşılığı verilmiştir. $SPCCF_{min-max}$ yöntemi H ~ T1 grubunda evrensel ortalama ve ortanca bilgi kütüphaneleri için sırasıyla %61.83 ve %62.08 ile en iyi sonucu vermiştir.

Tablo 5.7: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 5\,, \alpha = 0.5,\ \rho < 0$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	971	900	1676 (34.52)	1536 (34.13)
	$SWRS_cor$	4	6	6 (30)	9 (30)
	SWT	19	66	35 (36.84)	96 (29.09)
	SWT_cor	2	3	3 (30)	2 (13.33)
	$SF_{min-max}$	6348	6421	6348 (58.05)	6421 (58.72)
	$SPCCF_{min-max}$	3641	3783	**3641 (61.83)**	**3783 (62.08)**

Çerçeve boyutunun 5'ten 9'a arttırılmasıyla kesişen kümedeki gen sayılarının yüzdece daha çok arttığı, bu doğrultuda bilgi sağlama açısından daha etkin sonuçlar verdiği gözlemlenmiştir (Tablo 5.8). $SPCCF_{min-max}$ yönteminin genel olarak tüm gruplarda en iyi sonucu veren yöntem olduğu karşımıza çıkmıştır.

Tablo 5.8: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 9\,, \alpha = 0.5,\ \rho < 0$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1219	1045	3841 (35.01)	3200 (34.02)
	$SWRS_cor$	16	14	41 (28.47)	27 (21.43)
	SWT	7	82	20 (31.75)	266 (36.04)
	SWT_cor	1	3	5 (55.56)	1 (3.70)
	$SF_{min-max}$	3663	3769	3663 (60.30)	**3769 (62.04)**
	$SPCCF_{min-max}$	1753	1620	**1753 (62.12)**	1620 (61.60)

Çerçeve boyutunun 9'dan 15'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%66'lara varan), yine genel olarak $SPCCF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.9'da sunulmuştur.

Tablo 5.9: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 15\,, \alpha = 0.5,\ \rho < 0$).

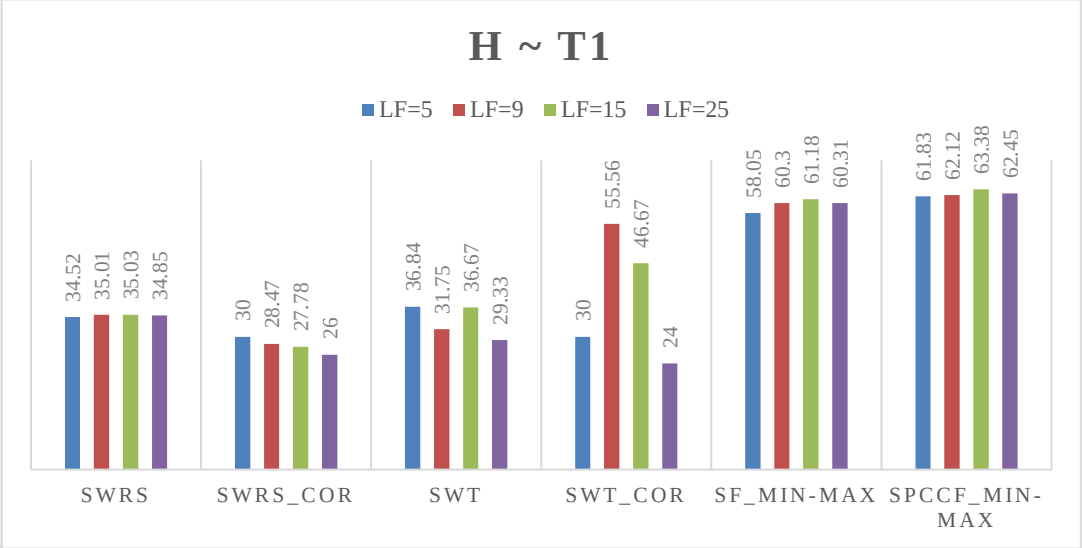
Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1107	819	5817 (35.03)	4208 (34.25)
	$SWRS_cor$	24	16	100 (27.78)	68 (28.33)
	SWT	2	87	11 (36.67)	480 (36.78)
	SWT_cor	1	3	7 (46.67)	17 (37.78)
	$SF_{min-max}$	2230	2316	2230 (61.18)	2316 (63.54)
	$SPCCF_{min-max}$	1265	1402	**1265 (63.38)**	**1402 (65.64)**

Çerçeve boyutunun 15'ten 25'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak azalsa da 15 boyutlu çerçeve yapısındaki ile yakın sonuçlar verdiği, yine genel olarak $SPCCF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.10'da sunulmuştur.

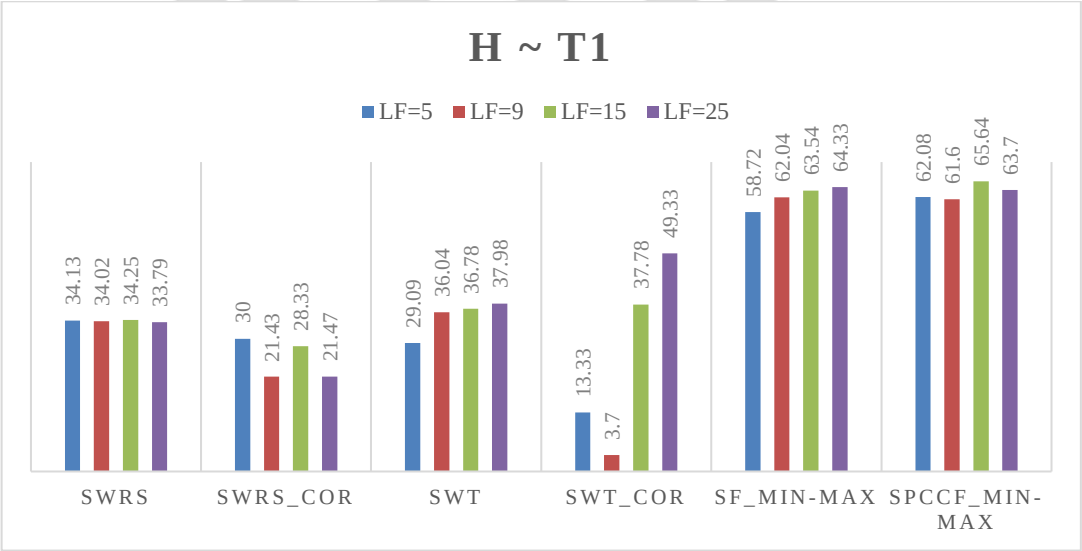
Tablo 5.10: Akciğer kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 25\,, \alpha = 0.5,\ \rho < 0$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	951	651	8286 (34.85)	5500 (33.79)
	$SWRS_cor$	30	19	195 (26)	102 (21.47)
	SWT	3	87	22 (29.33)	826 (37.98)
	SWT_cor	1	3	6 (24)	37 (49.33)
	$SF_{min-max}$	1319	1407	1319 (60.31)	**1407 (64.33)**
	$SPCCF_{min-max}$	710	530	**710 (62.45)**	530 (63.70)

SYMPES yöntemiyle oluşturulan evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.11'de verilmiştir. Benzer şekilde, SYMPES yöntemiyle oluşturulan evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.12'de verilmiştir.



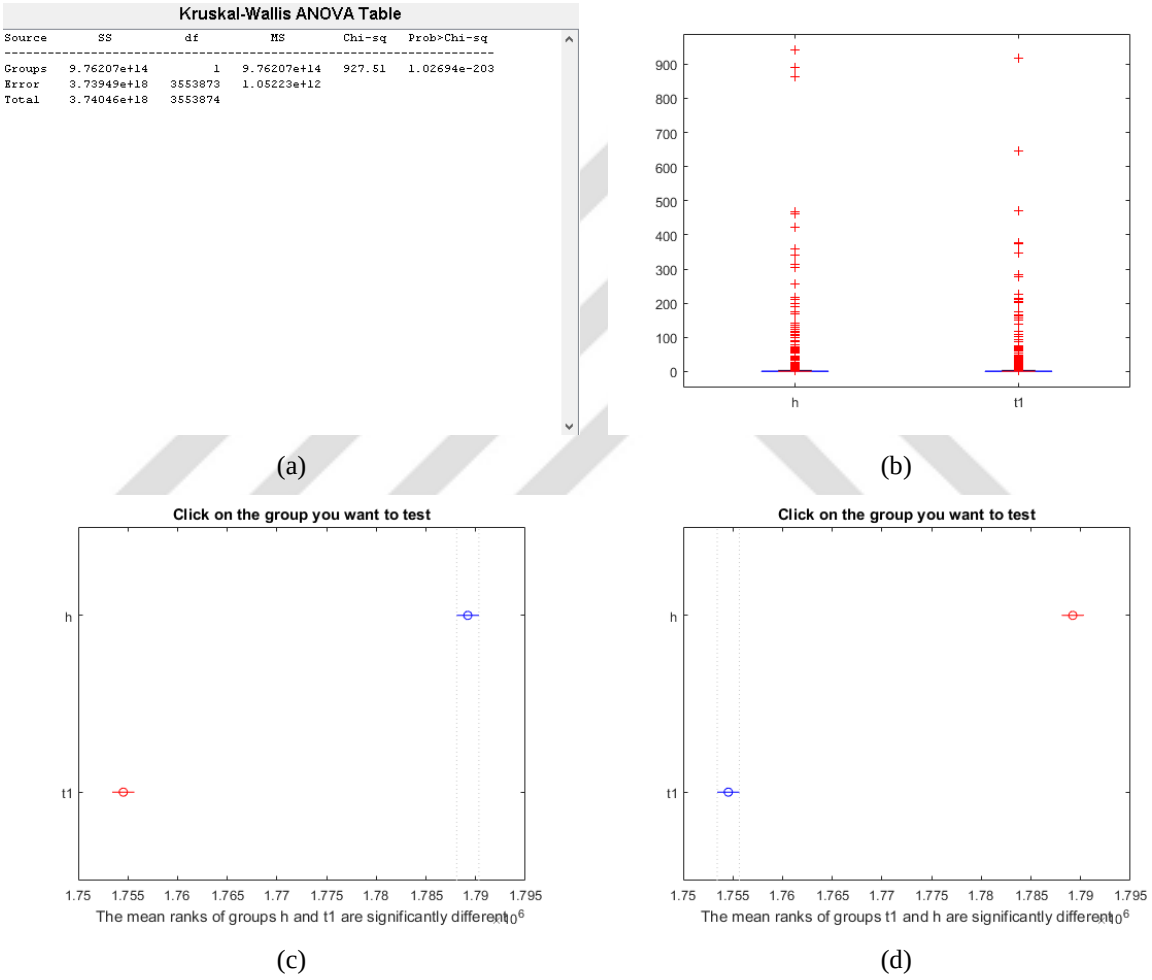
Şekil 5.11: Akciğer kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)



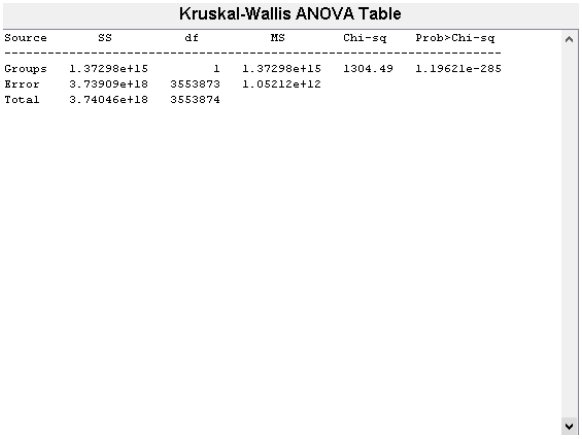
Şekil 5.12: Akciğer kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

Çerçeve boyutlarının değişimi ile akciğer kanseri veri kümesinin grupları arasında anlamlı farklılık olup olmadığı Kruskal – Wallis H testi ile gözlemlenmiştir. Sonuçlar ANOVA tablosu, kutu – bıyık grafiği ve seçilen her bir grubun diğerleri ile arasındaki anlamlı farklılığı gösteren grafikler şeklinde verilmiştir. Çerçeve boyutunun 5 olarak yapılandırılması ile bu testin uygulanması sonucunda elde edilen tüm sonuçlar Şekil 5.13'te verilmiştir. Aynı şekilde, çerçeve boyutunun 9 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen

veriler Şekil 5.14'te; çerçeve boyutunun 15 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.15'te ve son olarak çerçeve boyutunun 25 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.16'da verilmiştir. Özetle, tüm çerçeve boyutlarında gruplar arasında önemli seviyede anlamlı farklılık olduğu ($p \ll 0.05$), aykırı değerlerin dağılımının kutu – bıyık grafikleri ile gözlemlenebildiği (boyutlar arttıkça aykırı değerlerin büyüklüğünün de arttığı) ve çerçeve boyutlarının artması ile anlamlı farklılığın daha çok arttığı (grupların birbirlerinden farklılaşması) gözlemlenmiştir.



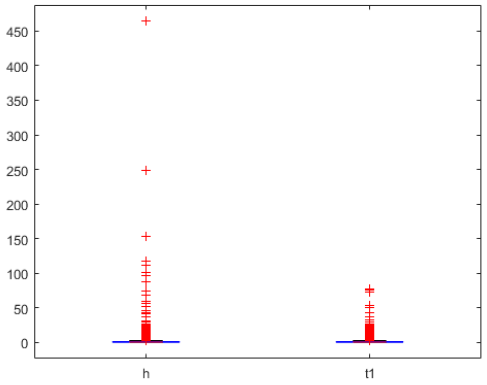
Şekil 5.13: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 5, \alpha = 0.05$).



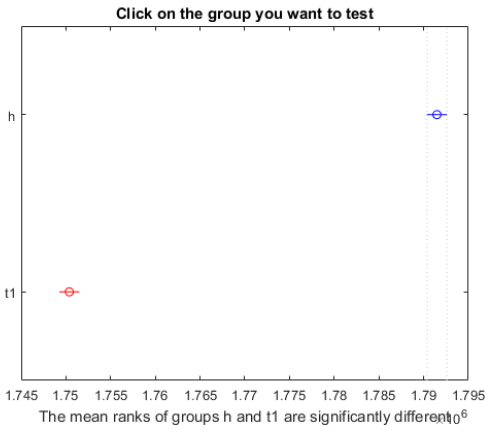
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.37298e+15	1	1.37298e+15	1304.49	1.19621e-285
Error	3.73909e+18	3553873	1.05212e+12		
Total	3.74046e+18	3553874			

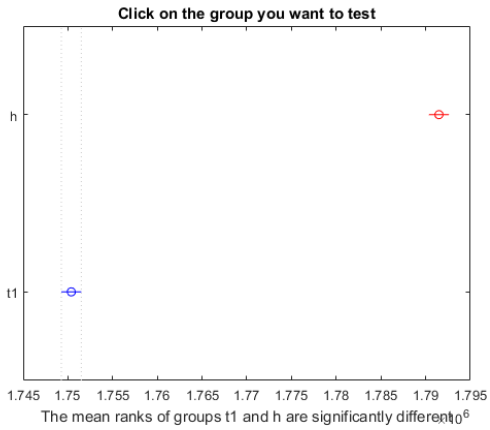
(a)



(b)

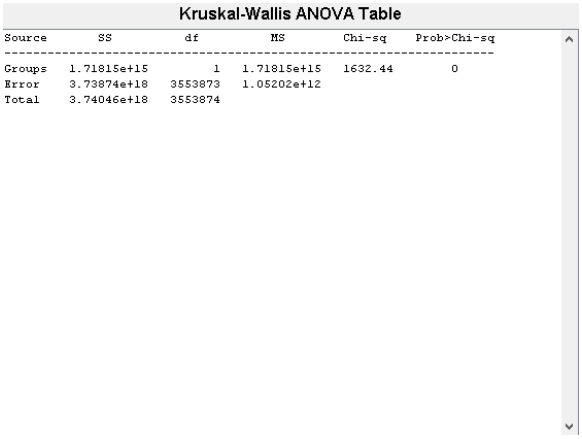


(c)

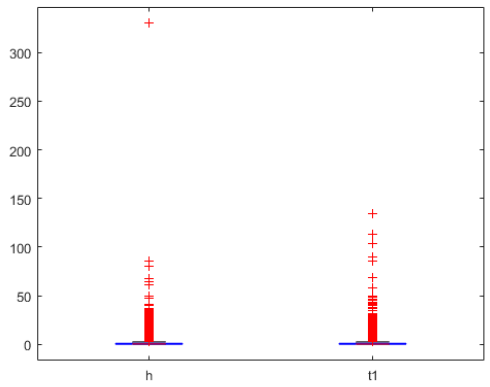


(d)

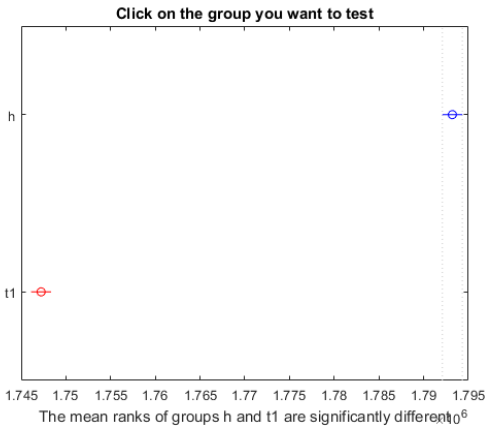
Şekil 5.14: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 9, \alpha = 0.05$).



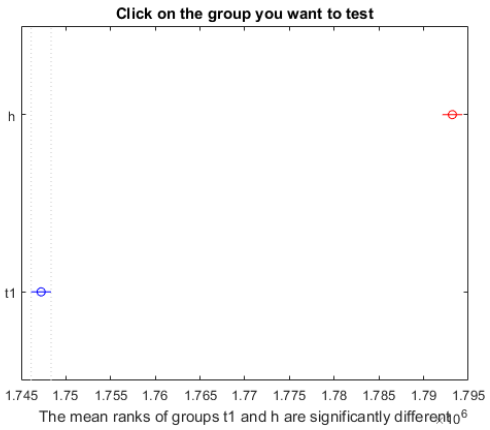
(a)



(b)

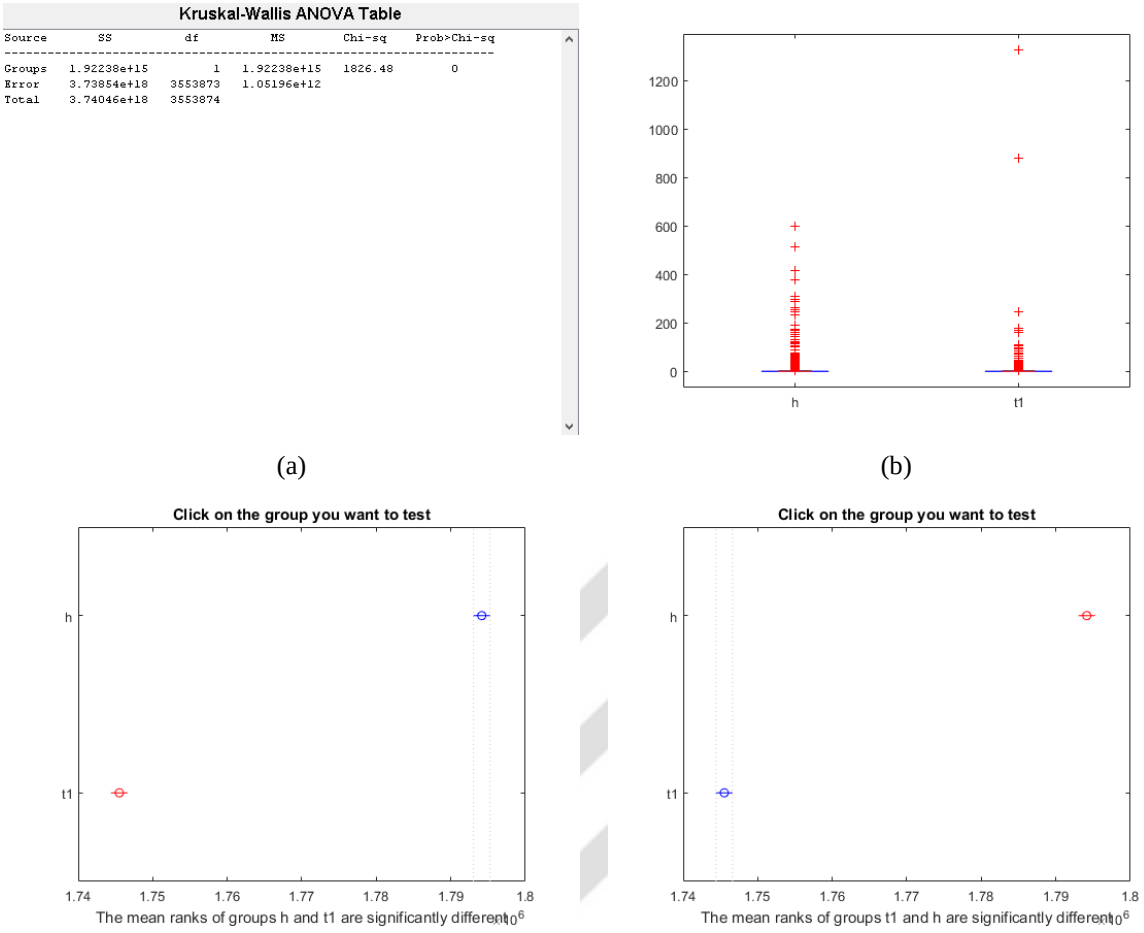


(c)



(d)

Şekil 5.15: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 15, \alpha = 0.05$).



