T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

DOKTORA TEZİ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE BİYOMEDİKAL VERİDEN KANSER TEŞHİSİ

Öznur Sinem SÖNMEZ

DANIŞMAN
Dr. Öğr. Üyesi Mustafa DAĞTEKİN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

İSTANBUL-2022

Bu çalışma, 11.01.2022 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Dr. Öğr. Üyesi Mustafa DAĞTEKİN(Danışman)
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Ahmet SERTBAŞ
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Doç. Dr. Gökhan BİLGİN
Yıldız Teknik Üniversitesi
Elektrik Elektronik Fakültesi

Dr. Öğr. Üyesi Tolga ENSARİ
Arkansas Tech University
College of Engineering and Applied Sciences

Doç. Dr. Selçuk SEVGEN
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

Bu çalışma “Makine Öğrenmesi Yöntemleriyle Biyomedikal Veriden Kanser Teşhisi” adlı doktora tez çalışmasını içermektedir.

Tez çalışmam sırasında bilgi ve tecrübeleriyle yardımlarını esirgemeyen değerli tez danışmanım Sayın Dr. Öğr. Üyesi Mustafa DAĞTEKİN’e ve değerli hocam Sayın Dr. Öğr. Üyesi Tolga ENSARİ’ye katkılarından dolayı teşekkür ederim.

Tez izleme komitesi üyelerim Sayın Prof. Dr. Ahmet SERTBAŞ’a ve Sayın Doç. Dr. Gökhan BİLGİN’e değerli önerileri ve yönlendirmeleri için teşekkür ederim.

Beni her zaman destekleyen aileme en içten teşekkürlerimi sunarım.

Ocak 2022 Öznur Sinem SÖNMEZ

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **viii**
TABLO LİSTESİ ..... **x**
SİMGE VE KISALTMA LİSTESİ ..... **xvi**
ÖZET ..... **xix**
SUMMARY ..... **xxi**
1. GİRİŞ **1**
2. GENEL KISIMLAR **4**
2.1. KANSER 4
2.2. GEN İFADESİ 5
2.3. MİKRODİZİ VE RNA DİZİLEME TEKNOLOJİLERİ 6
2.3. GEN İFADE SEVİYELERİNİ İÇEREN VERİ KÜMELERİNİN SINIFLANDIRILMASIYLA İLGİLİ LİTERATÜR ÇALIŞMALARI 9
3. MALZEME VE YÖNTEM **16**
3.1. ÖZNİTELİK SEÇİMİ YÖNTEMLERİ 16
3.1.1. Filtre Yöntemler 16
3.1.1.1. Pearson Korelasyonu 17
3.1.1.2. Karşılıklı Bilgi 17
3.1.1.3. Minimum Tekrar Maksimum İlgi 18
3.1.1.4. Bilgi Kazancı 20
3.1.1.5. Relief-F 21
3.1.1.6. Diğer Filtre Yöntemler 21
3.1.2. Sarmal Yöntemler 22
3.1.3. Gömülü Yöntemler 23
3.1.4. Hibrit Yöntemler 23
3.2. BOYUT İNDİRGEME YÖNTEMLERİ 24
3.2.1. Temel Bileşen Analizi 24
3.2.2. Lineer Diskriminant Analizi 24
3.3. SINIFLANDIRMA YÖNTEMLERİ 25

3.3.1. Destek vektör makinesi 25
3.3.2. K-en yakın komşu....26
3.3.3. Naive Bayes....26
3.3.4. İleri Beslemeli Yapay Sinir Ağları (İBYSA) 27
3.4. OPTİMİZASYON YÖNTEMLERİ....29
3.4.1. Genetik Algoritma....29
3.4.2. Parçacık Sürü Optimizasyonu 30
3.5. ÖNERİLEN HİBRİT YÖNTEMLER....31
3.6. VERİ KÜMELERİ....35
3.7. DEĞERLENDİRME ÖLÇÜTLERİ....38
3.7.1. Doğruluk Oranı....38
3.7.2. Kesinlik....38
3.7.3. Duyarlılık....38
3.7.4. Özgüllük 39
3.7.5. F1-Ölçütü....39
3.7.6. Negatif Kestirim Değeri 39
4. BULGULAR....40
4.1. ÖZNİTELİK SEÇİMİ UYGULANMADAN SINIFLANDIRMA SONUÇLARI....40
4.1.1. Lösemi Veri Kümesi için Sınıflandırma Sonuçları 41
4.1.2. Kolon Kanseri Veri Kümesi için Sınıflandırma Sonuçları....43
4.1.3. Prostat Kanseri Veri Kümesi için Sınıflandırma Sonuçları....46
4.1.4. Over Kanseri Veri Kümesi için Sınıflandırma Sonuçları....48
4.1.5. DLBCL Veri Kümesi için Sınıflandırma Sonuçları 50
4.1.6. SMK Veri Kümesi için Sınıflandırma Sonuçları 52
4.1.7. Meme Kanseri Veri Kümesi için Sınıflandırma Sonuçları....54
4.1.8. Meme Kanseri 2 Veri Kümesi için Sınıflandırma Sonuçları....56
4.1.9. CNS Veri Kümesi için Sınıflandırma Sonuçları 58
4.2. FİLTRE YÖNTEMLERLE ÖZNİTELİK SEÇİMİNİN SINIFLANDIRMA SONUÇLARINA ETKİSİNİN İNCELENMESİ....60
4.2.1. Kolon Kanseri Veri Kümesi Sonuçları....61
4.2.2. Prostat Kanseri Veri Kümesi Sonuçları....63
4.2.3. DLBCL Veri Kümesi Sonuçları 64
4.2.4. SMK Veri Kümesi Sonuçları....66
4.2.5. CNS Veri Kümesi Sonuçları 68
4.2.6. Meme Kanseri Veri Kümesi Sonuçları....69

4.3. TBA VE LDA YÖNTEMLERİYLE BOYUT İNDİRGEME SONRASI SINIFLANDIRMA SONUÇLARI....71
4.4. SARMAL YÖNTEMLERLE SINIFLANDIRMA SONUÇLARI....78
4.4.1. GA İçeren Sarmal Yöntemlerin Sınıflandırma Sonuçları....79
4.4.2. PSO İçeren Sarmal Yöntemlerin Sınıflandırma Sonuçları....83
4.5. ÖNERİLEN HİBRİT YÖNTEMLERİN SINIFLANDIRMA SONUÇLARI....88
5. TARTIŞMA VE SONUÇ....106
KAYNAKLAR....112
EKLER....122
EK.1. FİLTRE YÖNTEMLERLE ÖZNİTELİK SEÇİMİ SONRASI SINIFLANDIRMA SONUÇLARI....122
ÖZGEÇMİŞ....138

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1: DNA'dan proteine bilgi aktarımı aşamaları (Barton ve diğ., 2007).....5

Şekil 2.2: Mikrodizi hazırlanması ve deneyi (Somasundaram ve diğ., 2002). 7

Şekil 2.3: RNA dizileme deneyi temel aşamaları. 8

Şekil 3.1: Çok katmanlı ileri beslemeli YSA yapısı. 28

Şekil 3.2: GA tabanlı önerilen yöntemlerin akış şeması. 32

Şekil 3.3: PSO tabanlı önerilen yöntemlerin akış şeması. 34

Şekil 4.1: Lösemi veri kümesi için sınıflandırma sonuçları. 43

Şekil 4.2: Kolon kanseri veri kümesi için sınıflandırma sonuçları. 45

Şekil 4.3: Prostat kanseri veri kümesi için sınıflandırma sonuçları. 47

Şekil 4.4: Over kanseri veri kümesi için sınıflandırma sonuçları. 49

Şekil 4.5: DLBCL kanseri veri kümesi için sınıflandırma sonuçları. 51

Şekil 4.6: SMK veri kümesi için sınıflandırma sonuçları. 54

Şekil 4.7: Meme kanseri veri kümesi için sınıflandırma sonuçları. 56

Şekil 4.8: Meme kanseri 2 veri kümesi için sınıflandırma sonuçları. 58

Şekil 4.9: CNS kanseri veri kümesi için sınıflandırma sonuçları. 60

Şekil 4.10: Filtre yöntemlerin kolon kanseri veri kümesinde KNN yöntemi için gen sayısı-doğruluk oranı grafiği. 62

Şekil 4.11: Filtre yöntemlerin prostat kanseri veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği. 64

Şekil 4.12: Filtre yöntemlerin DLBCL veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği. 66

Şekil 4.13: Filtre yöntemlerin SMK veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği. 66

Şekil 4.14: Filtre yöntemlerin CNS veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği. 69

Şekil 4.15: Filtre yöntemlerin meme kanseri veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği. 71