(a) (b)

(c) (d)

Şekil 5.16: Akciğer kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 25, \alpha = 0.05$).

5.2.3. Over Kanseri Veri Kümesinin Sonuçları

SYMPES yöntemi ile iki sınıflı over kanserinin her bir sınıfı için oluşturulan evrensel ortalama ve ortanca bilgi kütüphanelerinde, çerçeve boyutlarına göre değişen GF sayısınca genlerden oluşan bir gen listesi gen çipinin sağladığı büyük gen listesi kullanılarak oluşturulmuştur. İki sınıflı olan over kanseri veri kümesinin çok sınıflı veri kümelerinden farklı olarak, Pearson korelasyon katsayısının ± 1 değerlerini içermesinden dolayı zıt ilişki durumu için geçerli olan $\rho < 0$ değerlerine karşılık gelen GF'ler anlamlı farklılık gösteren GF'ler olarak varsayılmıştır. $SPCCF_{min-max}$ yönteminde $\rho < 0$ değeri dikkate alınmıştır.

Tablo 5.11: SYMPES yönteminin değişen çerçeve boyutlarına göre over kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak).

Gruplar	Çerçeve Boyutu (L_F)	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$	$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$
H	5	5580	5593	5580 (51.03)	5593 (51.15)
	9	3207	3232	3207 (52.79)	3232 (53.20)
	15	1964	1976	1964 (53.88)	1976 (54.21)
	25	1245	1256	**1245 (56.93)**	**1256 (57.43)**
T1	5	5598	5577	5598 (51.19)	5577 (51)
	9	3182	3174	3182 (52.38)	3174 (52.25)
	15	1901	1920	1901 (52.15)	**1920 (52.67)**
	25	1153	1148	**1153 (52.72)**	1148 (52.49)

Her bir çerçevenin en büyük bilgi taşıyan indislerine karşılık gelen genlerinden oluşturulan bu yeni gen listesinin referans gen listesi ile kesişen genlerinden yeni bir küme oluşturulmuştur. Bu kesişim kümesindeki genlerin sayısı ve yüzdelik karşılığı Tablo 5.11'de verilmiştir. Çerçeve boyutunun artması ile iki gen listesinin kesişim kümesinin içerdiği önemli gen sayısı da artmıştır.

Tablo 5.12'de geliştirilen 6 yeni yöntemin her biri için yeni birer gen listesi oluşturulmuş, referans gen listesi ile kesişim kümelerindeki gen sayısı ve yüzdelik karşılığı verilmiştir. $SPCCF_{min-max}$ yöntemi H ~ T1 grubunda evrensel ortalama ve ortanca bilgi kütüphaneleri için sırasıyla %86.36 ve %85.46 ile en iyi sonucu vermiştir.

Tablo 5.12: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 5\,, \alpha = 0.5,\ \rho < 0$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$	$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$
H ~ T1	$SWRS$	1055	1125	2652 (50.27)	2695 (47.91)
	$SWRS_cor$	6	6	19 (63.33)	17 (56.67)
	SWT	17	142	49 (57.65)	214 (30.14)
	SWT_cor	1	3	1 (20)	1 (6.67)
	$SF_{min-max}$	8877	8839	8877 (81.18)	8839 (80.83)
	$SPCCF_{min-max}$	5165	5362	**5165 (86.36)**	**5362 (85.46)**

Çerçeve boyutunun 5'ten 9'a arttırılmasıyla kesişen kümedeki gen sayılarının yüzdece daha çok arttığı, bu doğrultuda bilgi sağlama açısından daha etkin sonuçlar verdiği gözlemlenmiştir (Tablo 5.13). $SPCCF_{min-max}$ yönteminin genel olarak tüm gruplarda en iyi sonucu veren yöntem olduğu karşımıza çıkmıştır.

Tablo 5.13: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 9\,, \alpha = 0.5, \rho < 0$).

Gruplar	**Yöntemler**	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1351	1253	5887 (48.42)	5383 (47.73)
	$SWRS_cor$	30	28	111 (41.11)	112 (44.44)
	SWT	8	56	39 (54.17)	149 (29.56)
	SWT_cor	1	3	7 (77.78)	2 (7.41)
	$SF_{min-max}$	5427	5291	5427 (89.33)	5291 (87.09)
	$SPCCF_{min-max}$	3099	3353	**3099 (93.32)**	**3353 (91.86)**

Çerçeve boyutunun 9'dan 15'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%97'lere varan), yine genel olarak $SPCCF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.14'te sunulmuştur.

Tablo 5.14: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 15\,, \alpha = 0.5,\ \rho < 0$).

Gruplar	**Yöntemler**	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1158	993	8283 (47.69)	7027 (47.18)
	$SWRS_cor$	48	42	331 (45.97)	281 (44.60)
	SWT	3	28	26 (57.78)	128 (30.48)
	SWT_cor	1	2	10 (66.67)	7 (23.33)
	$SF_{min-max}$	3389	3383	3389 (92.98)	3383 (92.81)
	$SPCCF_{min-max}$	1523	2206	**1523 (93.78)**	**2206 (96.16)**

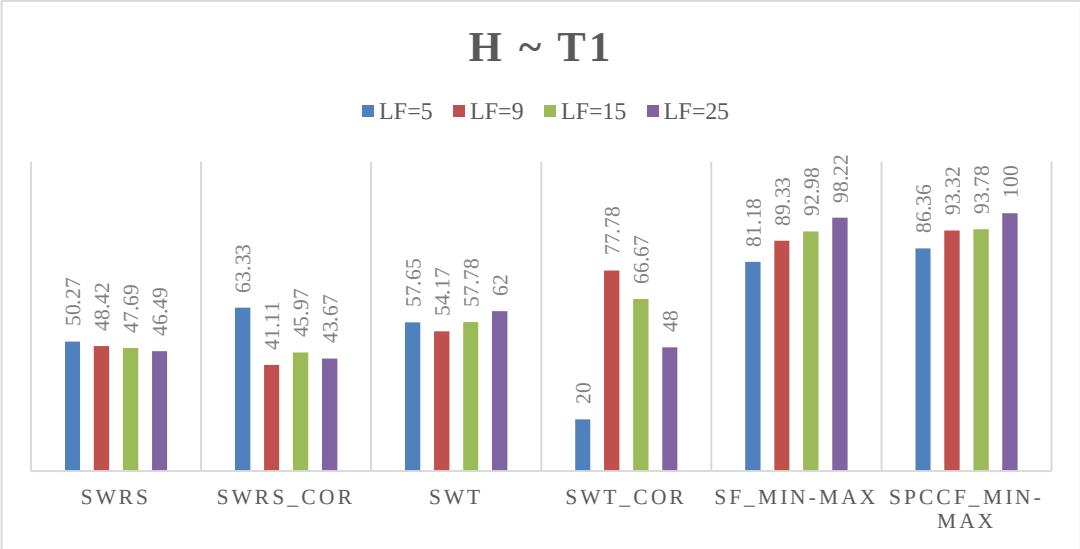
Çerçeve boyutunun 15'ten 25'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%100'e ulaştığı), yine genel olarak $SPCCF_{min-max}$

yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.15'te sunulmuştur.

Tablo 5.15: Over kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 25\,, \alpha = 0.5,\ \rho < 0$).

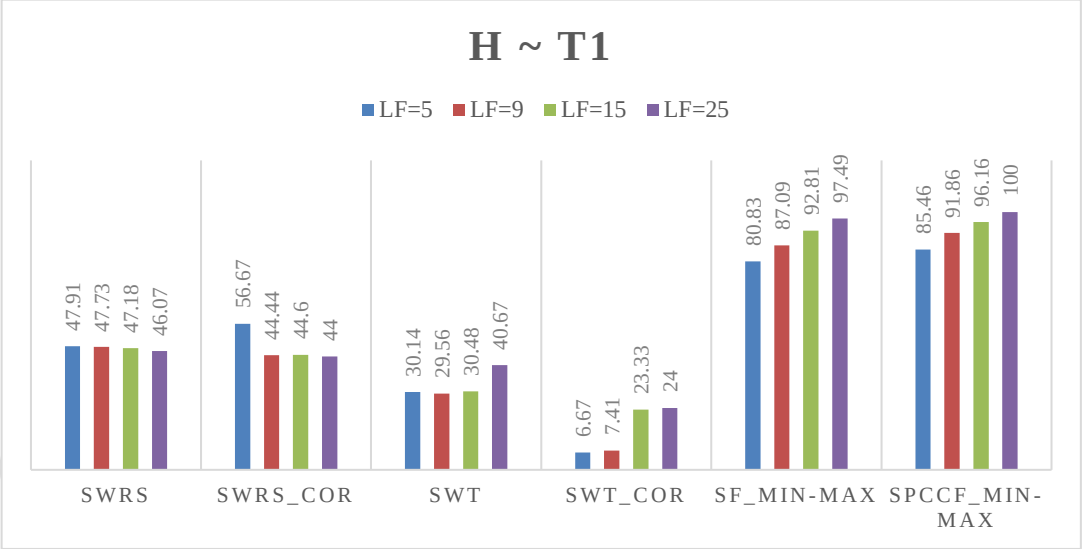
Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$	$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$
H ~ T1	$SWRS$	971	791	11286 (46.49)	9111 (46.07)
	$SWRS_cor$	61	49	666 (43.67)	539 (44)
	SWT	2	18	31 (62)	183 (40.67)
	SWT_cor	1	2	12 (48)	12 (24)
	$SF_{min-max}$	2148	2132	2148 (98.22)	2132 (97.49)
	$SPCCF_{min-max}$	976	1454	**976 (100)**	**1454 (100)**

SYMPES yöntemiyle oluşturulan evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.17'de verilmiştir.



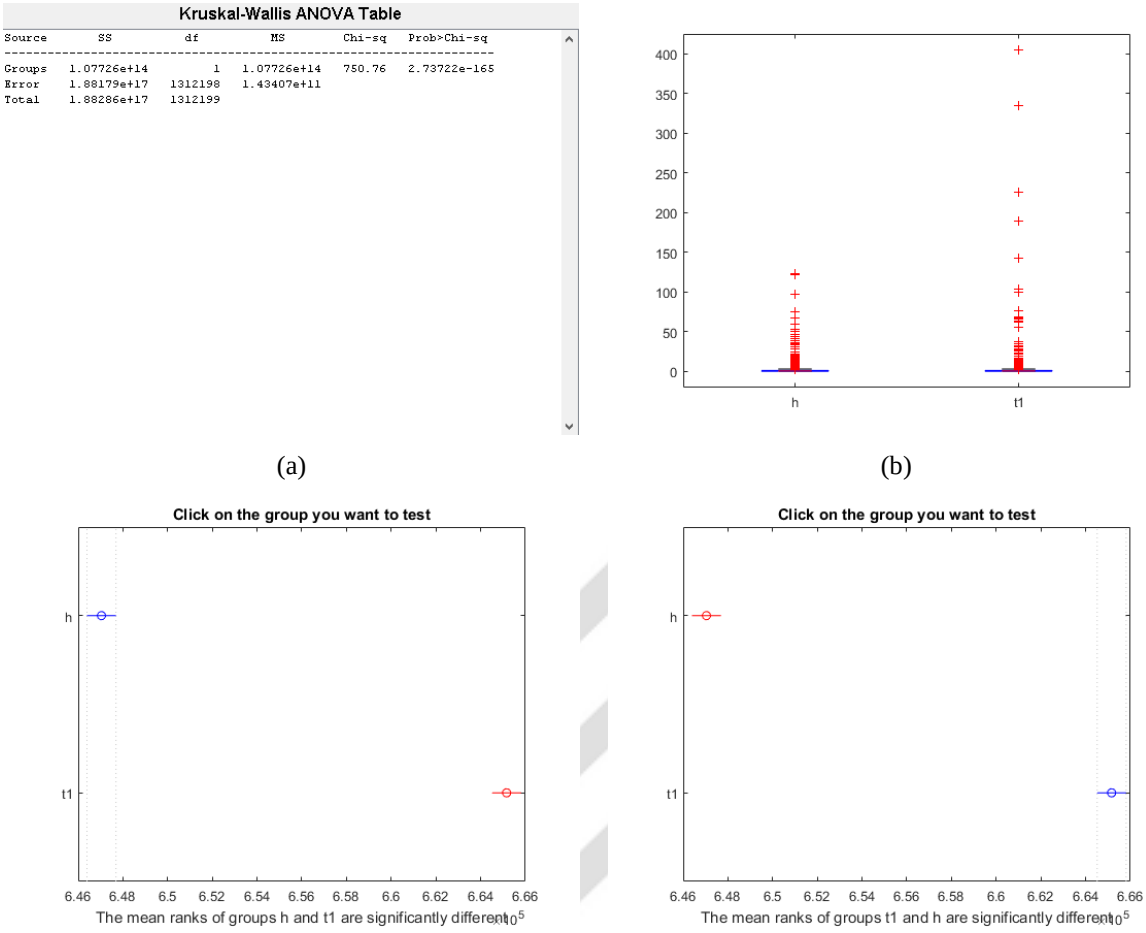
Şekil 5.17: Over kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

SYMPES yöntemiyle oluşturulan evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.18'de verilmiştir.

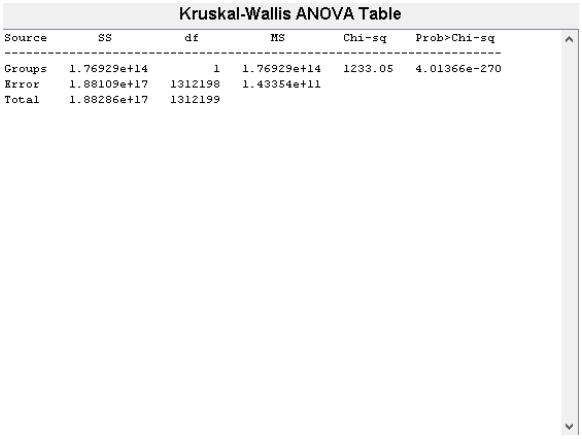


Şekil 5.18: Over kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

Çerçeve boyutlarının değişimi ile over kanseri veri kümesinin grupları arasında anlamlı farklılık olup olmadığı Kruskal – Wallis H testi ile gözlemlenmiştir. Sonuçlar ANOVA tablosu, kutu – bıyık grafiği ve seçilen her bir grubun diğerleri ile arasındaki anlamlı farklılığı gösteren grafikler şeklinde verilmiştir. Çerçeve boyutunun 5 olarak yapılandırılması ile bu testin uygulanması sonucunda elde edilen tüm sonuçlar Şekil 5.19'da verilmiştir. Aynı şekilde, çerçeve boyutunun 9 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.20'de; çerçeve boyutunun 15 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.21'de ve son olarak çerçeve boyutunun 25 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.22'de verilmiştir. Özetle, tüm çerçeve boyutlarında gruplar arasında önemli seviyede anlamlı farklılık olduğu ($p \ll 0.05$), aykırı değerlerin dağılımının kutu – bıyık grafikleri ile gözlemlenebildiği (boyutlar arttıkça aykırı değerlerin büyüklüğünün de arttığı) ve çerçeve boyutlarının artması ile anlamlı farklılığın daha çok arttığı (grupların birbirlerinden ayrıklaştığı) gözlemlenmiştir.



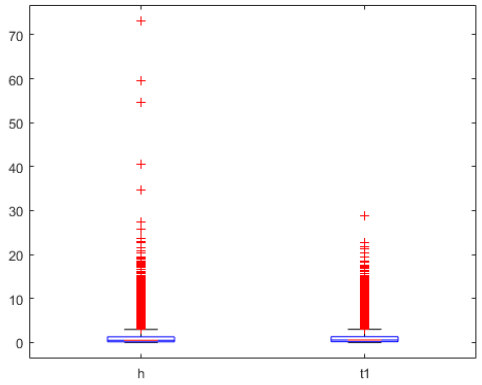
Şekil 5.19: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 5, \alpha = 0.05$).