TABLO LİSTESİ

Sayfa No

Tablo 2.1: Kanserin histolojik olarak ana grupları. 4

Tablo 2.2: Öznitelik seçiminde gömülü yöntemler. 13

Tablo 3.1: Önerilen yöntemler için GA parametreleri.....33

Tablo 3.2: Önerilen yöntemler için PSO parametreleri. 35

Tablo 3.3: Veri kümeleri. 36

Tablo 3.4: Sınıflandırmanın karışıklık matrisi.....38

Tablo 4.1: Lösemi veri kümesi için sınıflandırma sonuçları (hold-out).....41

Tablo 4.2: Lösemi veri kümesi için sınıflandırma sonuçları (CV).....42

Tablo 4.3: Lösemi veri kümesi için sınıflandırma sonuçları (LOOCV).....42

Tablo 4.4: Kolon kanseri veri kümesi için sınıflandırma sonuçları (hold-out). 44

Tablo 4.5: Kolon kanseri veri kümesi için sınıflandırma sonuçları (CV). 44

Tablo 4.6: Kolon kanseri veri kümesi için sınıflandırma sonuçları (LOOCV). 45

Tablo 4.7: Prostat kanseri veri kümesi için sınıflandırma sonuçları (hold-out). 46

Tablo 4.8: Prostat kanseri veri kümesi için sınıflandırma sonuçları (CV). 46

Tablo 4.9: Prostat kanseri veri kümesi için sınıflandırma sonuçları (LOOCV). 47

Tablo 4.10: Over kanseri veri kümesi için sınıflandırma sonuçları (hold-out). 48

Tablo 4.11: Over kanseri veri kümesi için sınıflandırma sonuçları (CV). 48

Tablo 4.12: Over kanseri veri kümesi için sınıflandırma sonuçları (LOOCV). 49

Tablo 4.13: DLBCL veri kümesi için sınıflandırma sonuçları (hold-out).....50

Tablo 4.14: DLBCL veri kümesi için sınıflandırma sonuçları (CV).....50

Tablo 4.15: DLBCL veri kümesi için sınıflandırma sonuçları (LOOCV). 51

Tablo 4.16: SMK veri kümesi için sınıflandırma sonuçları (hold-out). 52

Tablo 4.17: SMK veri kümesi için sınıflandırma sonuçları (CV). 53

Tablo 4.18: SMK veri kümesi için sınıflandırma sonuçları (LOOCV). 53

Tablo 4.19: Meme kanseri veri kümesi için sınıflandırma sonuçları (hold-out). 54

Tablo 4.20: Meme kanseri veri kümesi için sınıflandırma sonuçları (CV). 55

Tablo 4.21: Meme kanseri veri kümesi için sınıflandırma sonuçları (LOOCV). 55

Tablo 4.22: Meme kanseri veri 2 kümesi için sınıflandırma sonuçları (hold-out). 57

Tablo 4.23: Meme kanseri 2 veri kümesi için sınıflandırma sonuçları (LOOCV). 57

Tablo 4.24: CNS kanseri veri kümesi için sınıflandırma sonuçları (hold-out).....58

Tablo 4.25: CNS kanseri veri kümesi için sınıflandırma sonuçları (CV).....59

Tablo 4.26: CNS kanseri veri kümesi için sınıflandırma sonuçları (LOOCV). 59

Tablo 4.27: Kolon kanseri veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları. 61

Tablo 4.28: Prostat kanseri veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları. 63

Tablo 4.29: DLBCL veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları. 65

Tablo 4.30: SMK veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları. 67

Tablo 4.31: CNS veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları. 68

Tablo 4.32: Meme kanseri veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları. 70

Tablo 4.33: Lösemi veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 72

Tablo 4.34: Kolon kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 73

Tablo 4.35: Prostat kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 73

Tablo 4.36: Over kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 74

Tablo 4.37: DLBCL veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 74

Tablo 4.38: SMK veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 75

Tablo 4.39: Meme kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 75

Tablo 4.40: Meme kanseri 2 veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 76

Tablo 4.41: CNS veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 77

Tablo 4.42: COAD veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 77

Tablo 4.43: BRCA veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları. 78

Tablo 4.44: GA içeren sarmal yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları. 79

Tablo 4.45: GA içeren sarmal yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları. 79

Tablo 4.46: GA içeren sarmal yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları. 80

Tablo 4.47: GA içeren sarmal yöntemlerin SMK veri kümesi için sınıflandırma sonuçları. 80

Tablo 4.48: GA içeren sarmal yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları. 81

Tablo 4.49: GA içeren sarmal yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları. 81

Tablo 4.50: GA içeren sarmal yöntemlerin CNS veri kümesi sınıflandırma için sonuçları. 82

Tablo 4.51: GA içeren sarmal yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları. 82

Tablo 4.52: GA içeren sarmal yöntemlerin COAD veri kümesi için sınıflandırma sonuçları. 83

Tablo 4.53: PSO içeren sarmal yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları. 83

Tablo 4.54: PSO içeren sarmal yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları. 84

Tablo 4.55: PSO içeren sarmal yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları. 84

Tablo 4.56: PSO içeren sarmal yöntemlerin SMK veri kümesi için sınıflandırma sonuçları. 85

Tablo 4.57: PSO içeren sarmal yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları. 85

Tablo 4.58: PSO içeren sarmal yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları. 86

Tablo 4.59: PSO içeren sarmal yöntemlerin CNS veri kümesi için sınıflandırma sonuçları. 86

Tablo 4.60: PSO içeren sarmal yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları. 87

Tablo 4.61: PSO içeren sarmal yöntemlerin COAD veri kümesi için sınıflandırma sonuçları. 87

Tablo 4.62: GA tabanlı önerilen yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları. 88

Tablo 4.63: PSO tabanlı önerilen yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları. 89

Tablo 4.64: GA tabanlı önerilen yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları. 90

Tablo 4.65: PSO tabanlı önerilen yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları. 91

Tablo 4.66: GA tabanlı önerilen yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları. 92

Tablo 4.67: PSO tabanlı önerilen yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları. 93

Tablo 4.68: GA tabanlı önerilen yöntemlerin SMK veri kümesi için sınıflandırma sonuçları. 94

Tablo 4.69: PSO tabanlı önerilen yöntemlerin SMK veri kümesi için sınıflandırma sonuçları. 95

Tablo 4.70: GA tabanlı önerilen yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları. 96

Tablo 4.71: PSO tabanlı önerilen yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları. 97

Tablo 4.72: GA tabanlı önerilen yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları. 98

Tablo 4.73: PSO tabanlı önerilen yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları. 99

Tablo 4.74: GA tabanlı önerilen yöntemlerin CNS veri kümesi için sınıflandırma sonuçları. 100

Tablo 4.75: PSO tabanlı önerilen yöntemlerin CNS veri kümesi için sınıflandırma sonuçları. 101

Tablo 4.76: GA tabanlı önerilen yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları. 102

Tablo 4.77: PSO tabanlı önerilen yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları. 103

Tablo 4.78: GA tabanlı önerilen yöntemlerin COAD veri kümesi için sınıflandırma sonuçları. 104

Tablo 4.79: PSO tabanlı önerilen yöntemlerin COAD veri kümesi için sınıflandırma sonuçları. 105

Tablo 5.1: Başlıca sınıflandırma yöntemleri için doğruluk oranı (%) sonuçları (CV). 106

Tablo 5.2: Önerilen hibrit yöntemlerin ve sarmal yöntemlerin doğruluk oranı (%) sonuçları (CV). 108

Tablo 5.3: Yöntemlerin seçtiği sınıflandırmada kullanılan gen sayıları. 109

Tablo 5.4: Önerilen yöntemlerin ve diğer literatür çalışmalarının (%) doğruluk oranı (seçilen gen sayısı) sonuçları. 111

Tablo Ek.1.1: Kolon kanseri veri kümesi için DDVM yöntemi sınıflandırma sonuçları. 122

Tablo Ek.1.2: Kolon kanseri veri kümesi için PDVM yöntemi sınıflandırma sonuçları. ...123

Tablo Ek.1.3: Kolon kanseri veri kümesi için KNN yöntemi sınıflandırma sonuçları. 123

Tablo Ek.1.4: Kolon kanseri veri kümesi için NB yöntemi sınıflandırma sonuçları. 124

Tablo Ek.1.5: Kolon kanseri veri kümesi için KA yöntemi sınıflandırma sonuçları. 124

Tablo Ek.1.6: Kolon kanseri veri kümesi için İBYSA yöntemi sınıflandırma sonuçları. 125

Tablo Ek.1.7: Prostat kanseri veri kümesi için DDVM yöntemi sınıflandırma sonuçları. 125

Tablo Ek.1.8: Prostat kanseri veri kümesi için PDVM yöntemi sınıflandırma sonuçları. 126

Tablo Ek.1.9: Prostat kanseri veri kümesi için KNN yöntemi sınıflandırma sonuçları. 126

Tablo Ek.1.10: Prostat kanseri veri kümesi için NB yöntemi sınıflandırma sonuçları. 127

Tablo Ek.1.11: Prostat kanseri veri kümesi için KA yöntemi sınıflandırma sonuçları.....127