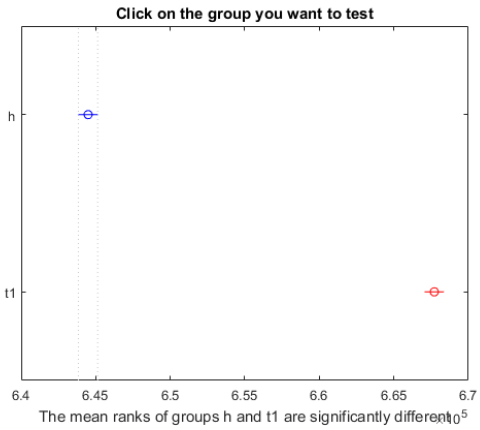
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.76929e+14	1	1.76929e+14	1233.05	4.01366e-270
Error	1.88109e+17	1312198	1.43354e+11		
Total	1.88286e+17	1312199			

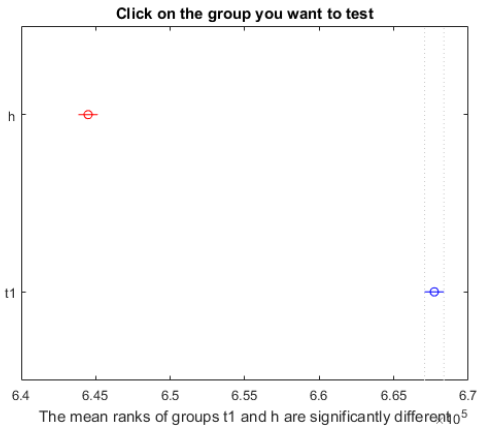
(a)



(b)

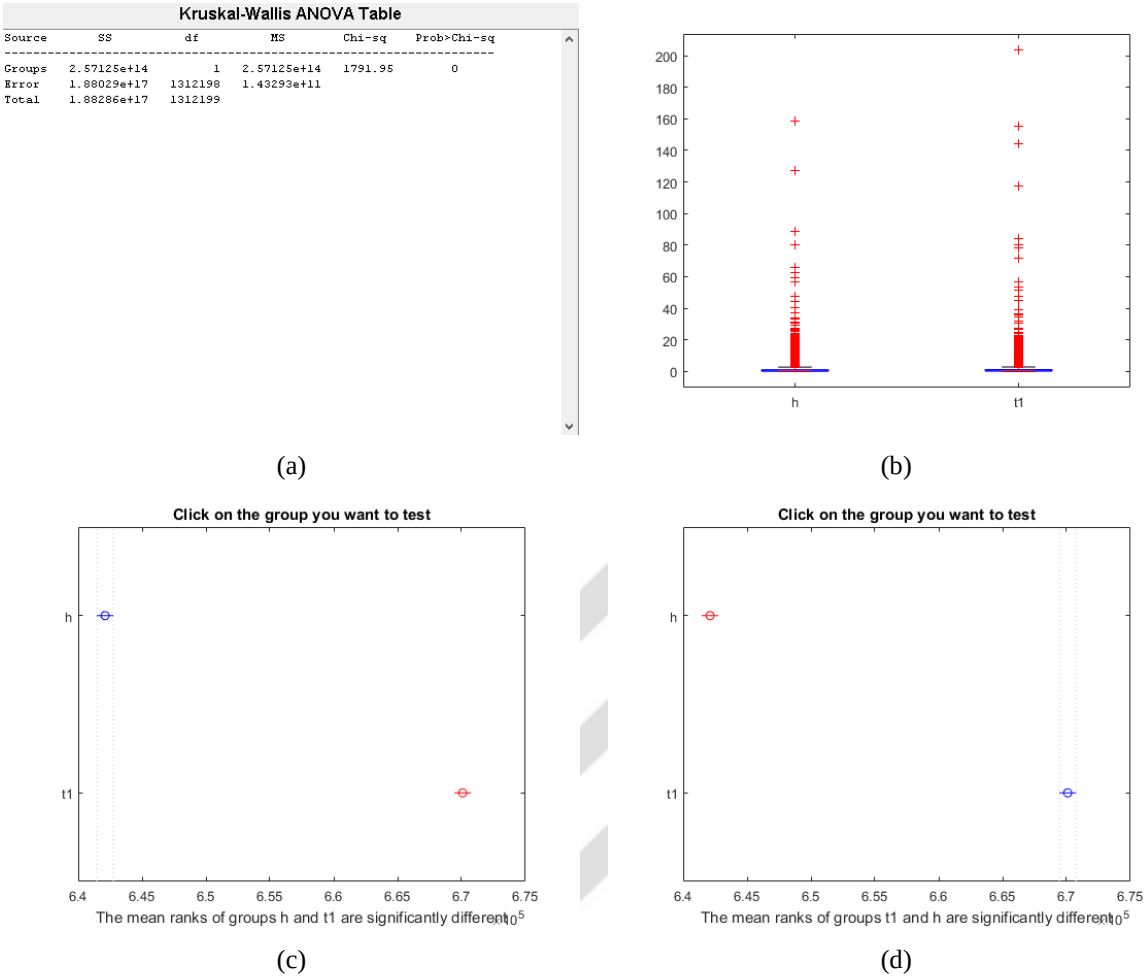


(c)



(d)

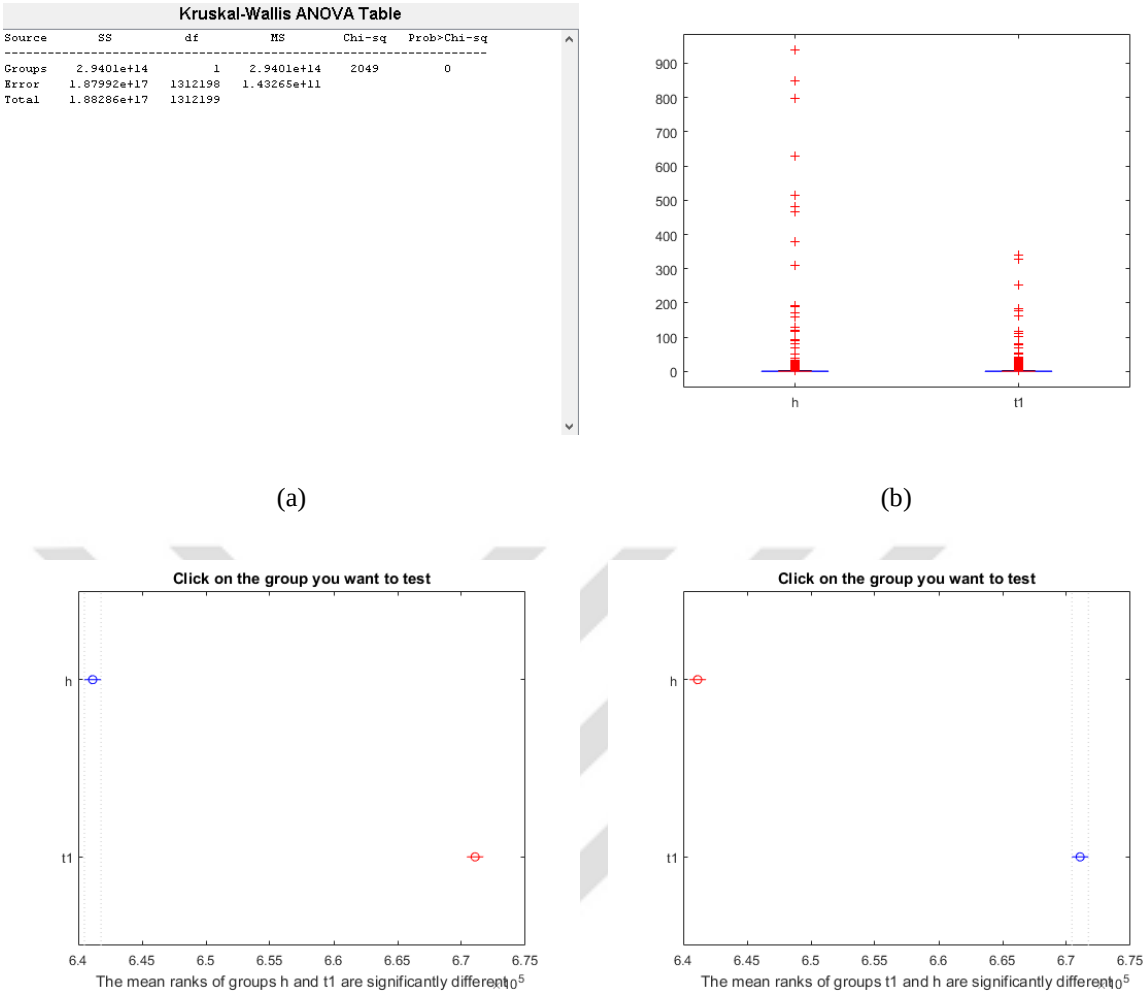
Şekil 5.20: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 9, \alpha = 0.05$).



(a) (b)

(c) (d)

Şekil 5.21: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 15, \alpha = 0.05$).



(a) (b)

(c) (d)

Şekil 5.22: Over kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 25, \alpha = 0.05$).

5.2.4. Prostat Kanseri Veri Kümesinin Sonuçları

SYMPES yöntemi ile iki sınıflı prostat kanserinin her bir sınıfı için oluşturulan evrensel ortalama ve ortanca bilgi kütüphanelerinde, çerçeve boyutlarına göre değişen GF sayısınca genlerden oluşan bir gen listesi gen çipinin sağladığı büyük gen listesi kullanılarak oluşturulmuştur. İki sınıflı olan over kanseri veri kümesinin çok sınıflı veri kümelerinden farklı olarak, Pearson korelasyon katsayısının ± 1 değerlerini içermesinden dolayı zıt ilişki durumu için geçerli olan $\rho < 0$ değerlerine karşılık gelen GF'ler anlamlı farklılık gösteren GF'ler olarak varsayılmıştır. $SPCCF_{min-max}$ yönteminde $\rho < 0$ değeri dikkate alınmıştır.

Her bir çerçevenin en büyük bilgi taşıyan indislerine karşılık gelen genlerinden oluşturulan bu yeni gen listesinin referans gen listesi ile kesişen genlerinden yeni bir küme oluşturulmuştur.

Bu kesişim kümesindeki genlerin sayısı ve yüzdelik karşılığı Tablo 5.16'da verilmiştir. Çerçeve boyutunun artması ile iki gen listesinin kesişim kümesinin içerdiği önemli gen sayısı da artmıştır.

Tablo 5.16: SYMPES yönteminin değişen çerçeve boyutlarına göre prostat kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak).

Gruplar	**Çerçeve Boyutu** (L_F)	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H	5	3579	3577	3579 (32.73)	3577 (32.71)
	9	2021	2007	2021 (33.27)	2007 (33.04)
	15	1218	1205	1218 (33.42)	1205 (33.06)
	25	742	744	**742 (33.93)**	**744 (34.02)**
T1	5	3478	3477	3478 (31.81)	3477 (31.80)
	9	1901	1906	1901 (31.29)	1906 (31.37)
	15	1145	1140	1145 (31.41)	1140 (31.28)
	25	719	715	**719 (32.88)**	**715 (32.69)**

Tablo 5.17'de, geliştirilen 6 yeni yöntemin her biri için yeni birer gen listesi oluşturulmuş, referans gen listesi ile kesişim kümelerindeki gen sayısı ve yüzdelik karşılığı verilmiştir. $SPCCF_{min-max}$ yöntemi H ~ T1 grubunda evrensel ortalama ve ortanca bilgi kütüphaneleri için sırasıyla %61.05 ve %59.69 ile en iyi sonucu vermiştir.

Tablo 5.17: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 5\,, \alpha = 0.5,\ \rho < 0$).

Gruplar	**Yöntemler**	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	376	381	650 (34.57)	621 (32.60)
	$SWRS_cor$	4	4	6 (30)	10 (50)
	SWT	--	1	0 (0)	1 (20)
	SWT_cor	--	1	0 (0)	1 (20)
	$SF_{min-max}$	6335	6298	6335 (57.93)	6298 (57.59)
	$SPCCF_{min-max}$	3233	3253	**3233 (61.05)**	**3253 (59.69)**

Çerçeve boyutunun 5'ten 9'a arttırılmasıyla kesişen kümedeki gen sayılarının yüzdece daha çok arttığı, bu doğrultuda bilgi sağlama açısından daha etkin sonuçlar verdiği gözlemlenmiştir

(Tablo 5.18). $SPCCF_{min-max}$ yönteminin genel olarak tüm gruplarda en iyi sonucu veren yöntem olduğu karşımıza çıkmıştır.

Tablo 5.18: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 9\,, \alpha = 0.5, \rho < 0$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	438	431	1176 (29.83)	1137 (29.31)
	$SWRS_cor$	7	8	21 (33.33)	21 (29.17)
	SWT	--	2	0 (0)	3 (16.67)
	SWT_cor	--	1	0 (0)	2 (22.22)
	$SF_{min-max}$	4174	4115	4174 (68.71)	4115 (67.74)
	$SPCCF_{min-max}$	2210	2185	**2210 (70.11)**	**2185 (69.83)**

Çerçeve boyutunun 9'dan 15'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%80'lere varan), yine genel olarak $SPCCF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.19'da sunulmuştur.

Tablo 5.19: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 15\,, \alpha = 0.5,\ \rho < 0$).

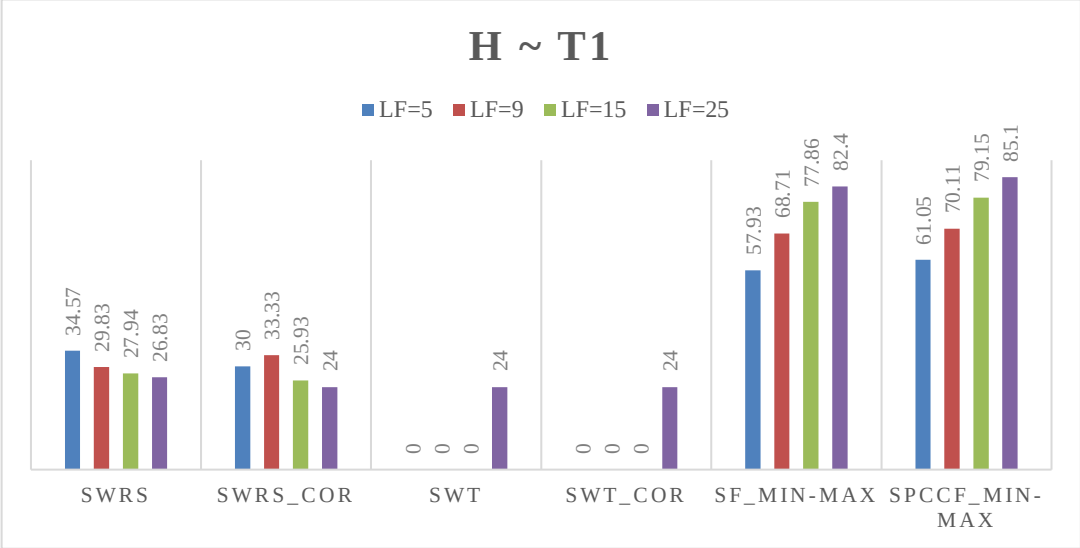
Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	310	286	1299 (27.94)	1179 (27.48)
	$SWRS_cor$	9	8	35 (25.93)	26 (21.67)
	SWT	--	--	0 (0)	0 (0)
	SWT_cor	--	--	0 (0)	0 (0)
	$SF_{min-max}$	2838	2784	2838 (77.86)	2784 (76.38)
	$SPCCF_{min-max}$	1427	1361	**1427 (79.15)**	**1361 (77.11)**

Çerçeve boyutunun 15'ten 25'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı, yine genel olarak $SPCCF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.20'de sunulmuştur.

Tablo 5.20: Prostat kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 25\,, \alpha = 0.5,\ \rho < 0$).

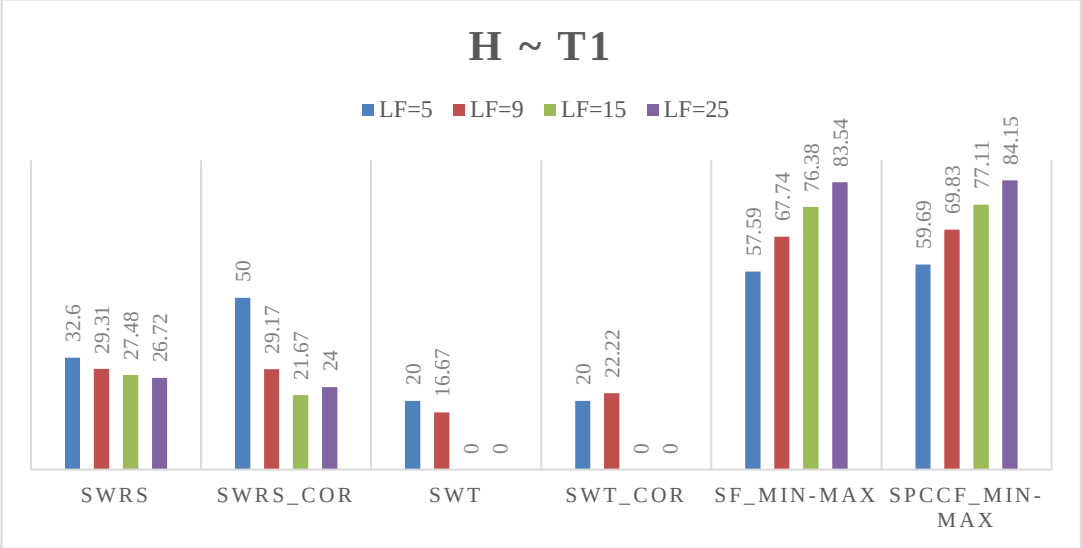
Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$	$\mu_{UE,c}$	$\tilde{\mu}_{UE,c}$
H ~ T1	$SWRS$	243	207	1630 (26.83)	1383 (26.72)
	$SWRS_cor$	11	10	66 (24)	60 (24)
	SWT	1	--	6 (24)	0 (0)
	SWT_cor	1	--	6 (24)	0 (0)
	$SF_{min-max}$	1802	1827	1802 (82.40)	1827 (83.54)
	$SPCCF_{min-max}$	937	961	**937 (85.10)**	**961 (84.15)**

SYMPES yöntemiyle oluşturulan evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.23'te verilmiştir.



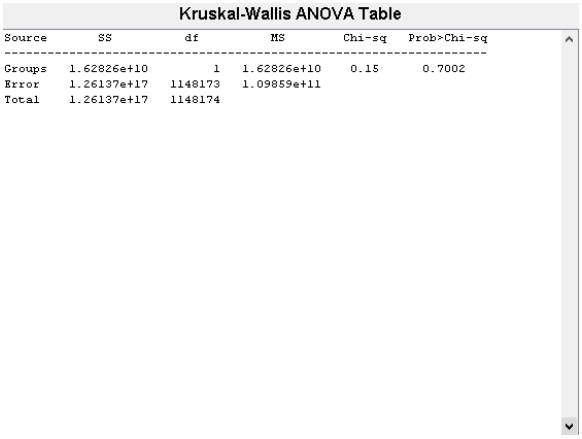
Şekil 5.23: Prostat kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

SYMPES yöntemiyle oluşturulan evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.24'te verilmiştir.

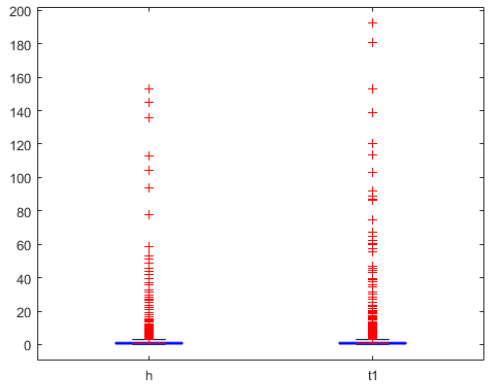


Şekil 5.24: Prostat kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

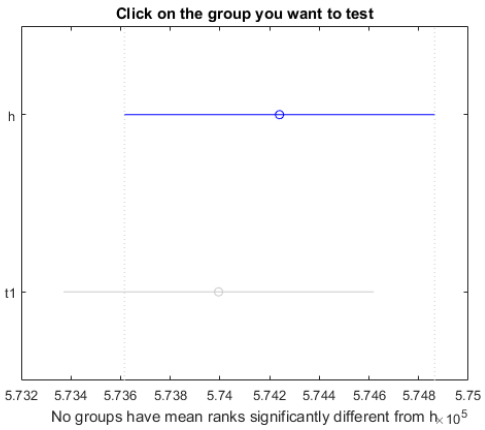
Çerçeve boyutlarının değişimi ile prostat kanseri veri kümesinin grupları arasında anlamlı farklılık olup olmadığı Kruskal – Wallis H testi ile gözlemlenmiştir. Sonuçlar ANOVA tablosu, kutu – bıyık grafiği ve seçilen her bir grubun diğerleri ile arasındaki anlamlı farklılığı gösteren grafikler şeklinde verilmiştir. Çerçeve boyutunun 5 olarak yapılandırılması ile bu testin uygulanması sonucunda elde edilen tüm sonuçlar Şekil 5.25'te verilmiştir. Aynı şekilde, çerçeve boyutunun 9 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.26'da; çerçeve boyutunun 15 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.27'de ve son olarak çerçeve boyutunun 25 olarak yapılandırıldığı durumda testin uygulanması sonucunda elde edilen veriler Şekil 5.28'de verilmiştir. Özetle, tüm çerçeve boyutlarında gruplar arasında önemli seviyede anlamlı farklılık olmadığı ($p > 0.05$) gözlemlenmiştir.