Tablo Ek.1.12: Prostat kanseri veri kümesi için İBYSA yöntemi sınıflandırma sonuçları. 128

Tablo Ek.1.13: DLBCL veri kümesi için DDVM yöntemi sınıflandırma sonuçları.....128

Tablo Ek.1.14: DLBCL veri kümesi için PDVM yöntemi sınıflandırma sonuçları. 129

Tablo Ek.1.15: DLBCL veri kümesi için KNN yöntemi sınıflandırma sonuçları. 129

Tablo Ek.1.16: DLBCL veri kümesi için NB yöntemi sınıflandırma sonuçları. 130

Tablo Ek.1.17: DLBCL veri kümesi için KA yöntemi sınıflandırma sonuçları. 130

Tablo Ek.1.18: DLBCL veri kümesi için İBYSA yöntemi sınıflandırma sonuçları. 131

Tablo Ek.1.19: SMK veri kümesi için DDVM yöntemi sınıflandırma sonuçları. 131

Tablo Ek.1.20: SMK veri kümesi için PDVM yöntemi sınıflandırma sonuçları. 132

Tablo Ek.1.21: SMK veri kümesi için KNN yöntemi sınıflandırma sonuçları. 132

Tablo Ek.1.22: SMK veri kümesi için NB yöntemi sınıflandırma sonuçları. 133

Tablo Ek.1.23: SMK veri kümesi için KA yöntemi sınıflandırma sonuçları. 133

Tablo Ek.1.24: SMK veri kümesi için İBYSA yöntemi sınıflandırma sonuçları. 134

Tablo Ek.1.25: Meme kanseri veri kümesi için DDVM yöntemi sınıflandırma sonuçları. 134

Tablo Ek.1.26: Meme kanseri veri kümesi için PDVM yöntemi sınıflandırma sonuçları. 135

Tablo Ek.1.27: Meme kanseri veri kümesi için KNN yöntemi sınıflandırma sonuçları.....135

Tablo Ek.1.28: Meme kanseri veri kümesi için NB yöntemi sınıflandırma sonuçları. 136

Tablo Ek.1.29: Meme kanseri veri kümesi için KA yöntemi sınıflandırma sonuçları.....136

Tablo Ek.1.30: Meme kanseri veri kümesi için İBYSA yöntemi sınıflandırma sonuçları. 137

SİMGE VE KISALTMA LİSTESİ

Simgeler	Açıklama
A()	: Sınıflandırıcının doğruluk oranı fonksiyonu
F()	: F-test fonksiyonu
f()	: Aktivasyon fonksiyonu
G()	: Bilgi Kazancı fonksiyonu
g()	: GA veya PSO için gen veya parçacık değeri
$gbest_j$	: PSO global en iyi pozisyonu
H()	: Entropi fonksiyonu
I()	: Karşılıklı bilgi fonksiyonu
J()	: LDA için maksimizasyon fonksiyonu
L	: Maksimum ilgi koşulu
M	: Kovaryans matrisi
net_i	: YSA gizli katmanının i. nöronu için net girdi değeri
$Pbest_{i,j}$	: PSO lokal en iyi pozisyon
r	: Pearson Korelasyon Katsayısı
R()	: Relief-F fonksiyonu
s()	: Birey veya parçacıkta 1 olan gen sayısı fonksiyonu
S_A	: Sınıflar arası saçılım matrisi
S_W	: Sınıf içi saçılım matrisi
T	: Minimum tekrar koşulu
U()	: Uygunluk fonksiyonu
V	: Gizli katmanın ağırlık katsayıları
v	: Kovaryans matrisinin özvektörü
$V_{(i,j)}$	: PSO i. parçacığın j. boyutu için hız değeri
W	: Çıkış katmanının ağırlık katsayıları
$X_{(i,j)}$	: PSO i. parçacığın j. boyutu için pozisyon değeri
y_i	: YSA gizli katmanının i. nöronu için çıkış değeri
Z	: Dönüşüm matrisi
Z_i	: YSA çıkış katmanın i. nöronu için değeri
λ	: Kovaryans matrisinin özdeğeri

Kısaltmalar	Açıklama
ABC	: Yapay Arı Kolonisi (Artificial Bee Colony)
ACO	: Karınca Kolonisi Optimizasyonu (Ant Colony Optimization)
AGA	: Uyarlamalı Genetik Algoritma (Adaptive Genetic Algorithm)
AKNN	: Uyarlamalı K En Yakın Komşu (Adaptive K Nearest Neighbour)
ALL	: Akut lenfoblastik lösemi (Acute Lymphoblastic Leukemia)
AML	: Akut miyeloid lösemi (Acute Myeloblastic Leukemia)
BBO	: Biyocoğrafya Tabanlı Optimizasyon (Biogeography Based Optimization)
BM	: Önyüklenmiş Mesafe (Bootstrapped Margin)
BRCA	: İnvaziv Meme Karsinomu (Breast Invasive Carsinoma)
cDNA	: Tamamlayıcı DNA (Complement DNA)
CHS	: Kaotik Armoni Arama (Chaotic Harmony Search)
COAD	: Kolon Adenokarsinomu (Colon Adenocarsinoma)
CV	: Çapraz Doğrulama (Cross Validation)
DLBCL	: Yaygın Büyük B-hücreli lenfoma (Diffuse Large B-Cell Lymphoma)
DNA	: Deoksiribonükleik Asit (Deoxyribonucleic Acid)
DVM	: Destek Vektör Makinesi (Support Vector Machine)
DDVM	: Doğrusal Destek Vektör Makinesi (Linear Support Vector Machine)
ELM	: Aşırı Öğrenme Makinesi (Extreme Learning Machine)
EO	: Evrimsel Operatörler (Evolutionary Operators)
FKF	: F-test Korelasyon Farkı (F-test Correlation Difference)
FKO	: F-test Korelasyon Oranı (F-test Correlation Quotient)
FL	: Foliküler lenfoma (Follicular Lymphoma)
FLD	: Fisher Lineer Diskriminant (Fisher Linear Discriminant)
FOIL	: Birinci Dereceden Tümevarımlı Öğrenici (First Order Inducticve Learner)
FRFS	: FOIL Kuralı Tabanlı Algoritma (FOIL Rule Based Algorithm)
FS	: Fisher Skoru (Fisher Score)
GA	: Genetik Algoritma (Genetic Algorithm)
GBC	: Genetik Arı Kolonisi (Genetic Bee Colony)
GPSO	: Geometrik Parçacık Sürü Optimizasyonu (Geometric Particle Swarm Optimization)
IG	: Bilgi Kazancı (Information Gain)
İBYSA	: İleri Beslemeli Yapay Sinir Ağı (Feedforward Artificial Neural Network)
KA	: Karar Ağacı (Decision Tree)

KBF : Karşılıklı Bilgi Farkı (Mutual Information Difference)

KBO : Karşılıklı Bilgi Oranı (Mutual Information Quotient)

KNN : K En Yakın Komşu (K Nearest Neighbour)

LDA : Lineer Diskriminant Analiz (Linear Discriminant Analysis)

LOOCV : Biri Hariç Çapraz Doğrulama (Leave One Out Cross Validation)

MBKH : Değiştirilmiş İkili Kril Sürüsü Algoritması (Modified Binary Krill Herd Algorithm)

MI : Karşılıklı Bilgi (Mutual Information)

MIM : Karşılıklı Bilgi Maksimizasyonu (Mutual Information Maximization)

MLP : Çok Katmanlı Algılayıcı (Multi-Layer Perceptron)

MOCEPO : Çok Amaçlı Kaotik İmparator Penguen Optimizasyonu (Multi-Objective Chaotic Emperor Penguin Optimization)

MOCS : Çok Amaçlı Guguk Kuşu Arama Algoritması (Multi-Objective Cuckoo Search Algorithm)

MRMR : Minimum Tekrar Maksimum İlgi (Minimum Redundancy Maximum Relevance)

mRNA : Mesajcı RNA (Messenger RNA)

NB : Naive Bayes (Naive Bayes)

OSA : Olasılıksal Sinir Ağları (Probabilistic Neural Networks)

PAC : Olası Yaklaşık Doğru (Probably Approximately Correct)

PDVM : Polinomal Destek Vektör Makinesi (Polynomial Support Vector Machine)

PSO : Parçacık Sürü Optimizasyonu (Particle Swarm Optimization)

RF : Rastgele Orman (Random Forest)

RFE : Özyinelemeli Öznitelik Eliminasyonu (Recursive Feature Elimination)

RNA : Ribonükleik Asit (Ribonucleic Acid)

SFS : Sıralı İleri Seçim (Sequential Forward Selection)

SNR : Sinyal-Gürültü Oranı (Signal to Noise Ratio)

SVM : Destek Vektör Makinesi (Support Vector Machine)

SUS : Stokastik Evrensel Seçim (Stochastic Universal Selection)

TBA : Temel Bileşen Analizi (Principal Component Analysis)

YSA : Yapay Sinir Ağı (Artificial Neural Network)

ÖZET

DOKTORA TEZİ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE BİYOMEDİKAL VERİDEN KANSER TEŞHİSİ

Öznur Sinem SÖNMEZ

İstanbul Üniversitesi-Cerrahpaşa

Lisansüstü Eğitim Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Dr. Öğr. Üyesi Mustafa DAĞTEKİN

Günümüzde kanser en çok ölüme neden olan hastalıklardandır. Kanserin erken seviyede ve doğru teşhisinin kanser tedavisi, kanserin ilerlemesini yavaşlatma ve hayatta kalma oranını arttırma açısından önemli olduğu bilinmektedir. Bu nedenle, genetik araştırmalar önem kazanmıştır. Gen ifade seviyelerinin belirlenmesini sağlayan mikrodizi ve RNA dizileme teknolojilerinin gelişmesiyle birlikte binlerce genin ifade seviyelerinin incelenmesi ve hastalıkların gen profillerinin çıkarılması gibi konularda çalışmalar yapılmaktadır. Gen ifade seviyeleri değerlendirilerek kanserli ve sağlıklı doku ayrımı veya kanserin alt türlerinin birbirinden ayrımı yapılabilmektedir. Bu değerlendirme sürecindeki birbiriyle ilişkili yüksek miktardaki veri nedeniyle veriyi analiz etmek, süreci daha verimli ve sonuçları daha doğru hale getirmek için bilgisayar algoritmalarına ve makine öğrenmesi yöntemlerine ihtiyaç duyulmaktadır. Bu çalışmada, gen ifade seviyelerini içeren veri kümelerinden öznitelik (gen) seçimi gerçekleştirilerek kanser teşhisi ve sınıflandırılması için yeni hibrit yöntemler önerilmiştir. Önerilen yöntemlerde, Genetik Algoritma - K En Yakın Komşu (GA-KNN), Genetik Algoritma - Destek Vektör Makinesi (GA-DVM), Parçacık Sürü Optimizasyonu - K En Yakın Komşu (PSO-KNN) ve PSO-DVM yöntemleri Pearson Korelasyonu, Karşılıklı Bilgi veya Relief-F yöntemleriyle geliştirilmiştir. GA'nın çaprazlama işlemi gen seçimi için özelleştirilmiştir. 9 farklı veri kümesi kullanılarak önerilen yöntemlerin sınıflandırma performansları değerlendirilmiştir. Ayrıca, literatürdeki başlıca filtre yöntemlerin, boyut indirgeme yöntemlerinin, sarmal yöntemlerin ve başlıca sınıflandırma yöntemlerinin sınıflandırma sonuçlarına olan etkileri karşılaştırılarak incelenmiştir. Önerilen yöntemlerin

sınıflandırma performansları diğer başlıca yöntemlerle ve diğer literatür çalışmalarıyla karşılaştırılmıştır. Sınıflandırma sonuçları, tüm veri kümeleri için tüm değerlendirme ölçütleri açısından önerilen yöntemlerle daha başarılı sonuçlar elde edildiğini göstermektedir.

Ocak 2022, 160 sayfa.

Anahtar kelimeler: Kanser teşhisi ve sınıflandırması, gen ifade seviyeleri, genetik algoritma, hibrit öznitelik seçimi yöntemleri

SUMMARY

Ph.D. THESIS

CANCER DIAGNOSIS FROM BIOMEDICAL DATA WITH MACHINE LEARNING METHODS

Öznur Sinem SÖNMEZ

Istanbul University-Cerrahpasa

Institute of Graduate Studies

Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Mustafa DAĞTEKİN

Nowadays, cancer is the leading cause of death among the diseases. It is known that early and accurate diagnosis of cancer is significant in terms of cancer treatment, slowing the progression of cancer and increasing the survival rate. Therefore, genetic research has gained importance. Studies are carried out on subjects such as analysing the expression levels of thousands of genes and gene profiling of diseases with the development of microarray and RNA sequencing technologies, which enable the determination of gene expression levels. Cancerous and healthy tissue or subtypes of cancer can be distinguished by evaluation of gene expression levels. Due to the high amount of interrelated data in this evaluation process, computer algorithms and machine learning methods are needed to analyse the data and make the process more efficient and the results more accurate. In this study, new hybrid methods are proposed for cancer diagnosis and classification by performing feature (gene) selection from datasets containing gene expression levels. In the proposed methods, Genetic Algorithm - K Nearest Neighbour (GA-KNN), Genetic Algorithm - Support Vector Machine (GA-SVM), Particle Swarm Optimization - K Nearest Neighbour (PSO-KNN) and PSO-SVM methods are improved with Pearson Correlation, Mutual Information or Relief-F methods. The crossover operation of GA is specialized for gene selection. The classification performances of proposed methods are evaluated by using nine different datasets. In addition, the effects of the main filter methods, dimension reduction methods, wrapper methods and main classification methods on classification results are analysed. The classification performances of proposed methods are

compared with the other main methods and other studies in the literature. Classification results have demonstrated that better results are obtained with the proposed methods in terms of all evaluation metrics for all datasets.

January 2022, 160 pages.

Keywords: Cancer diagnosis and classification, gene expression levels, genetic algorithm, hybrid feature selection methods

1. GİRİŞ

Kanser, 100'den fazla farklı türüyle en çok ölüme sebebiyet veren hastalıklardandır. Dünya Sağlık Örgütüne göre dünyada her yıl kanserden 8 milyon insan hayatını kaybetmektedir. Vakalardan 4 milyonu erken teşhis edilmesi halinde önlenebilecek kanser türlerinden oluşmaktadır. Kanser tedavisi, kanserin ilerlemesini yavaşlatma ve hayatta kalma oranını arttırma açısından erken ve doğru teşhisin son derece önemli olduğu bilinmektedir.

Günümüzde kanser teşhisinde geleneksel yöntemler tümör belirteçlerinin araştırılması, biyomedikal görüntüleme teknikleriyle elde edilen görüntülerin değerlendirilmesi ve belirlenen kitle veya lezyonlardan alınan doku örneklerinin histopatolojik incelemesidir. Bu geleneksel yöntemlerin, erken ve doğru teşhis konularında belirli bazı sınırlamaları bulunmaktadır. Örneğin görüntüleme teknikleri belirli bir boyutun altında kalan kitlelerin belirlenmesinde yetersiz kalabilmektedir. Tümör belirteçleri, teşhisten çok tedavi ve nüks takibinde daha belirleyici olmaktadır. Dolayısıyla erken teşhis, kanser türü ve alt türlerinin doğru olarak belirlenmesi süreçlerini iyileştirmek için kanseri moleküler düzeyde belirlemeye olanak sağlayan genetik araştırmalar önem kazanmaktadır. Son yıllarda mikrodizi ve RNA dizileme teknolojilerinin gelişmesiyle genetik araştırmalar hız kazanmıştır.

Bu teknolojiler, DNA molekülünün ve gen ifade seviyelerinin incelemesini sağlayan teknolojilerdir. Gen ifade seviyeleri değerlendirilerek kanserli ve sağlıklı doku ayrımı veya kanserin alt türlerinin birbirinden ayrımı yapılabilmektedir. Böylece, kanserle ilişkili genlerin belirlenmesi, hastalıkların gen profillerinin çıkarılması konularında çalışmalar gerçekleştirilmektedir. Mikrodiziler, tek deneyde binlerce genin ifade seviyelerinin incelenmesine olanak sağlar. Kanser teşhisi ve sınıflandırması için binlerce gen ifade seviyesinin birlikte değerlendirilmesi gerekmektedir. Bu değerlendirme sürecindeki birbiriyle ilişkili yüksek miktarlardaki veri nedeniyle bilgisayarlara ve makine öğrenmesi yöntemlerine ihtiyaç duyulmaktadır.

Gen ifade seviyelerini içeren verilerle oluşturulmuş herkese açık veri kümeleri bulunmaktadır. Bu veri kümeleri makine öğrenmesi yöntemleriyle gen ifade analizleri, kanser teşhisi ve sınıflandırması araştırmalarında sıklıkla kullanılmaktadır. Bu araştırmalarda karşılaşılan en önemli sorunlarından biri veri kümelerinin yüksek boyutlu olmasıdır. Özellikle, mikrodizi veri

kümeleri binlerce genden ve az sayıda örnekten oluşmaktadır. Bu genlerin çoğunluğu, çalışılan sınıflandırma süreci için sınıfı tanımlayan, sınıfla ilişkili olan genler arasında yer almamaktadır. Bu veri kümeleriyle çalışmanın bir diğer olumsuz yanı, genler arasındaki etkileşimin diğer veri kümelerine göre çok daha yüksek olmasıdır. Örneğin, bazı genler başka genleri aktive edebilir veya gen ifadesini tetikleyebilir. Dolayısıyla tekrarlı, ilgisiz ve gürültülü genler makine öğrenmesiyle kanser teşhisi ve sınıflandırması uygulamaları için hesaplama yükünü arttırmakta ve sınıflandırma başarısını olumsuz etkilemektedir. Bu olumsuz etkileri azaltmak için veriyi en iyi yansıtan, en az sayıda gen içeren öznitelik alt kümesinin seçilmesi işlemi sınıflandırmadan önce gerçekleştirilmektedir.