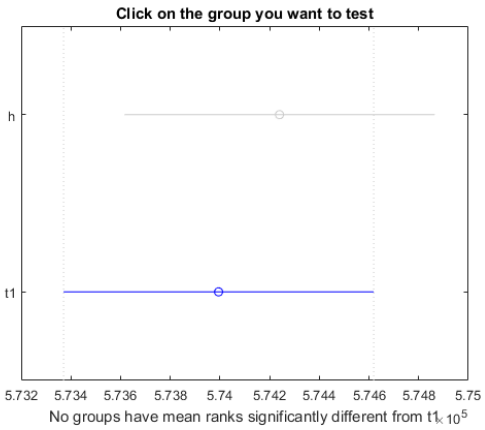
(a)



(b)

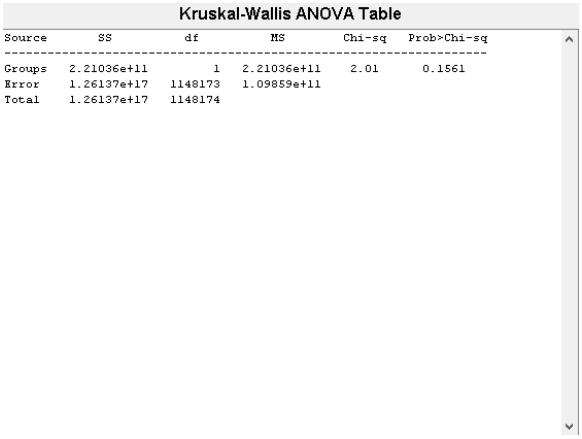


(c)



(d)

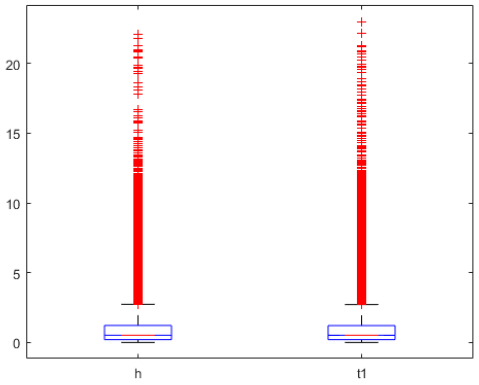
Şekil 5.25: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 5, \alpha = 0.05$).



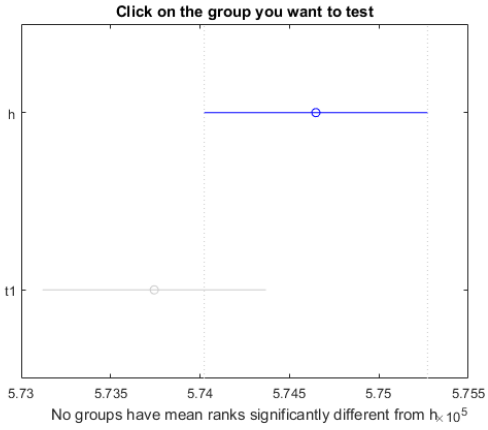
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	2.21036e+11	1	2.21036e+11	2.01	0.1561
Error	1.26137e+17	1148173	1.09859e+11		
Total	1.26137e+17	1148174			

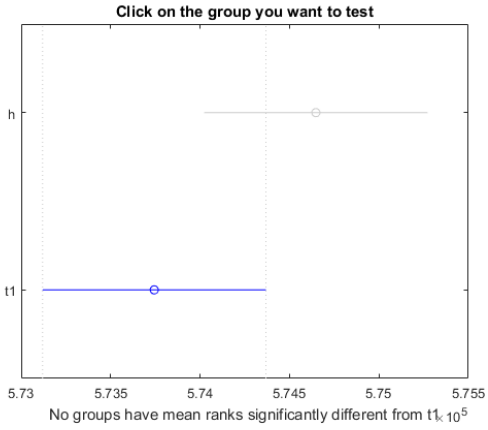
(a)



(b)

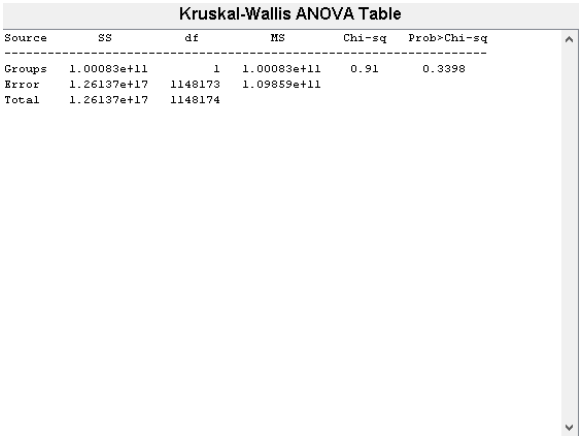


(c)



(d)

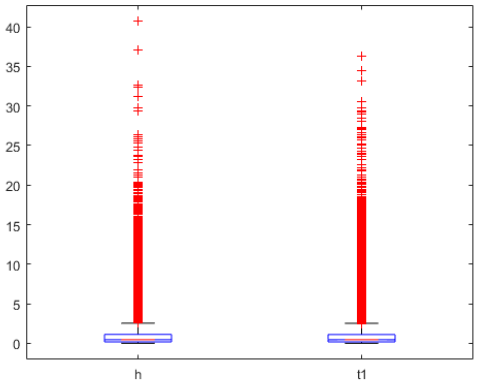
Şekil 5.26: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 9, \alpha = 0.05$).



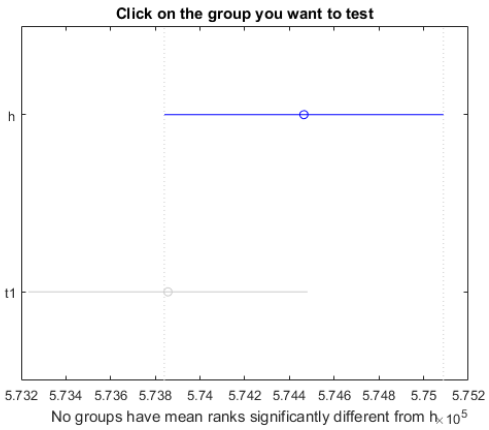
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	1.00083e+11	1	1.00083e+11	0.91	0.3398
Error	1.26137e+17	1148173	1.09859e+11		
Total	1.26137e+17	1148174			

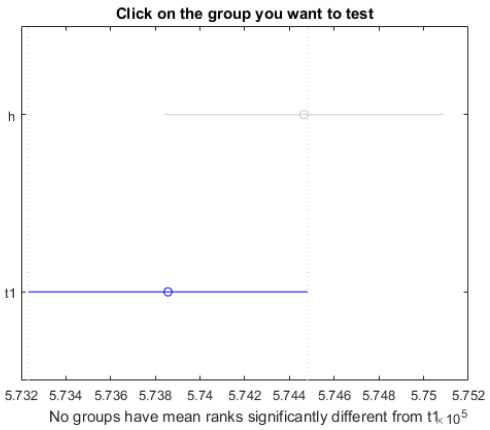
(a)



(b)

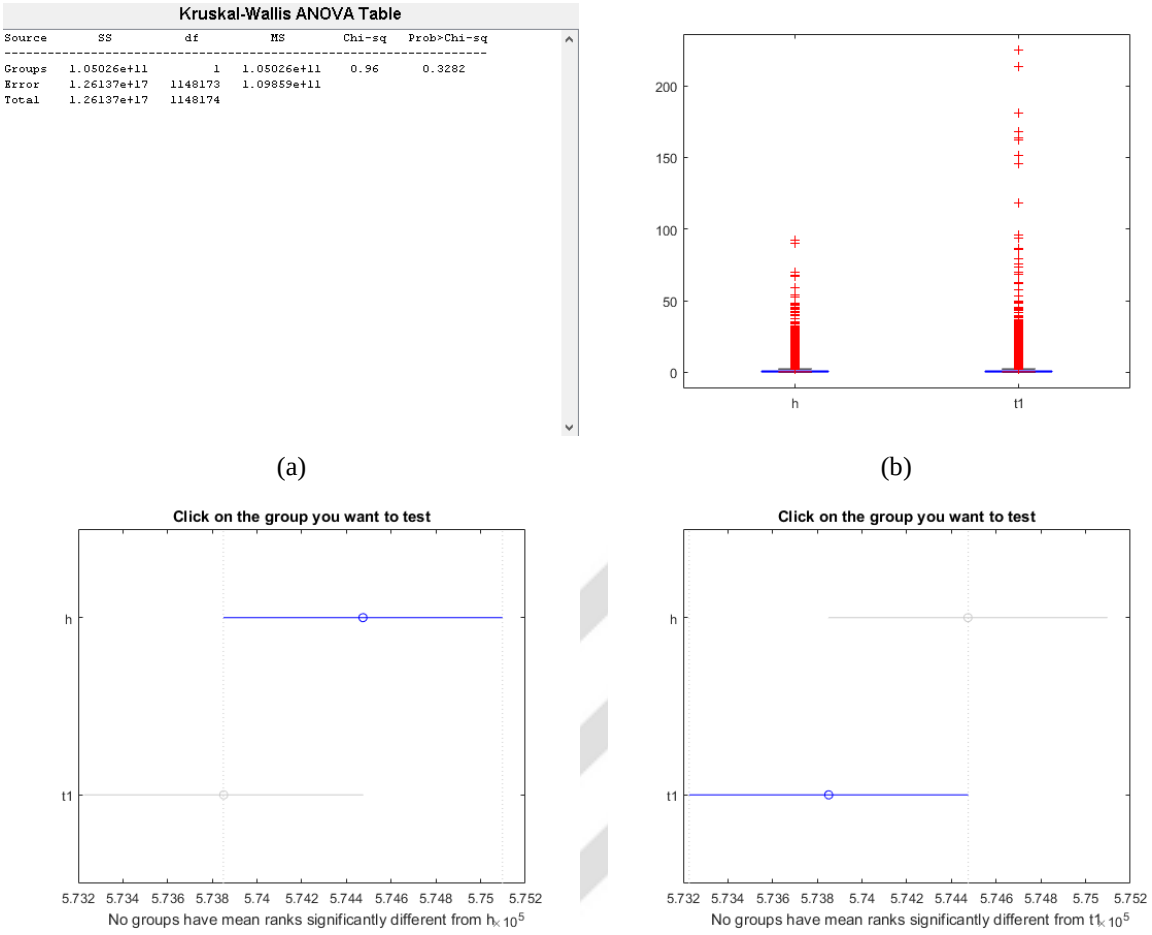


(c)



(d)

Şekil 5.27: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 15, \alpha = 0.05$).



(a) (b)

(c) (d)

Şekil 5.28: Prostat kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 25, \alpha = 0.05$).

5.2.5. Pankreas Kanseri Veri Kümesinin Sonuçları

SYMPES yöntemi ile dört sınıflı pankreas kanserinin her bir sınıfı için oluşturulan evrensel ortalama ve ortanca bilgi kütüphanelerinde, çerçeve boyutlarına göre değişen GF sayısınca genlerden oluşan bir gen listesi gen çipinin sağladığı büyük gen listesi kullanılarak oluşturulmuştur.

Her bir çerçevenin en büyük bilgi taşıyan indislerine karşılık gelen genlerinden oluşturulan bu yeni gen listesinin referans gen listesi ile kesişen genlerinden yeni bir küme oluşturulmuştur. Bu kesişim kümesindeki genlerin sayısı ve yüzdelik karşılığı Tablo 5.21'de verilmiştir. Çerçeve boyutunun artması ile iki gen listesinin kesişim kümesinin içerdiği önemli gen sayısı da artmıştır.

Tablo 5.21: SYMPES yönteminin değişen çerçeve boyutlarına göre pankreas kanseri veri kümesi için keşfettiği önemli GF ve gen sayısı (her bir çerçevenin en büyük bilgi taşıyan indisi esas alınarak).

Gruplar	**Çerçeve Boyutu** (L_F)	**Önemli GF Sayısı**		**Ortak Önemli Gen Sayısı (%)**	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H	5	5949	5953	5949 (54.40)	5953 (54.44)
	9	3399	3408	3399 (55.95)	3408 (56.10)
	15	2114	2118	2114 (58)	2118 (58.11)
	25	1307	1310	**1307 (59.76)**	**1310 (59.90)**
T1	5	5676	5692	5676 (51.91)	5692 (52.05)
	9	3297	3296	3297 (54.27)	3296 (54.26)
	15	2035	2044	2035 (55.83)	2044 (56.08)
	25	1261	1267	**1261 (57.66)**	**1267 (57.93)**
T2	5	5676	5677	5676 (51.91)	5677 (51.92)
	9	3242	3268	3242 (53.37)	3268 (53.79)
	15	2019	2045	2019 (55.39)	2045 (56.10)
	25	1256	1273	**1256 (57.43)**	**1273 (58.21)**
T3	5	5740	5737	5740 (52.49)	5737 (52.46)
	9	3351	3347	3351 (55.16)	3347 (55.09)
	15	2100	2088	2100 (57.61)	2088 (57.28)
	25	1295	1299	**1295 (59.21)**	**1299 (59.40)**

Tablo 5.22'de, geliştirilen 6 yeni yöntemin her biri için yeni birer gen listesi oluşturulmuş, referans gen listesi ile kesişim kümelerindeki gen sayısı ve yüzdelik karşılığı verilmiştir. $SPCCF_{min-max}$ yöntemi tüm grupların evrensel ortalama ve ortanca bilgi kütüphaneleri için en iyi sonucu vermiştir.

Tablo 5.22: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 5\,, \alpha = 0.5,\ |\rho| \leq 0.1$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1335	1251	2947 (44.15)	2809 (44.91)
	$SWRS_cor$	6	6	20 (66.67)	21 (70)
	SWT	22	103	58 (52.73)	166 (32.23)
	SWT_cor	2	2	2 (20)	1 (10)
	$SF_{min-max}$	8925	8872	8925 (81.62)	8872 (81.13)
	$SPCCF_{min-max}$	615	659	**615 (82.22)**	**659 (82.17)**
H ~ T2	$SWRS$	1251	1273	2697 (43.12)	2761 (43.38)
	$SWRS_cor$	6	6	10 (33.33)	12 (40)
	SWT	33	116	92 (55.76)	201 (34.66)
	SWT_cor	1	2	1 (20)	5 (50)
	$SF_{min-max}$	8619	8611	8619 (78.82)	8611 (78.75)
	$SPCCF_{min-max}$	626	707	**626 (83.69)**	**707 (88.15)**
H ~ T3	$SWRS$	1574	1590	3482 (44.24)	3463 (43.56)
	$SWRS_cor$	6	6	16 (53.33)	6 (20)
	SWT	55	201	142 (51.64)	350 (34.83)
	SWT_cor	2	4	5 (50)	5 (25)
	$SF_{min-max}$	8563	8568	8563 (78.31)	8568 (78.35)
	$SPCCF_{min-max}$	633	687	**633 (84.63)**	**687 (85.66)**

Çerçeve boyutunun 5'ten 9'a arttırılmasıyla kesişen kümedeki gen sayılarının yüzdece daha çok arttığı, bu doğrultuda bilgi sağlama açısından daha etkin sonuçlar verdiği gözlemlenmiştir (Tablo 5.23). $SPCCF_{min-max}$ yönteminin genel olarak tüm gruplarda en iyi sonucu veren yöntem olduğu karşımıza çıkmıştır.

Tablo 5.23: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 9\,, \alpha = 0.5,\ |\rho| \leq 0.1$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1621	1431	6093 (41.76)	5464 (42.43)
	$SWRS_cor$	46	39	182 (43.96)	180 (51.28)
	SWT	14	37	71 (56.35)	116 (34.83)
	SWT_cor	2	2	10 (55.56)	10 (55.56)
	$SF_{min-max}$	5630	5620	**5630 (92.67)**	5620 (92.51)
	$SPCCF_{min-max}$	349	405	349 (91.36)	**405 (99.02)**
H ~ T2	$SWRS$	1508	1491	5533 (40.77)	5514 (41.09)
	$SWRS_cor$	37	48	152 (45.65)	208 (48.15)
	SWT	12	44	59 (54.63)	153 (38.64)
	SWT_cor	2	2	13 (72.22)	6 (33.33)
	$SF_{min-max}$	5476	5444	5476 (90.14)	5444 (89.61)
	$SPCCF_{min-max}$	361	367	**361 (94.50)**	**367 (89.73)**
H ~ T3	$SWRS$	1928	1773	7291 (42.02)	6800 (42.61)
	$SWRS_cor$	48	44	180 (41.67)	182 (45.96)
	SWT	30	82	158 (58.52)	291 (39.43)
	SWT_cor	2	3	13 (72.22)	7 (25.93)
	$SF_{min-max}$	5406	5382	5406 (88.99)	5382 (88.59)
	$SPCCF_{min-max}$	359	368	**359 (93.98)**	**368 (89.98)**

Çerçeve boyutunun 9'dan 15'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (%100'e varan), yine genel olarak $SPCCF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.24'te sunulmuştur.

Tablo 5.24: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 15\,, \alpha = 0.5,\ |\rho| \leq 0.1$).

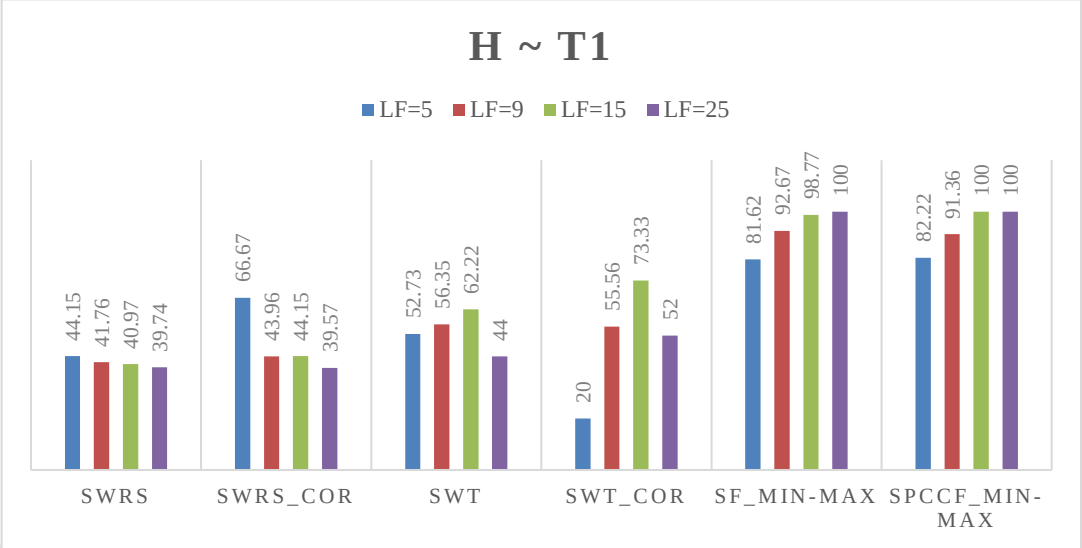
Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1389	1212	8537 (40.97)	7601 (41.81)
	$SWRS_cor$	61	56	404 (44.15)	367 (43.69)
	SWT	6	19	56 (62.22)	129 (45.26)
	SWT_cor	1	2	11 (73.33)	16 (53.33)
	$SF_{min-max}$	3600	3515	3600 (98.77)	3515 (96.43)
	$SPCCF_{min-max}$	220	253	**220 (100)**	**253 (100)**
H ~ T2	$SWRS$	1281	1166	7641 (39.77)	7038 (40.24)
	$SWRS_cor$	57	56	373 (43.63)	373 (44.40)
	SWT	7	25	60 (57.14)	167 (44.53)
	SWT_cor	2	2	25 (83.33)	23 (76.67)
	$SF_{min-max}$	3571	3531	**3571 (97.97)**	3531 (96.87)
	$SPCCF_{min-max}$	213	253	213 (96.82)	**253 (100)**
H ~ T3	$SWRS$	1646	1454	9955 (40.32)	9013 (41.33)
	$SWRS_cor$	72	64	465 (43.06)	420 (43.75)
	SWT	14	38	107 (50.95)	202 (35.44)
	SWT_cor	2	2	24 (80)	14 (46.67)
	$SF_{min-max}$	3455	3439	**3455 (94.79)**	3439 (94.35)
	$SPCCF_{min-max}$	198	253	198 (90)	**253 (100)**

Çerçeve boyutunun 15'ten 25'e çıkarılmasıyla birlikte, tüm gruplarda kesişen önemli gen kümelerinin yüzdelik olarak arttığı (kimi gruplarda %100'e varan), genel olarak $SF_{min-max}$ yönteminin en iyi sonuçları verdiği gözlemlenmiş ve elde edilen sonuçlar Tablo 5.25'te sunulmuştur.$SPCCF_{min-max}$ yönteminin de %96 ila %100 aralığında önemli genleri seçebildiği gruplar gözlemlenmiştir.