Literatürdeki öznitelik seçme yöntemleri filtre, sarmal, gömülü ve hibrit yöntemler olmak üzere dört kategoride incelenmektedir. Filtre yöntemler, genellikle öğrenme içermeyen istatistiksel yöntemlerdir. Sarmal yöntemler, bir arama algoritmasıyla sınıflandırma yönteminin birlikte kullanılmasıyla oluşturulan yöntemlerdir. Gömülü yöntemler, öznitelikleri derecelendirmek için bir kriter oluşturmak üzere sınıflandırıcıyı kullanan yöntemlerdir. Hibrit yöntemler, diğer yöntemlerin birkaçının birlikte kullanılmasıyla oluşturulmaktadır. Farklı yöntemlerin avantajlarını birleştirdiği için genellikle en başarılı yöntemler hibrit yöntemlerden çıkmaktadır. Öznitelik seçimi yöntemlerinin ayrıntıları Bölüm 3.1'de anlatılmaktadır.

Bu tez çalışmasında, gen ifade seviyelerini içeren veri kümelerinden öznitelik seçimi, bu veri kümelerinin sınıflandırılması ve kanser teşhisi için yeni hibrit yöntemler önerilmiştir. Bu kapsamda; öncelikle literatürde sıklıkla kullanılan filtre, sarmal öznitelik seçimi yöntemlerinin, boyut indirgeme yöntemlerinin ve sınıflandırma yöntemlerinin başarılarının karşılaştırılması gerçekleştirilmiştir. Bu karşılaştırmada, değerlendirme ölçütleri açısından en başarılı sonuçların elde edildiği yöntemler belirlenmiştir. Belirlenen bu yöntemleri içeren yeni hibrit yöntemler önerilmiştir. Bu hibrit yöntemlerde; GA-KNN, GA-DVM, PSO-KNN, PSO-DVM sarmal yöntemleri Pearson Korelasyonu, Karşılıklı Bilgi ve Relief-F filtre yöntemleriyle geliştirilmiştir. GA tabanlı önerilen yöntemlerde, GA seçim ve çaprazlama işlemleri filtre yöntemler kullanılarak özelleştirilmiştir. Önerilen yöntemlerin başarısı gen ifade seviyelerini içeren 9 farklı veri kümesi kullanılarak değerlendirilmiştir. Sonuçlar, diğer başlıca yöntemlerle, geleneksel sarmal yöntemlerle ve diğer literatür çalışmalarıyla karşılaştırılmıştır. Tüm veri kümeleri için tüm değerlendirme ölçütleri açısından önerilen yöntemlerle daha başarılı sonuçlar elde edildiği gösterilmiştir.

Genel Kısımlar bölümünde; kanser, gen ifadesi, gen ifade seviyelerinin belirlenmesine olanak sağlayan mikrodizi ve RNA dizileme teknolojilerinin temelleri ve literatürdeki diğer çalışmalar anlatılmıştır. Malzeme ve Yöntem bölümünde, bu çalışmada kullanılan öznitelik seçimi, boyut indirgeme, sınıflandırma ve optimizasyon yöntemlerinin ve önerilen yöntemlerin ayrıntıları anlatılmıştır. Ayrıca, yöntemlerin değerlendirilmesinde kullanılan veri kümeleri ve değerlendirme ölçütleri de bu bölümde açıklanmıştır. Bulgular bölümünde, başlıca öznitelik seçimi, boyut indirgeme, sınıflandırma yöntemlerinin ve önerilen yöntemlerin gen ifade seviyelerini içeren veri kümelerini sınıflandırma sonuçları sunulmuştur. Tartışma ve sonuç bölümünde önerilen yöntemlerin ve diğer yöntemlerin sonuçları karşılaştırılarak değerlendirilmiştir.

2. GENEL KISIMLAR

2.1. KANSER

Kanser, anormal hücrelerin düzensiz ve kontrolsüz çoğalmasıdır (Hejmadi, 2010). Normal hücrelerde hücre bölünmesini kontrol eden binlerce gen düzenli bir şekilde çalışmaktadır. Bu şekilde hücre bölünmesi, baskılanması ve ihtiyaç halinde hasarlı hücrelerin apoptozu (hücre ölümü) sağlanmaktadır. Hücrede bu kontrolü sağlayan genlerdeki birçok mutasyon sonucunda anormal (kanserli) hücreler oluşabilmektedir. Bu kanserli hücreler, birçok açıdan normal hücrelerden farklı olabilirler. Belli başlı farklılıklar; diğer hücrelerden gelen sinyallere tepki vermeme, aşırı büyüme ve çoğalma eğiliminde olma, normal hücrelerin fonksiyonlarını yerine getirememe, hücre yenilenmesi ve apoptoz işlemlerini gerçekleştirememe ve bulundukları dokulardan başka dokulara yayılabilme olarak sıralanabilir. Bu ve benzeri farklılıklar kanserli hücrelerin, öncelikle bulundukları doku ve organlarda kontrolsüz olarak çoğalmasına, sonrasında başka doku veya organlara yayılmasına (metastaz) neden olabilmektedir.

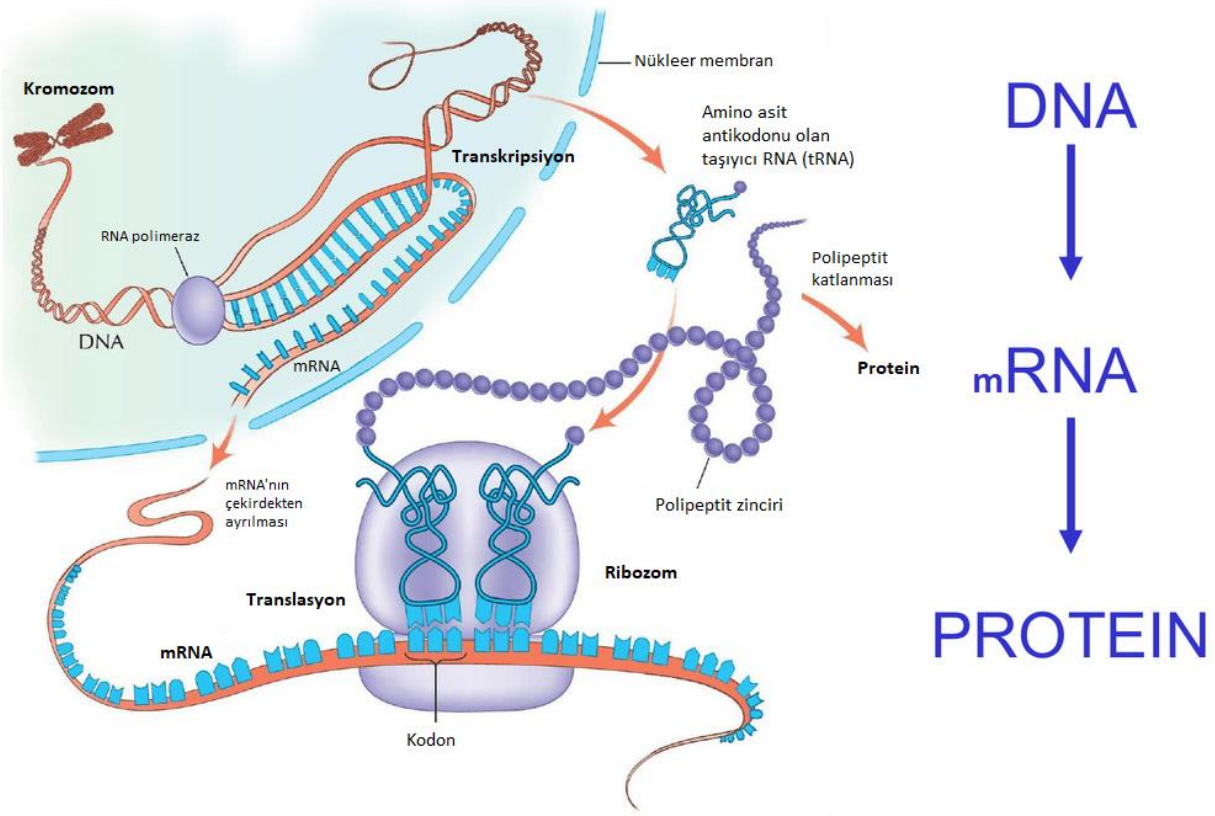
Kanser, histolojik olarak 6 ana gruba ayrılmaktadır (Devita ve diğ., 2011). Bunlar; karsinom, sarkom, miyelom, lenfoma, lösemi ve karma türler olarak adlandırılmaktadır. Tablo 2.1'de bu kanser sınıflarının temelleri gösterilmektedir.

Tablo 2.1: Kanserin histolojik olarak ana grupları.

Kanser türü	**Geliştiği doku**
Karsinom	Epitel kökenli malign neoplazm. İki alt türü vardır. Adenokarsinom: Bir organ veya bezden gelişir. Skuamöz hücreli karsinom: Skuamöz epitelden gelişir.
Sarkom	Kemik, tendon, kıkırdak, kas, yağ gibi dokulardan gelişir.
Miyelom	Kemik iliğinin plazma hücrelerinden gelişen kanserlerdir.
Lenfoma	Lenfositlerin kanseridir.
Lösemi	Kemik iliği kanserleridir.
Karma türler	Birden fazla bileşenden oluşan kanserlerdir. Örn: Karsinosarkom.

2.2. GEN İFADESİ

Gen ifadesi, bir genin bilgisinin fonksiyonel bir gen ürünü sentezinde kullanıldığı süreçtir (Raghavachari ve Garcia, 2018). Gen ürünleri proteinler veya kodlamayan RNA olabilir. DNA'dan proteine gen ifadesi aşamaları, Şekil 2.1'de gösterilmektedir. Bu süreç, hücre çekirdeğindeki DNA'dan transkripsiyon ile mRNA oluşturulması, mRNA'nın işlenerek kodlamada yer almayacak bölümlerinin çıkarılması, mRNA'nın çekirdekten ayrılarak sitoplazmaya geçmesi, ribozomlarda translasyon ile mRNA'dan polipeptit kodlanması ve polipeptitlerin katlanarak proteinleri oluşturması aşamalarını kapsamaktadır.



Şekil 2.1: DNA'dan proteine bilgi aktarımı aşamaları (Barton ve diğ., 2007).

Hücreler, normal işlevlerini yerine getirmek veya hayatta kalabilmek için gen ifadesini düzenleyerek gerektiğinde protein gibi fonksiyonel bir gen ürünü oluşturulmasını sağlamaktadırlar. Gen ifadesi sürecinin aşamalarının düzenlenmesiyle fizyolojik ve patolojik süreçlerdeki gen ifade zamanlaması, gen ürünü miktarı ve yeri kontrol edilebilmektedir (Raghavachari ve Garcia, 2018). Bu kontrol, gen fonksiyonlarını önemli ölçüde etkileyerek

hücre gelişimi ve farklılaşmasını düzenlemektedir. Örneğin, bazı durumlarda gen ifade seviyeleri artış gösterebilir veya baskılanabilir. Bu nedenle, gen ifade seviyeleri farklı fizyolojik ve patolojik durumlarda değişiklik göstermektedir.

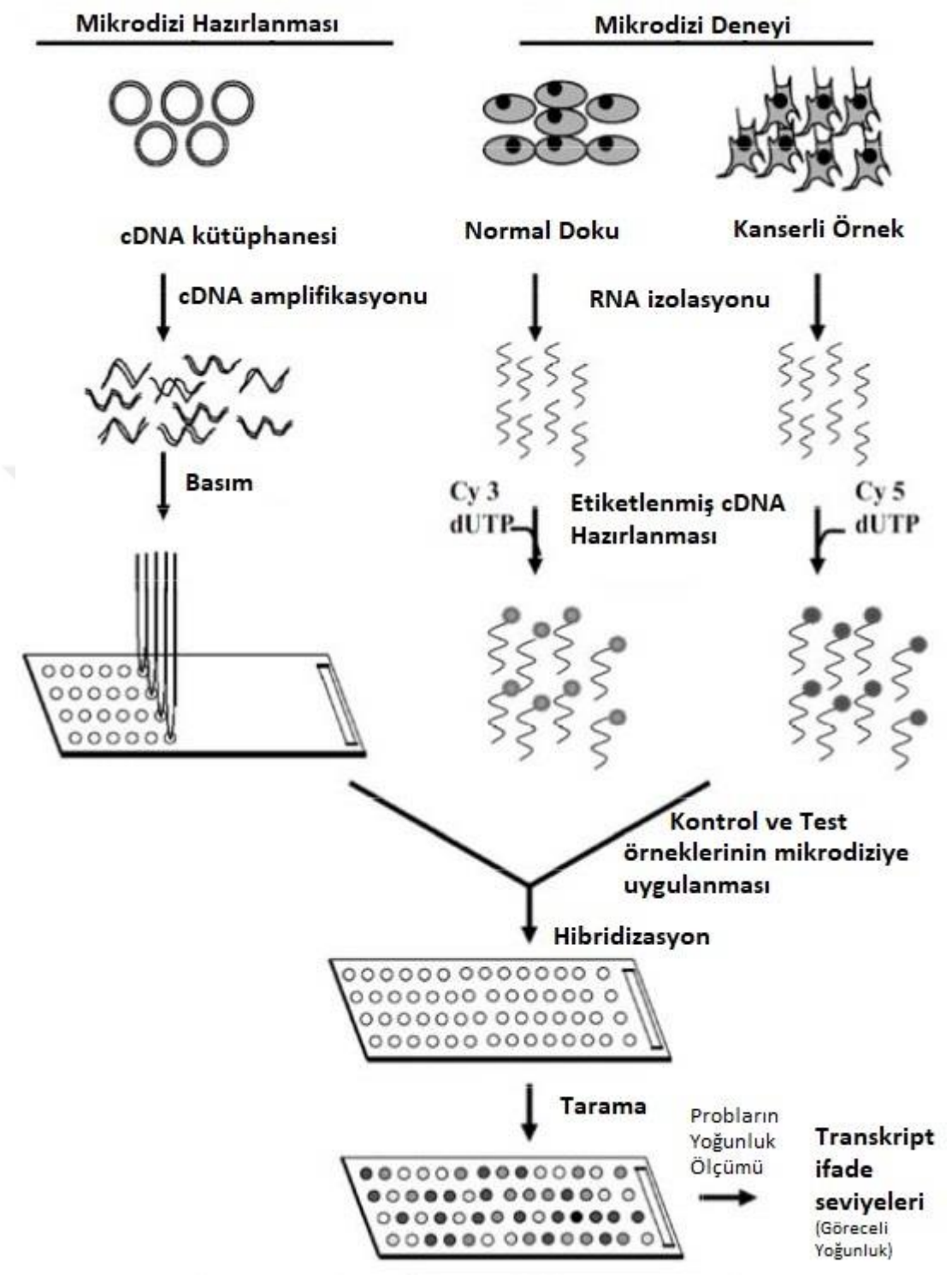
Gen ifade analizleri, bilinen genler için gen ifade seviyelerinin belirlenmesi ve nicelenmesine olanak sağlamaktadır. Gen ifade seviyelerinin analizi, gen profillerinin çıkarılması, hastalıkta etkili olan genlerin ve yolakların belirlenmesi gibi süreçlerde kullanılmaktadır. Bu sayede, moleküler düzeyde kanser teşhisi veya diğer yöntemlerle birbirinden ayrılması zor olan kanser alt türlerinin (örneğin: akut miyeblastik lösemi ve akut lenfoblastik lösemi) belirlenmesi gibi işlemler gerçekleştirilmektedir.

2.3. MİKRODİZİ VE RNA DİZİLEME TEKNOLOJİLERİ

Gen ifade seviyelerinin belirlenmesinde kullanılan en önemli teknolojiler mikrodizi ve RNA dizileme (RNA-sequencing) yöntemleridir. Mikrodizi teknolojisi, DNA veya oligonükleotit dizileri içeren ve aynı deneyde binlerce genin ifade seviyelerinin belirlenmesine olanak sağlayan teknolojidir. RNA dizileme yöntemi de örnekteki tüm genlerin ifade seviyelerinin aynı deneyde ölçülmesini sağlamaktadır.

Mikrodiziler; hastalık teşhisi, hastalıkların alt türlerinin sınıflandırılması, ilaç keşfi ve geliştirilmesi, biyokimyasal yolakların ve gen ifadelerinin profillenmesi gibi birçok araştırma ve klinik çalışmada kullanılmaktadır (Rabie ve diğ., 2021). Mikrodiziler, üretim teknolojilerine göre farklılık göstermekle birlikte belirli bir hücrenin genomunu temsil etmek üzere seçilen cDNA (tamamlayıcı DNA) veya oligonükleotitlerin cam slayt, plastik slayt veya işlevselleştirilmiş silikon plaka üzerine nokta basımı, inkjet teknolojisi, elektrokimyasal sentez gibi teknolojilerle yerleştirilmesiyle elde edilmektedirler (Lee, 2016; Rueda ve Ali, 2014).

Şekil 2.2'de mikrodizi hazırlanması ve deneyinin aşamaları gösterilmektedir. Mikrodiziler temel olarak cDNA veya oligonükleotit dizileri içermektedir. cDNA, mRNA'dan ters transkripsiyonla elde edilen DNA parçasıdır. Oligonükleotitler, DNA veya RNA'nın kısa nükleotit dizileridir. Mikrodizi deneyinin aşamaları; kontrol (normal doku) ve test (kanserli doku) örneklerinden RNA izolasyonu, mRNA'dan ters transkripsiyonla cDNA oluşturulması, cDNA'nın floresan boya (Cy3 ve Cy5) veya başka bir yöntemle etiketlenmesi, örneklerin mikrodiziye uygulanması, hibridizasyon, mikrodizinin taranması ve probların floresan yoğunluk ölçümüyle transkript seviyelerinin belirlenmesidir.