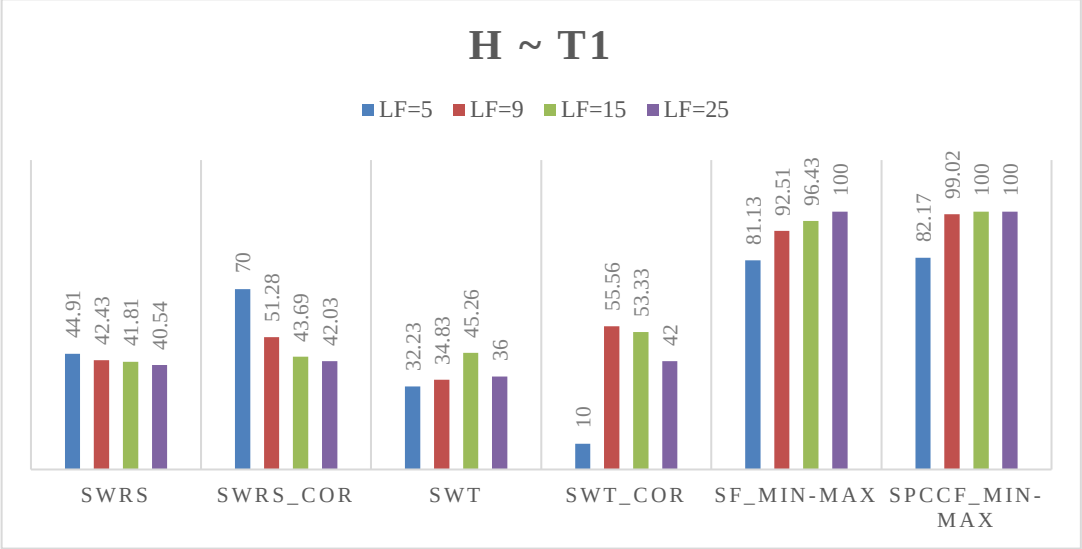
Tablo 5.25: Pankreas kanseri veri kümesi için yöntemlerin keşfettikleri önemli GF ve genlerin sayısı ($L_F = 25\,, \alpha = 0.5,\ |\rho| \leq 0.1$).

Gruplar	Yöntemler	Önemli GF Sayısı		Ortak Önemli Gen Sayısı (%)	
		$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$	$\boldsymbol{\mu_{UE,c}}$	$\boldsymbol{\tilde{\mu}_{UE,c}}$
H ~ T1	$SWRS$	1105	927	10978 (39.74)	9394 (40.54)
	$SWRS_cor$	75	59	742 (39.57)	620 (42.03)
	SWT	6	10	66 (44)	90 (36)
	SWT_cor	1	2	13 (52)	21 (42)
	$SF_{min-max}$	2187	2187	**2187 (100)**	**2187 (100)**
	$SPCCF_{min-max}$	142	101	**142 (100)**	**101 (100)**
H ~ T2	$SWRS$	1006	912	9717 (38.64)	8949 (39.25)
	$SWRS_cor$	65	59	636 (39.14)	594 (40.27)
	SWT	5	18	53 (42.40)	193 (42.89)
	SWT_cor	1	2	13 (52)	25 (50)
	$SF_{min-max}$	2187	2187	**2187 (100)**	**2187 (100)**
	$SPCCF_{min-max}$	140	97	140 (98.59)	97 (96.04)
H ~ T3	$SWRS$	1305	1112	12767 (39.13)	11157 (40.13)
	$SWRS_cor$	87	70	880 (40.46)	733 (41.89)
	SWT	5	17	45 (36)	131 (30.82)
	SWT_cor	1	2	10 (40)	25 (50)
	$SF_{min-max}$	2147	2156	**2147 (98.17)**	**2156 (98.58)**
	$SPCCF_{min-max}$	138	97	138 (97.18)	97 (96.04)

SYMPES yöntemiyle oluşturulan evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H ~ T1 karşılaştırma grubu için, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.29'da verilmiştir. Benzer şekilde, SYMPES yöntemiyle oluşturulan evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.30'da verilmiştir.

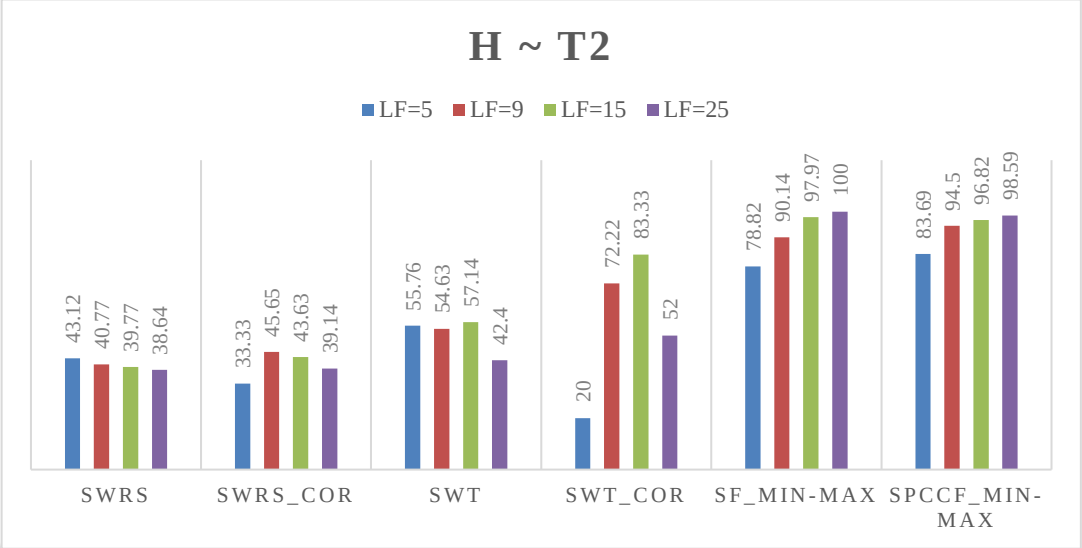


Şekil 5.29: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

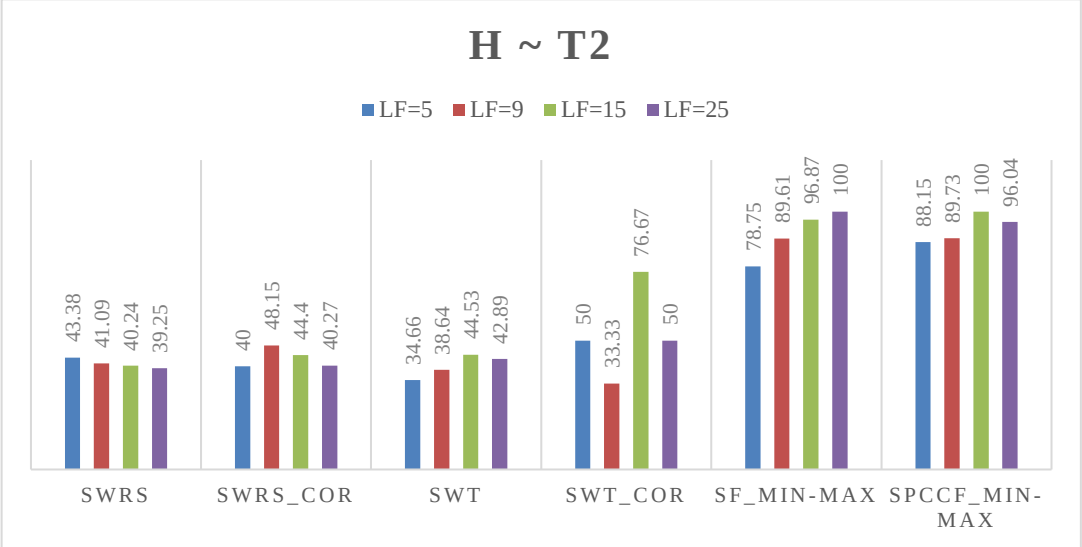


Şekil 5.30: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T1 karşılaştırma grubu için)

SYMPES yöntemiyle oluşturulan evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H ~ T2 karşılaştırma grubu için, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.31'de verilmiştir. Benzer şekilde, SYMPES yöntemiyle oluşturulan evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.32'de verilmiştir.

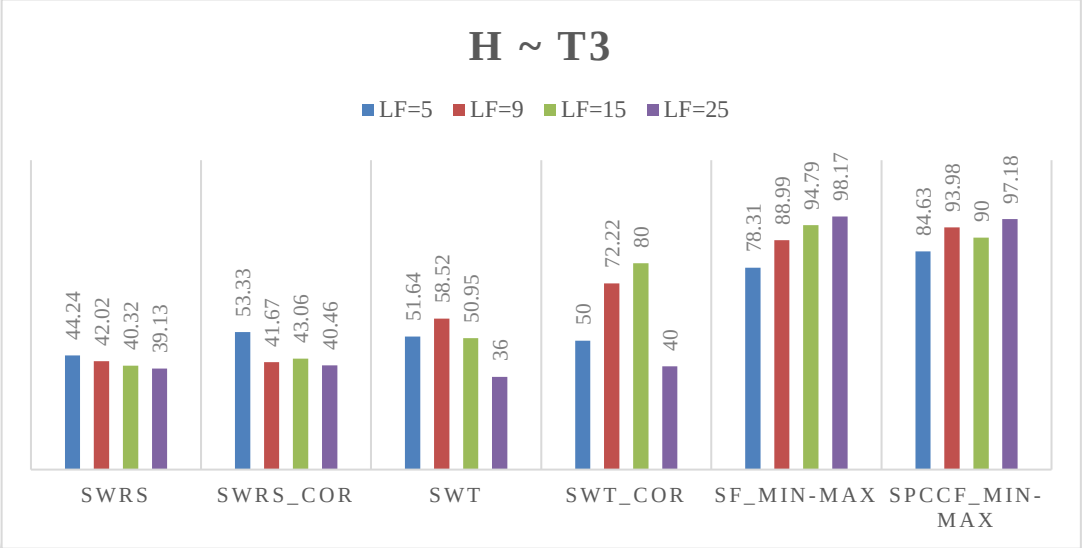


Şekil 5.31: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için)

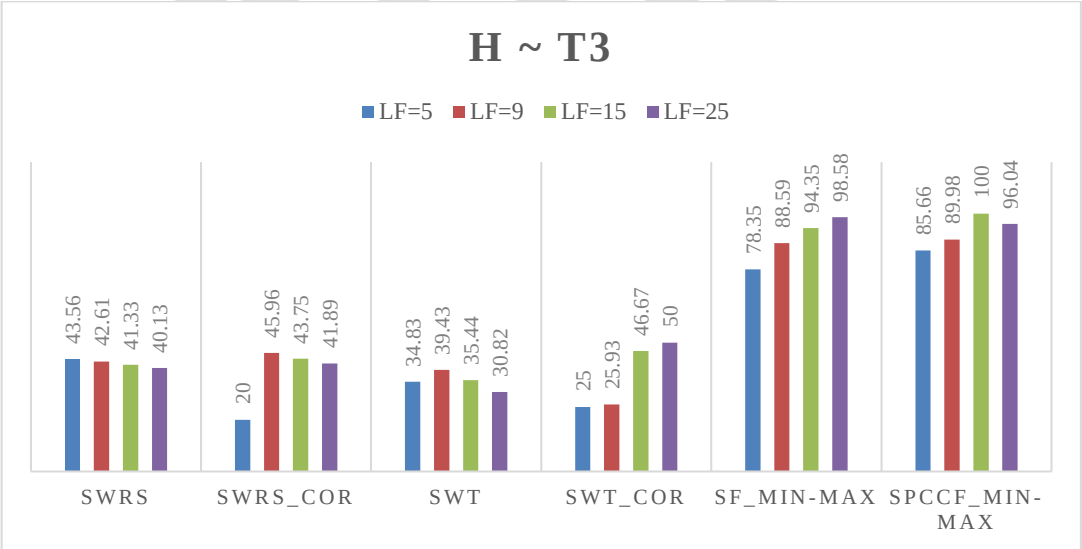


Şekil 5.32: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T2 karşılaştırma grubu için)

SYMPES yöntemiyle oluşturulan evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H ~ T3 karşılaştırma grubu için, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.33'te verilmiştir. Benzer şekilde, SYMPES yöntemiyle oluşturulan evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, 6 yöntemin değişen çerçeve uzunluklarına göre buldukları ortak önemli genlerin yüzdelik cinsinden oranları Şekil 5.34'te verilmiştir.



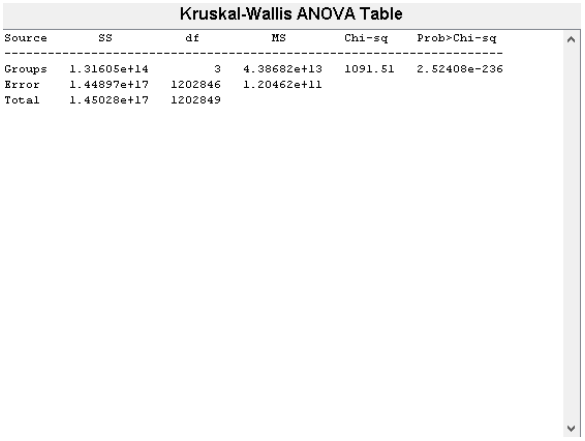
Şekil 5.33: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için)



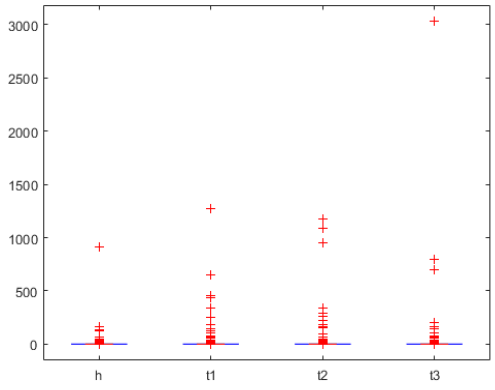
Şekil 5.34: Pankreas kanseri verisi için yöntemlerin çerçeve uzunluğuna bağlı olarak yüzdelik cinsinden önemli genleri bulma oranları (evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, H~T3 karşılaştırma grubu için)

Çerçeve boyutlarının değişimi ile prostat kanseri veri kümesinin grupları arasında anlamlı farklılık olup olmadığı Kruskal – Wallis H testi ile gözlemlenmiştir. Sonuçlar ANOVA tablosu, kutu – bıyık grafiği ve seçilen her bir grubun diğerleri ile arasındaki anlamlı farklılığı gösteren grafikler şeklinde verilmiştir. Çerçeve boyutunun 5 olarak yapılandırılması ile bu testin uygulanması sonucunda elde edilen tüm sonuçlar Şekil 5.35'te; çerçeve boyutunun 9 olarak yapılandırıldığı durumda elde edilen veriler Şekil 5.36'da; çerçeve boyutunun 15 olarak

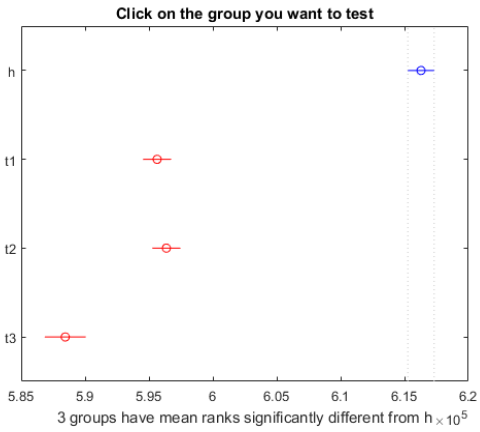
yapılandırıldığı durumda elde edilen veriler Şekil 5.37'de ve son olarak çerçeve boyutunun 25 olarak yapılandırıldığı durumda elde edilen veriler Şekil 5.38'de verilmiştir. Özetle, tüm çerçeve boyutlarında gruplar arasında önemli seviyede anlamlı farklılık olduğu ($p \ll 0.05$), aykırı değerlerin dağılımının kutu – bıyık grafikleri ile gözlemlenebildiği (boyutlar arttıkça aykırı değerlerin büyüklüğünün de arttığı) ve çerçeve boyutlarının artması ile anlamlı farklılığın daha çok arttığı (grupların birbirlerinden ayrıklaştığı) gözlemlenmiştir.



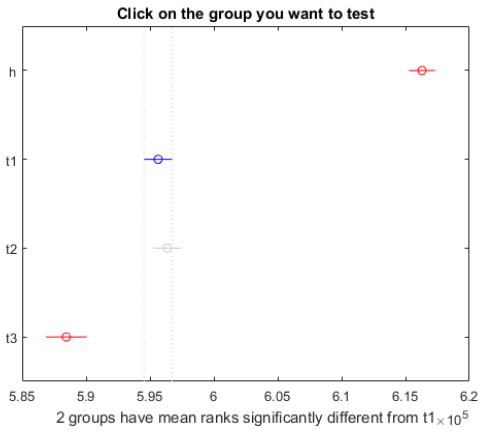
(a)



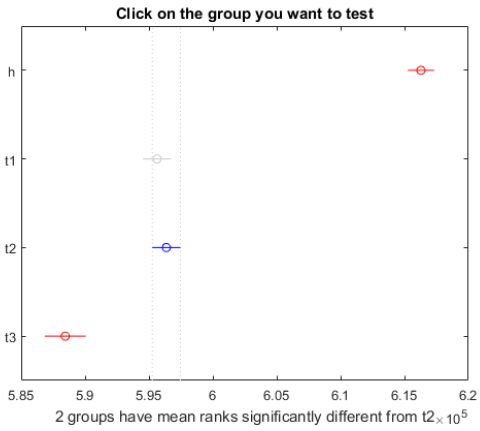
(b)



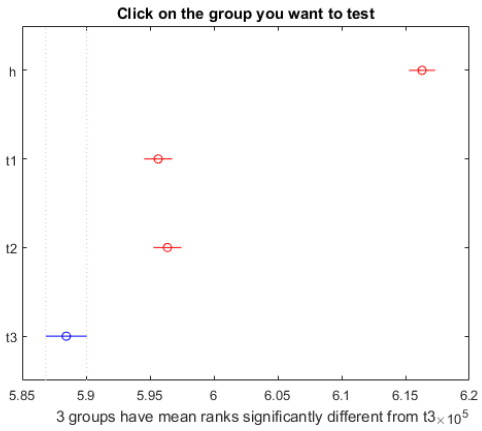
(c)



(d)

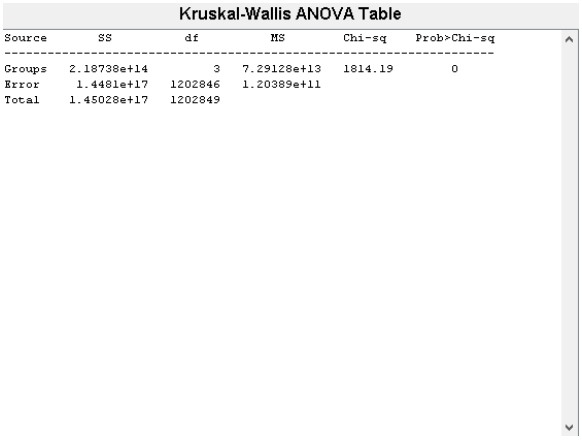


(e)

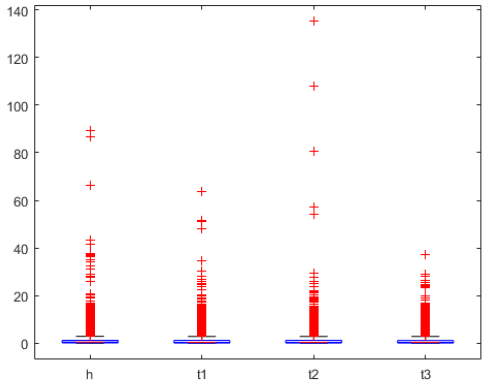


(f)

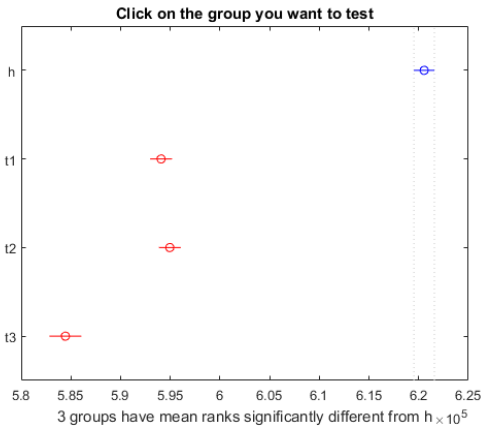
Şekil 5.35: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 5, \alpha = 0.05$).