Şekil 2.2: Mikrodizi hazırlanması ve deneyi (Somasundaram ve diğ., 2002).

RNA dizileme ile örnekte bulunan bütün genlerin transkript (mRNA) seviyeleri ölçülmektedir. RNA dizileme deneyinin temel aşamaları Şekil 2.3'de gösterilmektedir. Buna göre, hücre veya

dokudan izole edilen RNA parçalara ayrılarak ters transkripsiyonla cDNA parçalarına dönüştürülür. cDNA parçaları işaretlenir, yükseltilir ve okuma işlemiyle dizilenir. Diziler belirli genlere göre hizalanarak okuma sayıları ölçülmekte ve transkript seviyeleri belirlenmektedir.



Şekil 2.3: RNA dizileme deneyi temel aşamaları.

RNA dizileme yönteminin, mikrodizi yöntemine göre bazı avantajları vardır. Örneğin, bu yöntemde mikrodizi yönteminden farklı olarak içeriği bilinen prob dizilerine ihtiyaç duyulmaz. İçeriği bilinen veya bilinmeyen tüm transkriptlerin dizilemesi yapılabilir. Bu sayede, yeni gen ve transkriptler veya gen füzyonu gibi yapısal değişimler belirlenebilmektedir. Mikrodizilerde

prob yoğunluğu ölçülürken, RNA dizilemede belirli genlere hizalanan dizilerin okuma sayıları ölçülmektedir. Ayrıca, düşük bolluktaki transkriptlerin belirlenmesinde RNA dizilemenin daha etkili olduğu gösterilmiştir (Zhao ve diğ., 2014). Buna karşın, RNA dizileme daha maliyetli, yüksek miktardaki ham verinin depolanması daha zor ve ham veri analizi daha karmaşıktır.

2.3. GEN İFADE SEVİYELERİNİ İÇEREN VERİ KÜMELERİNİN SINIFLANDIRILMASIYLA İLGİLİ LİTERATÜR ÇALIŞMALARI

Literatürde, gen ifade seviyelerini içeren veri kümelerinin sınıflandırılması ve öznitelik seçimi ile ilgili birçok çalışma bulunmaktadır. Bu çalışmalar filtre, sarmal, gömülü ve hibrit yöntemler olmak üzere farklı öznitelik seçimi yöntemleri içermektedirler. Öznitelik seçimi yöntemlerinin ayrıntıları Bölüm 3.1'de anlatılmaktadır.

Literatürde, Pearson Korelasyonu, Karşılıklı Bilgi (Mutual Information-MI), Bilgi Kazancı (Information Gain-IG), Sinyal-Gürültü Oranı (Signal to Noise Ratio-SNR) gibi birçok yöntem filtre öznitelik seçimi yöntemi olarak kullanılmıştır. Bu yöntemler genellikle gen sayısını belirli bir seviyeye indirgemek için ön işlem olarak uygulanmaktadır.

Liu ve diğ. (2005) çalışmalarında genler arasındaki ilişki ve benzerliği ölçmek için MI yöntemine dayalı bir algoritma önermişlerdir. Bu yöntemle, seçilen alt kümedeki birbiriyle ilişkili gen sayısının azaltılabildiğini göstermişlerdir.

Ding ve Peng (2005) çalışmalarında genler arasındaki ilişkinin en az olduğu, genlerle sınıf arasındaki ilginin en fazla olduğu genlerin belirlenmesine yönelik Minimum Tekrar Maksimum İlgi (Minimum Redundancy Maximum Relevance-MRMR) yöntemini önermişlerdir. Farklı sınıflandırma yöntemleri kullanarak önerdikleri yöntemin sınıflandırma başarısına olan etkilerini incelemiş ve iyileşme elde edildiğini göstermişlerdir.

Al-Batah ve diğ. (2019) korelasyon tabanlı gen seçimi algoritması önermişler ve bu yöntemin doğrudan sınıflandırma yapılmasına göre benzer ve ilişkili genleri azalttığını göstermişlerdir.

Hwang ve diğ. (2010) T-test, entropi gibi istatistiksel ölçütlerden hangisinin o sırada uygulandığı mikrodizi veri kümesi için uygun olduğunu belirleyen filtre yöntem tabanlı bir algoritma önermişler ve bu yöntemle optimal gen alt kümesini belirlemeye çalışmışlardır. DVM ve Rastgele Orman (Random Forest-RF) sınıflandırma yöntemlerini kullanarak önerdikleri

yöntemi diğer filtre yöntemlerle karşılaştırmış ve çoğunlukla daha başarılı olduğunu göstermişlerdir.

Uzma ve Halim (2021) çalışmalarında; Ki-kare, Relief-F ve yerel arama algoritmasına dayalı yeni bir yöntem ileri sürmüşlerdir. DVM, KNN ve RF sınıflandırma yöntemleriyle seçilen alt kümelerin sınıflandırma performanslarını değerlendirip iyileşme olduğunu göstermişlerdir.

Literatürde filtre yöntemlerin öznitelik seçimi performanslarını karşılaştıran çalışmalar bulunmaktadır. Ryu ve Cho (2002) çalışmalarında, lösemi veri kümesi için Pearson Korelasyonu, IG, MI, SNR filtre yöntemlerinin başarılarını karşılaştırmışlardır. DVM, Çok Katmanlı Algılayıcı (Multi-Layer Perceptron-MLP), KNN, Karar Ağacı (KA) sınıflandırma yöntemleri için elde edilen sonuçlara göre Pearson Korelasyonu-MLP yöntemiyle en yüksek sonuçlara ulaşıldığını göstermişlerdir.

Gao ve diğ. (2017) çalışmalarında önerdikleri yöntem için IG, kazanç oranı, korelasyon ve Relief-F yöntemlerinin başarılarını karşılaştırmışlardır. Filtre yöntemlerle seçilen genler, DVM sınıflandırma yöntemi kullanılarak değerlendirilmiştir. Bu değerlendirmeden elde edilen gen alt kümesiyle yapılan sınıflandırma sonuçlarına göre 5 veri kümesinin tamamında IG-DVM ile en yüksek doğruluk oranı sonuçlarına ulaşmışlardır. Korelasyon ve Relief-F yöntemleriyle de bu yönteme yakın sonuçlar elde etmişlerdir.

Relief-F, IG, Kazanç Oranı ve Ki-kare filtre yöntemlerini karşılaştıran Wang ve Makedon (2004), çalışmalarında DVM sınıflandırma yöntemini kullanmışlardır. Sınıflandırma sonuçlarına göre Relief-F ve kazanç oranı yöntemlerinin öne çıktığını göstermişlerdir. Özellikle gen sayısı arttıkça, Relief-F yöntemi diğerlerine üstünlük sağlamaktadır.

Literatürde filtre yöntemlerin başarılarını karşılaştıran çalışmalar olmasına karşın bu çalışmalarda genellikle sadece 3-4 yöntem karşılaştırılmıştır. Daha fazla yöntemin karşılaştırıldığı, yöntemlerin hem farklı veri kümeleri için hem de farklı sınıflandırma yöntemleri için birlikte değerlendirildiği geniş kapsamlı çalışmalara ihtiyaç duyulmaktadır. Bu tez çalışmasında, 11 farklı filtre yöntemin öznitelik seçimi performansları 6 farklı sınıflandırma yöntemi kullanılarak karşılaştırılmıştır. Elde edilen sonuçlar Bulgular 4.2'de sunulmuştur.

Literatürde, arama algoritması ve sınıflandırma yönteminden oluşan birçok sarmal yöntem bulunmaktadır. En sık kullanılan sarmal yöntemler; GA-DVM (Chen, 2006; Alba ve diğ., 2008;

Yıldız ve diğ., 2012), GA-KNN (Gunavathi ve Premalatha, 2014), PSO-DVM (Alba ve diğ., 2008; Li ve diğ., 2013) yöntemleridir. Bunların dışında; GA-ELM (Arunkumar ve diğ., 2017), AGA-KNN (Lee ve diğ., 2011), BM (Chen, 2006), PSO-AKNN (Kar ve diğ., 2015), MMACO-DVM (Alba ve diğ., 2008), EA-KNN (Juliusdottir ve diğ., 2005) gibi birçok farklı sarmal yöntem önerilmiştir.

Arunkumar ve diğ. (2017), GA ile Aşırı Öğrenme Makinesi (Extreme Learning Machine-ELM) yöntemlerini bir arada kullanarak GA-ELM sarmal yöntemini önermişlerdir. Bu yöntemin öznitelik seçme başarısını GA-DVM ve GA-KNN yöntemleriyle karşılaştırmışlardır. GA-ELM yöntemiyle lösemi ve meme kanseri veri kümeleri için daha başarılı sonuçların elde edildiğini göstermişlerdir.