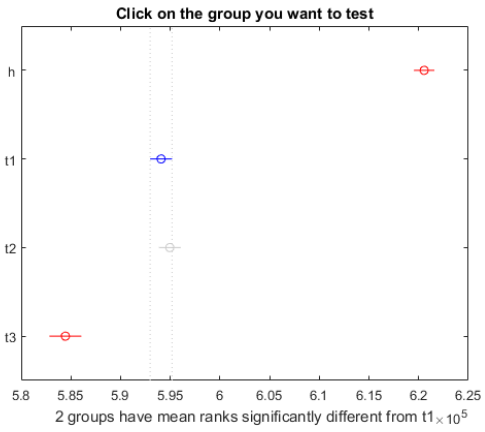
(a)



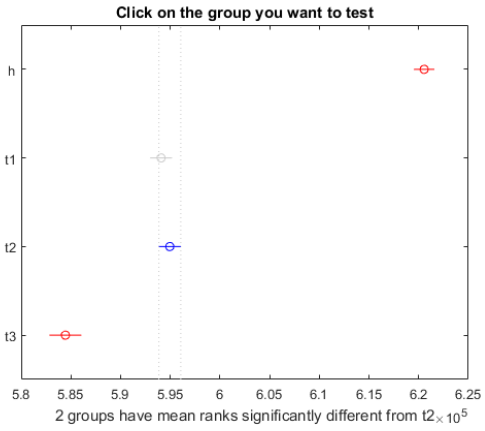
(b)



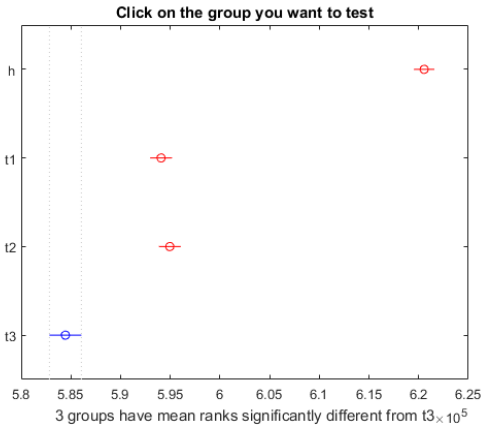
(c)



(d)

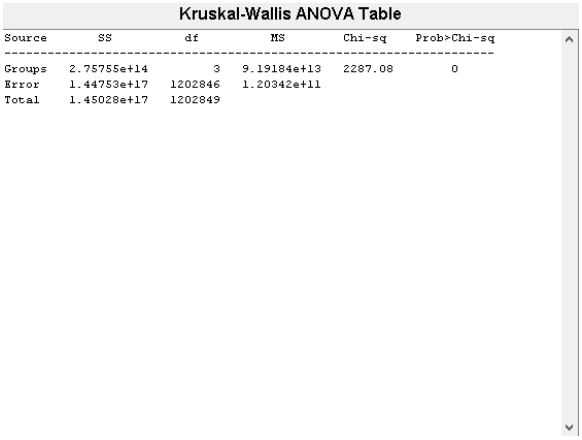


(e)



(f)

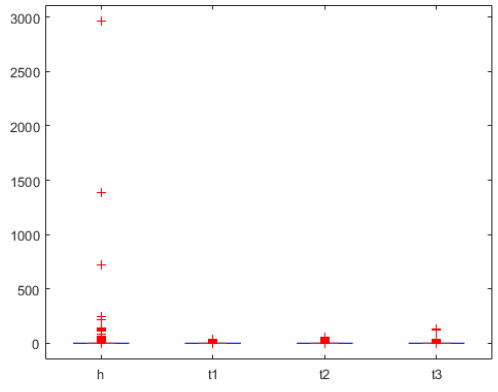
Şekil 5.36: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 9, \alpha = 0.05$).



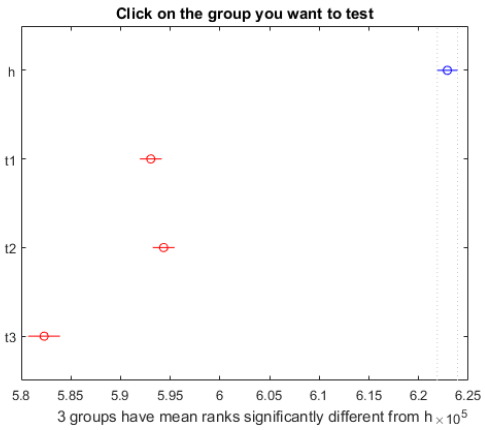
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	2.75755e+14	3	9.19184e+13	2287.08	0
Error	1.44753e+17	1202846	1.20342e+11		
Total	1.45028e+17	1202849			

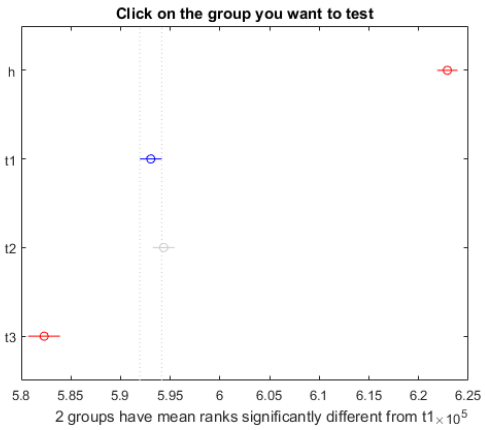
(a)



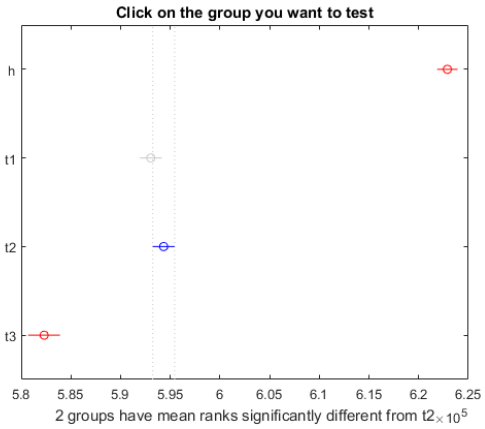
(b)



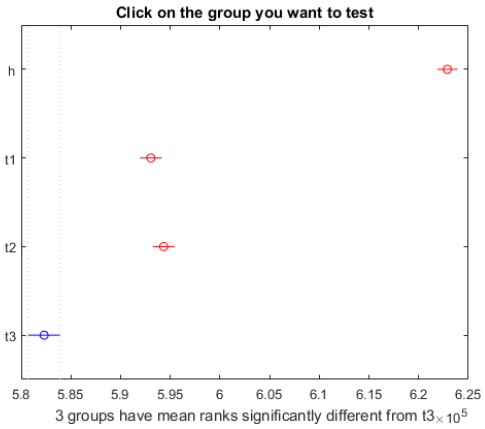
(c)



(d)

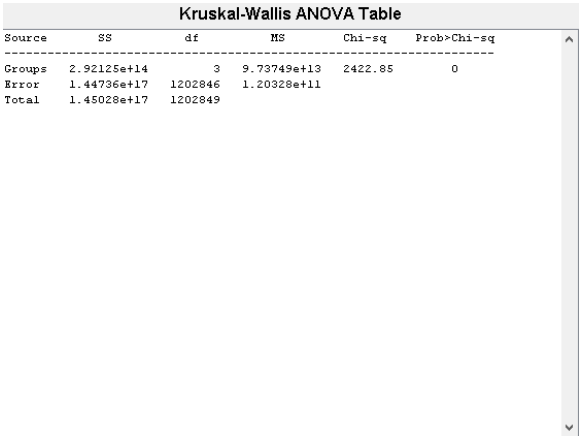


(e)



(f)

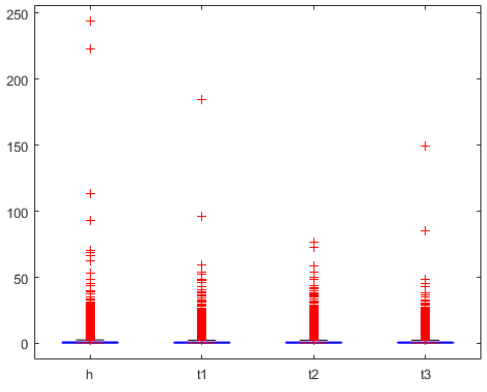
Şekil 5.37: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 15, \alpha = 0.05$).



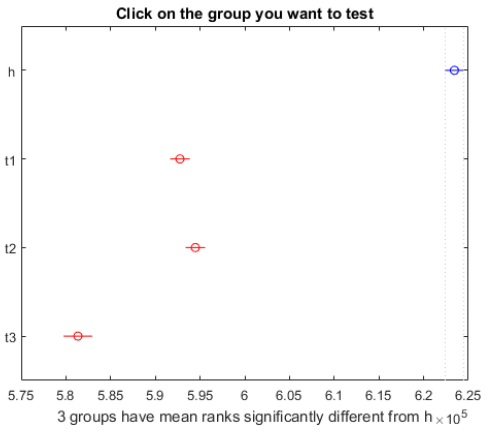
Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	2.92125e+14	3	9.73749e+13	2422.85	0
Error	1.44736e+17	1202846	1.20328e+11		
Total	1.45028e+17	1202849			

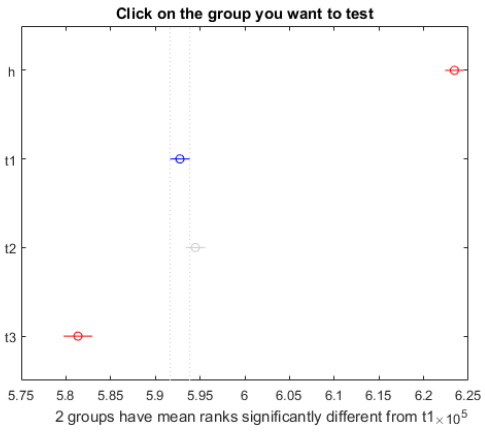
(a)



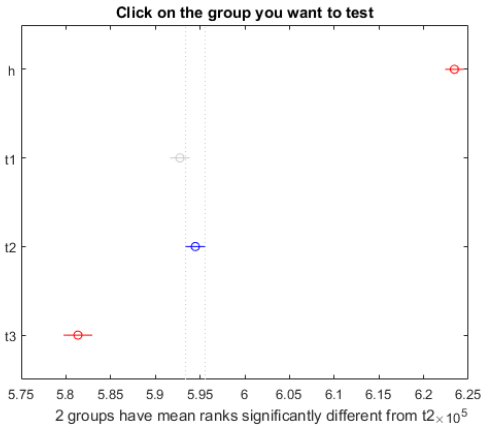
(b)



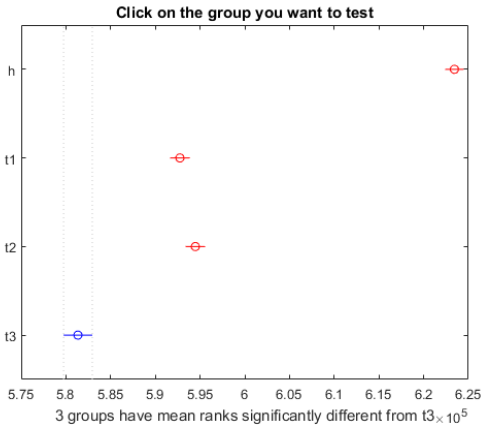
(c)



(d)



(e)



(f)

Şekil 5.38: Pankreas kanseri veri kümesi için Kruskal – Wallis H testi sonuçları ($L_F = 25, \alpha = 0.05$).

5.3.SINIFLANDIRMA

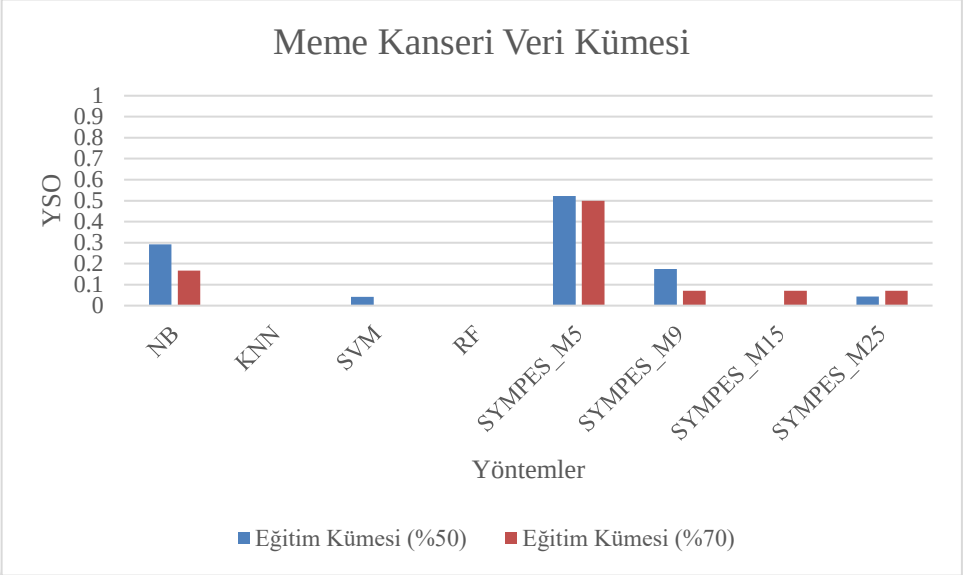
SYMPES yönteminin 5, 9, 15 ve 25 çerçeve boyutlu kütüphaneleri ile (sırasıyla SYMPES_M5, SYMPES_M9, SYMPES_M15 ve SYMPES_M25) sınıflandırma işleminin başarım sonuçları bu bölümde verilecektir. SYMPES yöntemi dışındaki karşılaştırmaya konu olan diğer sınıflandırma yöntemleri (NB, KNN, SVM, RF) R programlama ile gerçekleştirilmiştir. Bu yöntemlerin düşük (%50'den küçük) ve yüksek (%50'den büyük) yüzdelikli eğitim kümelerini kullanarak test kümelerini doğru sınıflamadaki başarımları gözlemlenmiştir. Başarım ölçütü olarak yanlış sınıflandırma oranı (YSO) kullanılmıştır.

5.3.1. Meme Kanseri Veri Kümesinin Sonuçları

Meme kanseri veri kümesinde sınıflardan birinin 2 örneklem içermesinden dolayı diğer veri kümelerinden farklı olarak %50 ve %70 oranlarında örneklem içeren eğitim kümeleri oluşturulmuş ve sınıflandırma işlemi yapılmıştır. Meme kanseri veri kümesinin sınıflandırılmasında yöntemlerin başarımları Tablo 5. 26'da ve Şekil 5.39'da sunulmuştur. SYMPES_M15 ve SYMPES_M25 yöntemlerinin diğer yöntemlerle kıyaslanabilir düzeyde başarım elde ettiği görülmüştür.

Tablo 5.26: Meme kanseri veri kümesi için sınıflandırma sonuçları.

Yöntemler	**Yanlış Sınıflandırma Oranı**	
	Eğitim Kümesi (%50)	**Eğitim Kümesi (%70)**
NB	0.292	0.167
KNN	**0.000**	**0.000**
SVM	0.042	**0.000**
RF	**0.000**	**0.000**
SYMPES_M5	0.522	0.500
SYMPES_M9	0.174	**0.071**
SYMPES_M15	**0.000**	**0.071**
SYMPES_M25	**0.044**	**0.071**



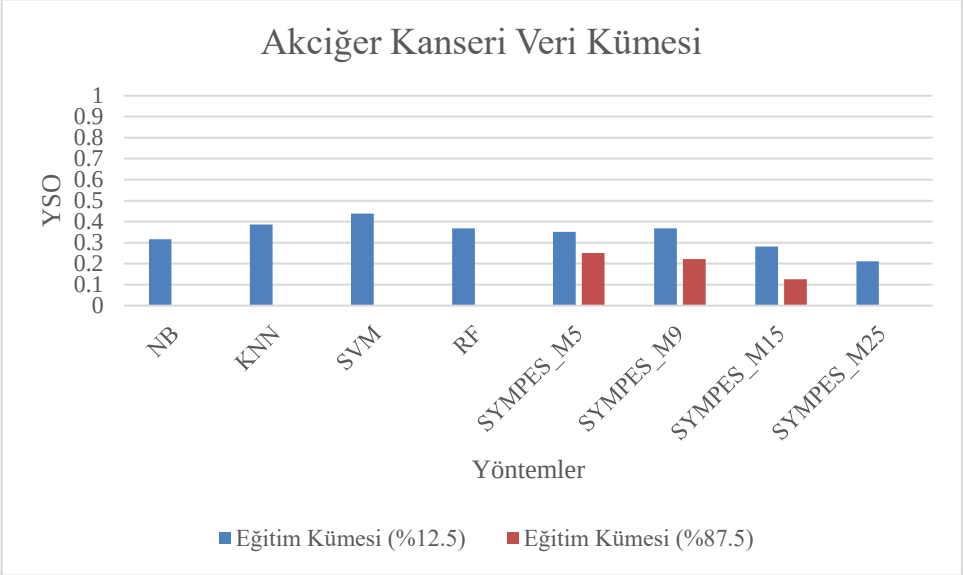
Şekil 5.39: Meme kanseri veri kümesi için sınıflandırma sonuçları.

5.3.2. Akciğer Kanseri Veri Kümesinin Sonuçları

Akciğer kanseri veri kümesi için yapılan sınıflandırma çalışmalarında düşük yüzdelikli (%12.5) örneklem içeren eğitim kümesini kullanan yöntemlerin her birinin yanlış sınıflandırma işlemi yaptığı görülmüştür. Eğitim kümesinin yüzdece test kümesinden daha az örneklem içerdiği durumlarda yöntemlerin başarımının oldukça düştüğü bilinmektedir. SYMPES yönteminin çerçeve boyutunun 25 olarak yapılandırıldığı durumda sınıflandırmada en başarılı yöntem olduğu görülmüştür (En az yanlış sınıflandırma oranına sahip olmuştur). Tablo 5.27'de ve Şekil 5.40'ta yöntemlerin sınıflandırma başarımları görülmektedir.

Tablo 5.27: Akciğer kanseri veri kümesi için sınıflandırma sonuçları.

Yöntemler	**Yanlış Sınıflandırma Oranı**	
	Eğitim Kümesi (%12.5)	**Eğitim Kümesi (%87.5)**
NB	0.316	**0.000**
KNN	0.386	**0.000**
SVM	0.439	**0.000**
RF	0.368	**0.000**
SYMPES_M5	0.351	0.250
SYMPES_M9	0.368	0.222
SYMPES_M15	0.281	0.125
SYMPES_M25	**0.211**	**0.000**



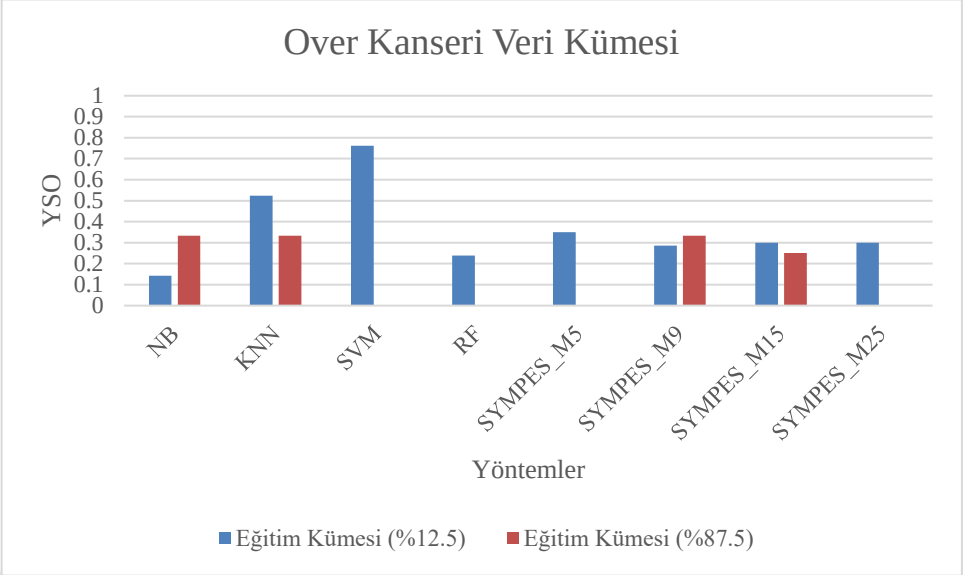
Şekil 5.40: Akciğer kanseri veri kümesi için sınıflandırma sonuçları.

5.3.3. Over Kanseri Veri Kümesinin Sonuçları

SYMPES yönteminin çerçeve boyutunun 9 olarak yapılandırıldığı ve düşük yüzdelikli örneklem içeren eğitim kümesinin kullanılması durumunda sınıflandırmada diğer yöntemlerle kıyaslanabilir düzeyde başarılı bir yöntem olduğu görülmüştür. Eğitim kümesinin örneklem sayısının arttırılması ile SYMPES yönteminin 5 ve 25'li çerçeve boyutlarında %100 başarıma sahip olduğu görülmüştür (SVM ve RF yöntemleri de %100 başarımla sınıflandırma yapmışlardır). Tablo 5.28'de ve Şekil 5.41'de yöntemlerin sınıflandırma başarımları görülmektedir.

Tablo 5.28: Over kanseri veri kümesi için sınıflandırma sonuçları.

Yöntemler	**Yanlış Sınıflandırma Oranı**	
	Eğitim Kümesi (%12.5)	**Eğitim Kümesi (%87.5)**
NB	**0.143**	0.333
KNN	0.524	0.333
SVM	0.762	**0.000**
RF	**0.238**	**0.000**
SYMPES_M5	0.350	**0.000**
SYMPES_M9	**0.286**	0.333
SYMPES_M15	0.300	0.250
SYMPES_M25	0.300	**0.000**



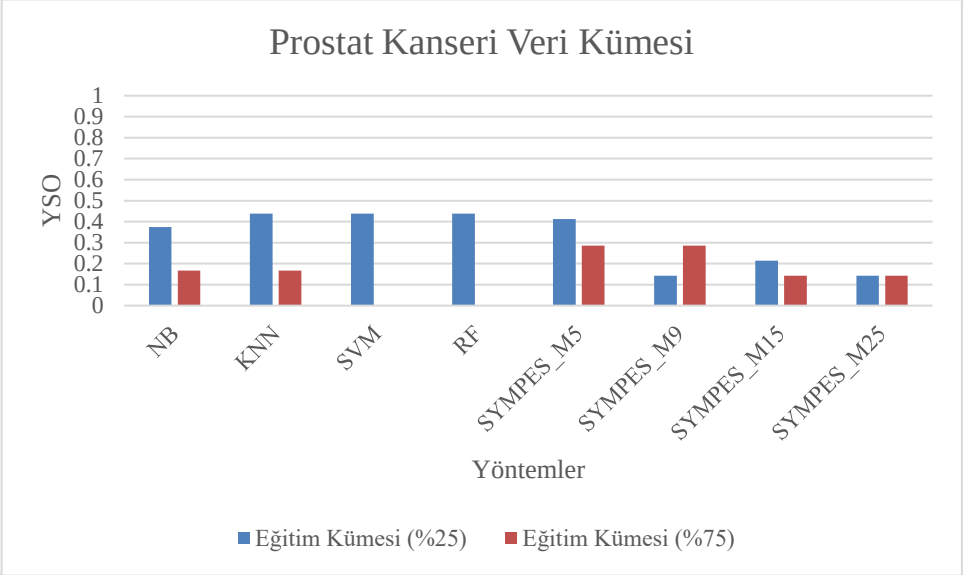
Şekil 5.41: Over kanseri veri kümesi için sınıflandırma sonuçları.

5.3.4. Prostat Kanseri Veri Kümesinin Sonuçları

SYMPES yönteminin çerçeve boyutunun 25 olarak yapılandırıldığı ve düşük yüzdelikli örneklem içeren eğitim kümesinin kullanılması durumunda sınıflandırmada diğer yöntemlere kıyasla daha başarılı bir yöntem olduğu görülmüştür. Prostat kanseri veri kümesinin bu şartlarda başarımı en yüksek yöntemi SYMPES_M25 olmuştur. Eğitim kümesinin örneklem sayısının arttırılması ile SYMPES yönteminin 25'li çerçeve boyutunda diğer yöntemlerle kıyaslanabilir düzeyde yaklaşık %86 başarıma sahip olduğu görülmüştür. Tablo 5.29'de ve Şekil 5.42'de yöntemlerin sınıflandırma başarımları görülmektedir.

Tablo 5.29: Prostat kanseri veri kümesi için sınıflandırma sonuçları.

Yöntemler	**Yanlış Sınıflandırma Oranı**	
	Eğitim Kümesi (%25)	**Eğitim Kümesi (%75)**
NB	0.375	0.167
KNN	0.438	0.167
SVM	0.438	**0.000**
RF	0.438	**0.000**
SYMPES_M5	0.412	0.286
SYMPES_M9	0.143	0.286
SYMPES_M15	0.214	0.143
SYMPES_M25	**0.143**	**0.143**



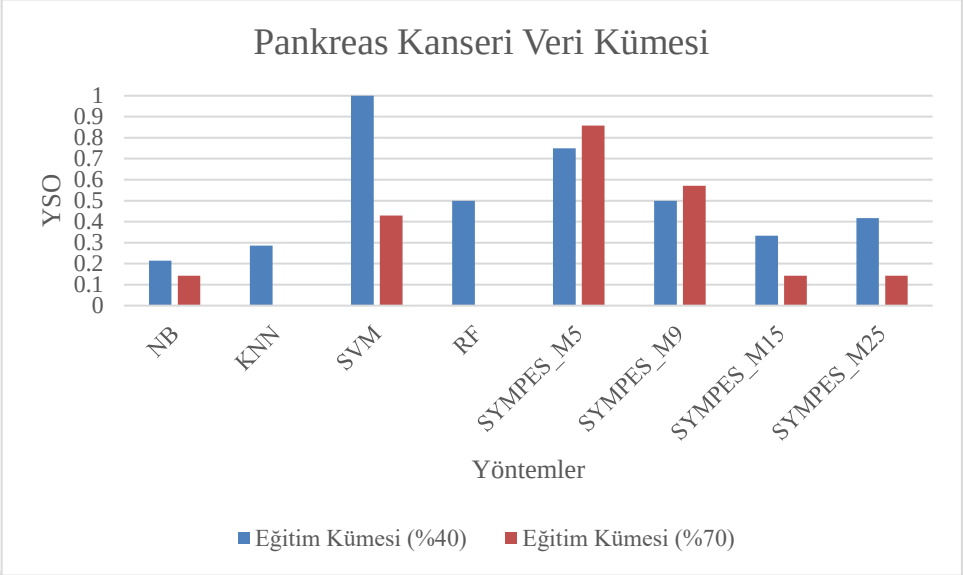
Şekil 5.42: Prostat kanseri veri kümesi için sınıflandırma sonuçları.

5.3.5. Pankreas Kanseri Veri Kümesinin Sonuçları

Pankreas kanseri veri kümesinde sınıflardan birinin 3 örneklem içermesinden dolayı diğer veri kümelerinden farklı olarak %40 ve %70 oranlarında örneklem içeren eğitim kümeleri oluşturulmuş ve sınıflandırma işlemi yapılmıştır. Pankreas kanseri veri kümesinin sınıflandırılmasında yöntemlerin başarımları Tablo 5. 30'da ve Şekil 5.43'te sunulmuştur. SYMPES_M15 ve SYMPES_M25 yöntemlerinin diğer yöntemlerle kıyaslanabilir düzeyde başarım elde ettiği görülmüştür.

Tablo 5.30: Pankreas kanseri veri kümesi için sınıflandırma sonuçları.

Yöntemler	**Yanlış Sınıflandırma Oranı**	
	Eğitim Kümesi (%40)	**Eğitim Kümesi (%70)**
NB	**0.214**	0.143
KNN	**0.286**	**0.000**
SVM	1.000	0.429
RF	0.500	**0.000**
SYMPES_M5	0.750	0.857
SYMPES_M9	0.500	0.571
SYMPES_M15	**0.333**	**0.143**
SYMPES_M25	0.417	0.143



Şekil 5.43: Pankreas kanseri veri kümesi için sınıflandırma sonuçları.

6. TARTIŞMA VE SONUÇ

Literatüre kazandırılmış önceki çalışmalar incelendiğinde SYMPES yönteminin biyomedikal alanındaki biyopotansiyel işaretler (EMG, EKG, EEG), konuşma ses işaretleri, tıbbi cihazlardan alınan görüntüler gibi tek ve iki boyutlu işaretlerin modellenmesinde kullanıldığı ve yüksek başarımlar sağladığı görülmüştür.

Mikrodizi teknolojisi, organizmaların genomik dizi bilgilerini kullanarak, sağlık ve hastalık durumlarının daha iyi anlaşılması, daha iyi teşhisler koyulması, yeni terapötik müdahaleler sunacak işlevsel biyolojik mekanizmalara dönüştürülmesi gibi görevleri yerine getirecek eşsiz niteliklere sahip olması sayesinde bilimsel araştırmaların önemli bir aracı haline gelmiştir. Mikrodiziler, giderek artan genomik bilgiyi etkili bir şekilde ölçeklendiren güvenilir, hızlı ve uygun maliyetli bir yöntem sağlarlar. Kanserlerin gelişiminde bir çok genin rol aldığı bilinmektedir. Mikrodizi teknolojileri ile aynı anda ve tek seferde tüm genomun analizinin gerçekleştirilmesinin mümkün olmasıyla, sağlıklı ve hasta bireylerin gen ifade profilleri çıkarılarak hastalık durumunda etkilenen genlerin belirlenmesinde önemli bir adım atılmıştır. Bu tezde, NCBI GEO veri tabanından ücretsiz olarak edinilmiş 5 kanser türüne ait (meme, akciğer, over, prostat ve pankreas) mikrodizi gen ifade veri kümeleri kullanılmıştır.

Tez çalışmaları kapsamında, mikrodizi gen ifade verileri üzerinde çalışılan iki önemli konu ele alınmıştır. SYMPES yöntemi ile oluşturulan her bir grup (sınıf veya kategori de denilebilir) için önceden tanımlı enerji ve bilgi blokları kullanılarak evrensel ortalama ve ortanca bilgi kütüphaneleri üretilmiştir. Bu kütüphanelerin çerçevelerinde en büyük bilgiyi taşıyan indis değerine karşılık gelen genlerin önemli genleri gösterdiği varsayımından yola çıkılarak farksal olarak ifade edilen genler tespit edilmeye çalışılmıştır. Farksal olarak ifade edilen genlerin analizinde SYMPES'in en büyük bilgi taşıyan genlerinden oluşturulan gen listesi ile mikrodizi platformunun limma yöntemiyle bulunmuş önemli gen listesinin kesişimini oluşturan gen sayısının yüzdelik değerleri incelendiğinde, genlerin yaklaşık yarısını tespit edebildiğini göstermiştir. Bunun üzerine SYMPES yöntemi ile birlikte Welch t-testi, Wilcoxon Sıra Sayıları – Toplam testi, ve bu testlerin FWER düzeltmesi yapan biçimleri, kat değişimi oranı (minimum ve maksimum oranlar) ve Pearson özilinti katsayısı kullanılarak 6 yeni yöntem geliştirilmiştir.

Meme kanseri veri kümesinin dört sınıflı yapısından dolayı oluşturulan üç karşılaştırma grubu üzerinde yapılan çalışmalardan elde edilen bulgular doğrultusunda, H~T1 karşılaştırma grubu

için, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin Pearson özilinti katsayısı ve kat değişimi parametreleri ile birlikte kullanılması ($SPCCF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphanelerinin kullanılması ile başarım oranı %89.02 olarak elde edilirken; evrensel ortanca bilgi kütüphanelerinin kullanılması ile başarım oranı %85.34 olarak elde edilmiştir. H~T2 karşılaştırma grubu için, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin kat değişimi parametresi ile birlikte kullanılması ($SF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphanelerinin kullanılması ile başarım oranı %89.80 olarak elde edilirken; evrensel ortanca bilgi kütüphanelerinin kullanılması ile başarım oranı %88.52 olarak elde edilmiştir. Benzer şekilde, H~T3 karşılaştırma grubu için, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin kat değişimi parametresi ile birlikte kullanılması ($SF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphanelerinin kullanılması ile başarım oranı %97.67 olarak elde edilirken; evrensel ortanca bilgi kütüphanelerinin kullanılması ile başarım oranı %96.84 olarak elde edilmiştir.

Akciğer kanseri veri kümesinin iki sınıflı yapısından dolayı oluşturulan tek karşılaştırma grubu (H~T1) üzerinde yapılan çalışmalardan elde edilen bulgular doğrultusunda, çerçeve boyutunun 15 olarak yapılandırılması ve SYMPES yönteminin Pearson özilinti katsayısı ve kat değişimi parametreleri ile birlikte kullanılması ($SPCCF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphanelerinin kullanılması ile başarım oranı %63.38 olarak elde edilirken; evrensel ortanca bilgi kütüphanelerinin kullanılması ile başarım oranı %65.64 olarak elde edilmiştir.

Over kanseri veri kümesinin iki sınıflı yapısından dolayı oluşturulan tek karşılaştırma grubu (H~T1) üzerinde yapılan çalışmalardan elde edilen bulgular doğrultusunda, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin Pearson özilinti katsayısı ve kat değişimi parametreleri ile birlikte kullanılması ($SPCCF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphaneleri ve evrensel ortanca bilgi kütüphanelerinin her ikisinin de kullanılması ile başarım oranı %100 olarak elde edilmiştir.

Prostat kanseri veri kümesinin iki sınıflı yapısından dolayı oluşturulan tek karşılaştırma grubu (H~T1) üzerinde yapılan çalışmalardan elde edilen bulgular doğrultusunda, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin Pearson özilinti katsayısı ve kat değişimi

parametreleri ile birlikte kullanılması ($SPCCF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphanelerinin kullanılması ile başarım oranı %85.10 olarak elde edilirken; evrensel ortanca bilgi kütüphanelerinin kullanılması ile başarım oranı %84.15 olarak elde edilmiştir.

Pankreas kanseri veri kümesinin dört sınıflı yapısından dolayı oluşturulan üç karşılaştırma grubu üzerinde yapılan çalışmalardan elde edilen bulgular doğrultusunda, H~T1 karşılaştırma grubu için, çerçeve boyutunun 15 ya da 25 olarak yapılandırılması ve SYMPES yönteminin Pearson özilinti katsayısı ve kat değişimi parametreleri ile birlikte kullanılması ($SPCCF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphaneleri ve evrensel ortanca bilgi kütüphanelerinin her ikisinin de kullanılması ile başarım oranı %100 olarak elde edilmiştir. H~T2 karşılaştırma grubu için, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin kat değişimi parametresi ile birlikte kullanılması ($SF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür. Evrensel ortalama bilgi kütüphaneleri ve evrensel ortanca bilgi kütüphanelerinin her ikisinin de kullanılması ile başarım oranı %100 olarak elde edilmiştir. Benzer şekilde, H~T3 karşılaştırma grubu için, evrensel ortalama bilgi kütüphanelerinin kullanılması durumunda, çerçeve boyutunun 25 olarak yapılandırılması ve SYMPES yönteminin kat değişimi parametresi ile birlikte kullanılması ($SF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür (%98.17). Aynı karşılaştırma grubu için, evrensel ortanca bilgi kütüphanelerinin kullanılması durumunda, çerçeve boyutunun 15 olarak yapılandırılması ve SYMPES yönteminin Pearson özilinti katsayısı ve kat değişimi parametreleri ile birlikte kullanılması ($SPCCF_{\min-\max}$) koşuluyla en yüksek başarımın elde edildiği görülmüştür (%100).

SYMPES yönteminin bu tez kapsamındaki araştırma konularından bir diğeri olan sınıflandırma işlemlerinde kullanılabilir olmasını sağlamak için optimum çerçeve boyutunun ve gerekli diğer işlemlerin veya parametrelerin belirlenmesi amaçlanmıştır.Genel olarak şimdiye kadar yapılan sınıflandırma çalışmalarında oldukça başarılı sonuçlar veren KNN, RF, SVM ve NB yöntemlerinin tez çalışmasında incelenen 5 veri kümesi üzerinde belirli yüzdelik oranda örneklem içeren eğitim kümelerini kullanarak sınıflandırma yaptığı bu çalışmamızda, SYMPES yöntemi ile oluşturulan önceden tanımlı enerji ve bilgi bloklarının 5, 9, 15 ve 25 çerçeve boyutları arasında çerçeve boyutlarının arttıkça başarım oranının yükseldiği, veri kümelerinden bazılarında SVM, RF gibi önemli algoritmalardan daha iyi sonuçlar verdiği gözlemlenmiştir.

Meme kanseri veri kümesinde, eğitim kümesinin %50 oranında örneklem içermesi durumunda, KNN ve RF yöntemleri gibi çerçeve boyutu 15 olarak yapılandırılan SYMPES yönteminin de %100 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. Eğitim kümesinin %70 oranında örneklem içermesi durumunda ise, KNN, SVM ve RF yöntemleri %100 doğrulukla sınıflandırma yapabilirken; çerçeve boyutu 9, 15 ya da 25 olarak yapılandırılan SYMPES yönteminin %93 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. NB yönteminin %83 doğrulukla sınıflandırma yaptığı da bulgular arasındadır.

Akciğer kanseri veri kümesinde, eğitim kümesinin %12.5 gibi çok küçük bir oranda örneklem içermesi durumunda, çerçeve boyutu 25 olarak yapılandırılan SYMPES yönteminin %100 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. Diğer tüm yöntemlerin %68'in altında doğrulukla sınıflandırma yaptıkları gözlemlenmiştir (NB: %68, KNN: %61, SVM: %56, RF: %63). Eğitim kümesinin %87.5 oranında örneklem içermesi durumunda ise, NB, KNN, SVM ve RF yöntemleri %100 doğrulukla sınıflandırma yapabilirken; aynı başarımı çerçeve boyutu 25 olarak yapılandırılan SYMPES yönteminin de elde ettiği gözlemlenmiştir.

Over kanseri veri kümesinde, eğitim kümesinin %12.5 gibi çok küçük bir oranda örneklem içermesi durumunda, çerçeve boyutu 9 olarak yapılandırılan SYMPES yönteminin %71 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. Diğer yöntemlerin doğru sınıflandırma oranları, NB için %86, KNN için %48, SVM için %24 ve RF için %76 olarak elde edilmiştir. SYMPES yönteminin KNN ve SVM yöntemlerine göre daha yüksek başarım elde ettiği açıkça görülmektedir. Eğitim kümesinin %87.5 oranında örneklem içermesi durumunda ise, SVM ve RF yöntemleri %100 doğrulukla sınıflandırma yapabilirken; aynı başarımı çerçeve boyutu 5 ya da 25 olarak yapılandırılan SYMPES yönteminin de elde ettiği gözlemlenmiştir. NB ve KNN yöntemleri bu eğitim kümeleri için, %67 doğrulukla sınıflandırma yapabilmiştir.

Prostat kanseri veri kümesinde, eğitim kümesinin %25 oranında örneklem içermesi durumunda, çerçeve boyutu 9 ya da 25 olarak yapılandırılan SYMPES yönteminin %86 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. Diğer yöntemlerin doğru sınıflandırma oranları, NB için %62, KNN için %56, SVM için %56 ve RF için %56 olarak elde edilmiştir. SYMPES yönteminin NB, KNN, SVM ve RF yöntemlerine göre daha yüksek başarım elde ettiği açıkça görülmektedir. Eğitim kümesinin %75 oranında örneklem içermesi durumunda ise, SVM ve RF yöntemleri %100 doğrulukla sınıflandırma yapabilirken; çerçeve boyutu 15 ya da 25 olarak yapılandırılan SYMPES yönteminin %86 doğrulukla sınıflandırma yapabildiği

gözlemlenmiştir. NB ve KNN yöntemlerinin %83 doğrulukla sınıflandırma yaptığı da bulgular arasındadır.

Panreas kanseri veri kümesinde, eğitim kümesinin %40 oranında örneklem içermesi durumunda, çerçeve boyutu 15 olarak yapılandırılan SYMPES yönteminin %67 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. Diğer yöntemlerin doğru sınıflandırma oranları, NB için %79, KNN için %71, SVM için %0 ve RF için %50 olarak elde edilmiştir. SYMPES yönteminin SVM ve RF yöntemlerine göre daha yüksek başarım elde ettiği açıkça görülmektedir. Eğitim kümesinin %70 oranında örneklem içermesi durumunda ise, KNN ve RF yöntemleri %100 doğrulukla sınıflandırma yapabilirken; çerçeve boyutu 15 ya da 25 olarak yapılandırılan SYMPES yöntemi ile NB yöntemlerinin %86 doğrulukla sınıflandırma yapabildiği gözlemlenmiştir. SVM yönteminin bu eğitim kümesi için, %57 doğrulukla sınıflandırma yaptığı da bulgular arasındadır.

SYMPES yöntemi için çerçeve boyutunun 25 olarak yapılandırıldığı durumlarda, tüm veri kümelerinde optimum sonuç elde edildiği söylenebilir. Sonuç olarak, hiç bir yöntem tüm veri kümelerinde en iyi sonucu veremez. Bu duruma elbette ki, bilgi kaynağı olan veri kümelerinin doğası sebep olmaktadır. Bir veri kümesinde en iyi sonucu veren yöntem başka bir veri kümesinde başarımı en düşük yöntem olabilmektedir. Yukarıda belirtilen sonuçlardan da bu durum anlaşılmaktadır.

Sonuç olarak, tez kapsamında, SYMPES yönteminin kanser türlerine ilişkin gen ifade verileri üzerinde kanserlerin gelişiminde etkili olduğu bilinen çok sayıda farksal olarak ifade edilen genlerin tespitinin yapılmasında kullanılabileceği; kat değişimi ve buna ek olarak Pearson özilinti katsayısı gibi ek parametrelerin de kullanılması durumunda SYMPES yönteminin başarımının artırıldığı gözlemlenmiştir. Bunun dışında, SYMPES yönteminin sınıflandırma çalışmalarında da umut vaat eden sonuçlar ortaya koyduğu gözlemlenmiştir. Tüm örneklemin yüzdece yarısından az sayıda örneklem içeren (%12,5, %25, %40, gibi) eğitim kümesini kullanarak, kanser türlerini sınıflandırmada yüksek başarım elde etmesi dikkat çekici bulunmuştur. Bunun yanında, eğitim kümesindeki örneklem sayısının arttırılması beklenildiği gibi başarımın artmasını sağlamıştır. Bunların yanı sıra, karşılaştırılan diğer yöntemlerde de elde edilen yüksek yanlış sınıflandırma oranlarının, örneklemlerin sınıflara göre dengesiz dağılımından (*İng.* imbalanced class problem) kaynaklandığı düşünülmektedir.

KAYNAKLAR

[1]Nussbaum, R. L., Mcinnes, R. R., Williard, H. F., Thompson, J. S. & Thompson, M. W. 2001. *Genetics in medicine*, Saunders Philadelphia.

[2]Mutch, D. M., Berger, A., Mansourian, R., Rytz, A. & Roberts, M.-A. 2002. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC bioinformatics,* 3**,** 17.

[3]Dalman, M. R., Deeter, A., Nimishakavi, G. & Duan, Z.-H. 2012. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC bioinformatics,* 13**,** S11.

[4]Butte, A. J., Ye, J., Häring, H., Stumvoll, M., White, M. & Kohane, I. 2000. Determining significant fold differences in gene expression analysis. *Biocomputing 2001*. World Scientific.

[5]Pavlidis, P. 2003. Using ANOVA for gene selection from microarray studies of the nervous system. *Methods,* 31**,** 282-289.

[6]Zhao, Y. & Pan, W. 2003. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics,* 19**,** 1046-54.

[7]Chen, J., Wang, S., Tsai, C. & Lin, C. 2007. Selection of differentially expressed genes in microarray data analysis. *The pharmacogenomics journal,* 7**,** 212-220.

[8]Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D. & Altman, R. B. 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics,* 18**,** 1454-1461.

[9]Parrish, R. S. & Spencer Iii, H. J. 2004. Effect of normalization on significance testing for oligonucleotide microarrays. *Journal of biopharmaceutical statistics,* 14**,** 575-589.

[10]Smyth, G. K. 2005. Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer.

[11]Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research,* 43**,** e47-e47.

[12]Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. & Betel, D. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology,* 14**,** 3158.

[13]Robinson, M. D., Mccarthy, D. J. & Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-140.

[14]Statnikov, A., Wang, L. & Aliferis, C. F. 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics,* 9**,** 319.

[15]Mei, Z., Shen, Q. & Ye, B. 2009. Hybridized KNN and SVM for gene expression data classification. *Life Science Journal,* 6**,** 61-66.

[16]Turgut, S., Dağtekin, M. & Ensari, T. Microarray breast cancer data classification using machine learning methods. 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018. IEEE, 1-3.

[17]Aydadenta, H. On the classification techniques in data mining for microarray data classification. Journal of Physics: Conference Series, 2018. IOP Publishing, 012004.

[18]Wang, X. & Simon, R. 2011. Microarray-based cancer prediction using single genes. *BMC bioinformatics,* 12**,** 391.

[19]Yarman, B., Gurkan, H., Guz, U. & Aygun, B. A novel method to represent ECG signals via predefined personalized signature and envelope functions. 2001 Conference

Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2001. IEEE, 2150-2153.

[20]Gurkan, H., Guz, U. & Yarman, B. S. An efficient ECG data compression technique based on predefined signature and envelope vector banks. 2005 IEEE International Symposium on Circuits and Systems, 2005. IEEE, 1334-1337.

[21]Güz, Ü., Gürkan, H. & Yarman, B. S. 2006. A new method to represent speech signals via predefined signature and envelope sequences. *EURASIP Journal on Advances in Signal Processing,* 2007**,** 1-17.

[22]Yarman, B. S., Güz, Ü. & Gürkan, H. 2006. On the comparative results of "SYMPES: A new method of speech modeling". *AEU-International Journal of Electronics and Communications,* 60**,** 421-427.

[23]Guz, U., Gurkan, H. & Yarman, B. S. A new speech modeling method: SYMPES. 2006 IEEE International Symposium on Circuits and Systems, 2006. IEEE, 4 pp.

[24]Guz, U., Gurkan, H. & Yarman, B. S. A novel noise robust and low bit rate speech coding algorithm. 2009 24th International Symposium on Computer and Information Sciences, 2009. IEEE, 471-474.

[25]Yarman, B. S., Ulger, C. & Aslan, A. B. SYMPES technique encoded IP-based secure voice communication system. 2017 International Symposium on Signals, Circuits and Systems (ISSCS), 2017. IEEE, 1-3.

[26]Gürkan, H., Guz, U. & Yarman, B. S. 2009. EEG signal compression based on classified signature and envelope vector sets. *International Journal of Circuit Theory and Applications,* 37**,** 351-363.

[27]Gokbay, I. Z., Gezer, M., Guz, U., Gurkan, H. & Yarman, B. S. compression of the CT images using classified energy and pattern blocks. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013. IEEE, 6462-6465.

[28]Gökbay Zaim, İ. 2013. *SYMPES Yöntemiyle Biyomedikal Görüntülerin Modellenmesi.* PhD, İstanbul Üniversitesi.

[29]Watson, J. D. & Crick, F. H. The structure of DNA. Cold Spring Harbor symposia on quantitative biology, 1953. Cold Spring Harbor Laboratory Press, 123-131.

[30]*GlobalRPh: Medical Terminology – Word Building Reference* [Online]. Available: https://globalrph.com/medical-terminology/ [Accessed 01.05.2020].

[31]Dalma-Weiszhausz, D. D., Warrington, J., Tanimoto, E. Y. & Miyada, C. G. 2006. [1] The Affymetrix GeneChip® Platform: An Overview. *Methods in enzymology,* 410**,** 3-28.

[32]Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M. & Holko, M. 2012. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research,* 41**,** D991-D995.

[33]Edgar, R., Domrachev, M. & Lash, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research,* 30**,** 207-210.

[34]Barrett, T. & Edgar, R. 2006. [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis. *Methods in enzymology,* 411**,** 352-369.

[35]Richardson, A. L., Wang, Z. C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J. D., Livingston, D. M. & Ganesan, S. 2006. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell,* 9**,** 121-32.

[36]Alimonti, A., Carracedo, A., Clohessy, J. G., Trotman, L. C., Nardella, C., Egia, A., Salmena, L., Sampieri, K., Haveman, W. J., Brogi, E., Richardson, A. L., Zhang, J. &

Pandolfi, P. P. 2010. Subtle variations in Pten dose determine cancer susceptibility. *Nat Genet,* 42**,** 454-8.

[37]Sato, T., Kaneda, A., Tsuji, S., Isagawa, T., Yamamoto, S., Fujita, T., Yamanaka, R., Tanaka, Y., Nukiwa, T., Marquez, V. E., Ishikawa, Y., Ichinose, M. & Aburatani, H. 2013. PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. *Sci Rep,* 3**,** 1911.

[38]Bowen, N. J., Walker, L. D., Matyunina, L. V., Logani, S., Totten, K. A., Benigno, B. B. & Mcdonald, J. F. 2009. Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. *BMC medical genomics,* 2**,** 71.

[39]Arredouani, M. S., Lu, B., Bhasin, M., Eljanne, M., Yue, W., Mosquera, J.-M., Bubley, G. J., Li, V., Rubin, M. A. & Libermann, T. A. 2009. Identification of the transcription factor single-minded homologue 2 as a potential biomarker and immunotherapy target in prostate cancer. *Clinical cancer research,* 15**,** 5794-5802.

[40]Hiraoka, N., Yamazaki–Itoh, R., Ino, Y., Mizuguchi, Y., Yamada, T., Hirohashi, S. & Kanai, Y. 2011. CXCL17 and ICAM2 are associated with a potential anti-tumor immune response in early intraepithelial stages of human pancreatic carcinogenesis. *Gastroenterology,* 140**,** 310-321. e4.

[41]Davis, S. & Meltzer, P. S. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics,* 23**,** 1846-1847.

[42]Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics,* 17**,** 520-525.

[43]Celton, M., Malpertuy, A., Lelandais, G. & De Brevern, A. G. 2010. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC genomics,* 11**,** 15.

[44]Oh, S., Kang, D. D., Brock, G. N. & Tseng, G. C. 2011. Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Bioinformatics,* 27**,** 78-86.

[45]Chiu, C.-C., Chan, S.-Y., Wang, C.-C. & Wu, W.-S. 2013. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. *BMC systems biology,* 7**,** S12.

[46]Luengo, J., García, S. & Herrera, F. 2012. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems,* 32**,** 77-108.

[47]De Souto, M. C., Jaskowiak, P. A. & Costa, I. G. 2015. Impact of missing data imputation methods on gene expression clustering and classification. *BMC bioinformatics,* 16**,** 64.

[48]Tuikkala, J., Elo, L. L., Nevalainen, O. S. & Aittokallio, T. 2008. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics,* 9**,** 202.

[49]Brock, G. N., Shaffer, J. R., Blakesley, R. E., Lotz, M. J. & Tseng, G. C. 2008. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC bioinformatics,* 9**,** 12.

[50]Sarıkaş, A., Odabaşıoğlu, N. & Altay, G. Comparison of estimation methods for missing value imputation of gene expression data. 2016 Medical Technologies National Congress (TIPTEKNO), 2016. IEEE, 1-4.

[51]Oba, S., Sato, M.-A., Takemasa, I., Monden, M., Matsubara, K.-I. & Ishii, S. 2003. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics,* 19**,** 2088-2096.

[52]Lee, M.-L. T. 2007. *Analysis of microarray gene expression data*, Springer Science & Business Media.

[53]Bartlett, M. S. 1947. The use of transformations. *Biometrics,* 3**,** 39-52.

[54]Box, G. E. & Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological),* 26**,** 211-243.

[55]Al Shalabi, L., Shaaban, Z. & Kasasbeh, B. 2006. Data mining: A preprocessing engine. *Journal of Computer Science,* 2**,** 735-739.

[56]Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* 19**,** 185-193.

[57]Tusher, V. G., Tibshirani, R. & Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences,* 98**,** 5116-5121.

[58]Witten, D. & Tibshirani, R. 2007. A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis,* 1776**,** 58-85.

[59]Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. & Halfon, M. S. 2005. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome biology,* 6**,** R16.

[60]Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., Herman, D. & Goodsaid, F. M. 2006. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature biotechnology,* 24**,** 1162-1169.

[61]Welch, B. L. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika,* 29**,** 350-362.

[62]Ge, Y., Sealfon, S. C. & Speed, T. P. 2009. Multiple testing and its applications to microarrays. *Statistical methods in medical research,* 18**,** 543-563.

[63]Walpole, R. E., Myers, R.H., Myers, S.L., Ye, K. 2011. *Probability and statistics for engineers and scientists*, Pearson Education.

[64]Churchill, G. A. 2004. Using ANOVA to analyze microarray data. *Biotechniques,* 37**,** 173-177.

[65]Kerr, M. K., Martin, M. & Churchill, G. A. 2000. Analysis of variance for gene expression microarray data. *Journal of computational biology,* 7**,** 819-837.

[66]Grant, G., Manduchi, E. & Stoeckert, C. 2002. Using non-parametric methods in the context of multiple testing to determine differentially expressed genes. *Methods of microarray data analysis.* Springer.

[67]Cribbie, R. A. & Keselman, H. 2003. Pairwise multiple comparisons: A model comparison approach versus stepwise procedures. *British Journal of Mathematical and Statistical Psychology,* 56**,** 167-182.

[68]Doğan, N. & Doğan, İ. 2014. Birinci Tür Hata'nın Kontrolü ve Adımsal (Stepwise) Çoklu Karşılaştırma Testleri. *Düzce Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi,* 1**,** 28-33.

[69]Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics***,** 65-70.

[70]Holland, B. S. & Copenhaver, M. D. 1987. An improved sequentially rejective Bonferroni test procedure. *Biometrics***,** 417-423.

[71]Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association,* 62**,** 626-633.

[72]Cover, T. & Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory,* 13**,** 21-27.
[73]Vapnik, V. N. 1995. The nature of statistical learning. *Theory*.
[74]Breiman, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.
[75]Semmlow, J. L. & Griffel, B. 2014. *Biosignal and medical image processing,* CRC press.
[76]Akansu, A. N., Haddad, P. A., Haddad, R. A. & Haddad, P. R. 2001. *Multiresolution signal decomposition: transforms, subbands, and wavelets,* Academic press.

EKLER

EK-A: İSTATİSTİKSEL TESTLER

Test	Eşitlik	Terimlerin Açıklaması
SAM testi (*İng.* Significance analysis of microarrays)	$d(i) = \frac{\bar{x}_A(i) - \bar{x}_B(i)}{s(i) + s_0}$	$d(i)$: Bağıl fark $\bar{x}_A(i), \bar{x}_B(i)$: A ve B örneklemlerinin i. gen için ortalaması $s(i)$: Standart sapma s_0: varyasyon katsayısını minimize edecek küçük pozitif bir sabit.
M testi (*İng.* Mean differences)	$d(i) = \bar{x}_A(i) - \bar{x}_B(i)$	$d(i)$: Bağıl fark $\bar{x}_A(i), \bar{x}_B(i)$: A ve B örneklemlerinin i. gen için ortalaması M- testi: bağıl fark anlamında kullanılan kat değişimi parametresinin diğer adıdır. Bu testte, logaritma-2 tabanında ölçeklenmiş örneklemlerin ortalamalarının farkının bulunması esastır.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Ali SARIKAŞ
Doğum Yeri	ADANA
Doğum Tarihi	02.05.1988
Uyruğu	☑ T.C. ❒ Diğer:
Telefon	0216 777 40 70
E-Posta Adresi	ali.sarikas@marmara.edu.tr
Web Adresi	https://avesis.marmara.edu.tr/ali.sarikas



Eğitim Bilgileri	
Lisans	
Üniversite	Marmara Üniversitesi
Fakülte	Teknik Eğitim Fakültesi
Bölümü	Elektronik ve Bilgisayar Eğitimi (İng.)
Mezuniyet Yılı	2010

Lisans	
Üniversite	İstanbul Teknik Üniversitesi
Fakülte	Elektrik Elektronik Fakültesi
Bölümü	Elektronik ve Haberleşme Mühendisliği
Mezuniyet Yılı	2018

Yüksek Lisans	
Üniversite	İstanbul Teknik Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Elektronik ve Haberleşme Mühendisliği
Programı	Biyomedikal Mühendisliği
Mezuniyet Yılı	2014

Makale ve Bildiriler
https://avesis.marmara.edu.tr/ali.sarikas/yayinlar