Chen (2006) çalışmasında, GA ve DVM yöntemlerinin öznitelik seçiminde kullanıldığı bir algoritma önermiştir. Bu algoritmanın GA-DVM sarmal yönteminden temel farkı, önyükleme (bootstrapping) yöntemiyle orijinal veri kümesinden rastgele alınan verilerle yeni veri kümelerinin oluşturulmasıdır. Bu veri kümeleri kullanılarak öznitelik seçimi GA-DVM yöntemlerinin birleştirilmesiyle gerçekleştirilmektedir. Önerilen bu yeni sarmal algoritma "Bootstrapped Margin" (BM) algoritması olarak adlandırılmaktadır. BM yöntemiyle kolon kanseri ve lösemi veri kümeleri için GA-KNN yöntemine göre daha yüksek doğruluk oranına ulaşıldığı gösterilmiştir.

Lee ve diğ. (2011), Uyarlamalı Genetik Algoritma (Adaptive Genetic Algorithm-AGA) ile KNN sarmal yöntemlerini birleştirmiş ve AGA-KNN yöntemini önermişlerdir. Bu yöntemin, GA-KNN yöntemine göre daha başarılı olduğunu göstermişlerdir.

Kar ve diğ. (2015), PSO ve Uyarlamalı KNN (Adaptive K Nearest Neighbour-AKNN) yöntemlerinin birlikte kullanıldığı PSO-AKNN yöntemini ileri sürmüşlerdir. Bu yöntemin lösemi veri kümesi doğruluk oranının diğer yöntemlerden daha yüksek olduğu belirtilmiştir.

Fahrudin ve diğ. (2016), GA, PSO, Karınca Kolonisi Optimizasyonu (Ant Colony Optimization-ACO) yöntemlerinin öznitelik seçme başarılarını karşılaştırmışlardır. Sınıflandırma yöntemi olarak Naive Bayes (NB) kullanıldığında F1 ölçütü değerlerine göre, ACO yöntemiyle en başarılı sonuçların elde edildiğini göstermişlerdir. GA ve PSO karşılaştırıldığında PSO, 3 veri kümesi için en iyi sonuçları oluşturmuştur.

Li ve diğ. (2013), karınca sistemi ve karınca kolonisi sistemini içeren ACO tabanlı iki aşamalı bir sarmal yöntem önermişlerdir. ACO-S olarak adlandırılan algoritmanın ilk aşamasında değiştirilmiş karınca sistemiyle ilgisiz genler çıkarılmaktadır. İkinci aşamasında, değiştirilmiş karınca kolonisi sistemi ile alt kümede bulunacak genler geri kalan genlerden seçilmektedir. ACO-S yönteminin; DVM, KNN, NB ve KA sınıflandırma yöntemleriyle oluşturulan sarmal versiyonlarının başarıları incelenmiştir. Ayrıca, çalışmada ACO-S, GA, PSO, ACO yöntemlerinin başarıları karşılaştırılmıştır. ACO-S yönteminin diğerlerine göre daha başarılı olduğu gösterilmiştir.

Yu ve diğ. (2009), tümör belirteç genleri bulmak için değiştirilmiş ACO (MMACO) yöntemini ileri sürmüşlerdir. MMACO-DVM sarmal yönteminin başarısını ACO-DVM ve GA-DVM ile karşılaştırarak MMACO-DVM yönteminin daha başarılı olduğunu göstermişlerdir.

Gömülü yöntemler, öznitelikleri derecelendirmek için bir kriter oluşturmak üzere sınıflandırıcıyı kullanan yöntemlerdir. Başlıca gömülü yöntemler; Guyon ve diğ. (2002) tarafından önerilen DVM'ye dayalı özyineli öznitelik eleme yöntemi (DVM-RFE), Maldonado ve diğ. (2011) tarafından geliştirilen kernel cezalı DVM, RF yönteminin kullanımı (Uriarte ve Andres, 2006; Anaissi ve Kennedy, 2011), kural tabanlı öğrenme yöntemi olan Birinci Dereceden Tümevarımlı Öğrenici (First Order Inducticve Learner-FOIL) algoritmasına dayanan alt küme seçme yöntemi FRFS (Wanga ve diğ., 2013), Olası Yaklaşık Doğru (Probably Approximately Correct-PAC) öğrenmeye dayalı bayes yaklaşımı (Shah ve diğ., 2011), T-test ile geliştirilmiş cezalı DVM (Chan ve diğ., 2016) olarak sıralanabilir. Tablo 2.2'de başlıca gömülü yöntemlerin işleyişi özetlenmiştir.

Tablo 2.2: Öznitelik seçiminde gömülü yöntemler.

Yöntem	**Çalışma Prensibi**
DVM-RFE Guyon ve diğ. (2002)	• Sınıflandırıcı model oluşturulurken elde edilen öznitelik ağırlıklarına göre öznitelikler seçilir. • Algoritma, tüm özniteliklerle başlayarak ayırıcı olmayanları eleyerek devam eder.
KP-DVM Maldonado ve diğ. (2011)	Sınıflandırıcın oluşturulması sırasında: • Öznitelik seçimi ve çekirdek optimizasyonu için gradyan azalma yaklaşımını kullanır. • Destek vektör makinesinin ikili formülasyonundaki özniteliklerin kullanımını cezalandırarak ayırıcı özniteliklerin seçilmesini sağlar.
RF Uriarte ve Andres (2006)	• Orijinal verinin farklı örnekleri kullanılarak birçok karar ağacı oluşturulur. • En az sayıda öznitelik içeren ve hata sayısı en düşük olan seçilir. • Bir diğer yöntem, elde edilen ortalama başarıya göre özniteliklerin seçilmesidir.
FRFS Wanga ve diğ. (2013)	• İlgisiz ve tekrarlı öznitelikleri elemenin yanı sıra öznitelik etkileşimlerini de dikkate alır. • Veri kümesi için kısıtlı FOIL algoritması ile sınıflandırma kurallarını çıkartır. • Bu kuralların tümündeki öncül öznitelikleri birleştirerek tekrarlı öznitelikleri eleyen ve etkileşim içindekileri koruyan bir aday öznitelik alt kümesi elde eder. • Makalede önerilen ölçüt kullanılarak aday alt kümedeki öznitelikler değerlendirilir. • Bu ölçüte göre, ilgisiz öznitelikleri belirleyip çıkararak sonuç öznitelik alt kümesini oluşturur.
PAC-Bayes Shah ve diğ. (2011)	• Eğitim veri kümesi için her bir öznitelik üzerinde tanımlanan karar kütükleri oluşturulur. • Her bir karar kütüğü tek bir öznitelik için tanımlanmış eşik sınıflandırıcıdır. • PAC-Bayes yaklaşımı, karar kütüklerinin birleşiminin öğrenilmesi için kullanılır.

Filtre, sarmal ve gömülü yöntemlerin birkaçının birlikte kullanılmasıyla hibrit yöntemler oluşturulmaktadır. Literatürde MRMR ve DVM-RFE yöntemini birlikte kullanan (Mundra ve Rajapakse, 2010), Relief-F, MRMR ve GA sarmal yöntemlerini birleştiren (Shreem ve diğ., 2012), çoklu filtre ve çoklu sarmal yöntemleri kullanan (Leung ve Hung, 2010), dinamik parametreli GA ve Ki-kare testini birleştiren (Lee ve Leu, 2011), korelasyon tabanlı filtreleme ve Taguchi GA'yı birleştiren (Chuang ve diğ., 2011), Relief-F filtre yöntemiyle ön işlem sonrası ELM ile sınıflandırma gerçekleştiren (Khadijah ve diğ., 2018), DVM-RFE yöntemini F-test ile geliştiren (Luo ve diğ., 2019) çalışmalar gibi çok sayıda hibrit yöntem önerilmiştir. Ayrıca, filtre yöntemleri kullanmadan farklı optimizasyon algoritmalarını, arama algoritmalarını, Evrimsel Algoritmaları (EA) birleştiren çeşitli hibrit yöntemler de bulunmaktadır. Örneğin Geometrik PSO ve Geri Beslemeli YSA'yı birleştiren (GPSO-GBYSA) (Leenavinmalar ve

Kumarkombaiya, 2016), GA ve ACO yöntemlerini Tabu arama algoritmasıyla hibritleştiren (Meenachi ve Ramakrishnan, 2021), elastik ağı DVM ile belirlemek için farklı optimizasyon algoritmalarını kullanan (Qaraad ve diğ., 2021), Çok Amaçlı Guguk Kuşu Arama Algoritmasını (Multi-Objective Cuckoo Search Algorithm-MOCS) mutasyon ve çaprazlama evrimsel operatörleriyle birleştiren (Othman ve diğ., 2020) çalışmalar bu hibrit yöntemlerdendir.

Huerta ve diğ. (2006) çalışmalarında, bulanık mantık tabanlı filtreleme yöntemini ön işlem olarak kullanmış ve GA-DVM ile gen seçimi gerçekleştirmişlerdir. Elde ettikleri sonuçlar karşılaştırdıkları diğer yöntemlerle benzer seviyededir.

Lee (2008) çalışmasında regresyon analizi, GA, PSO, varyans analizi ve DVM yöntemlerini birleştirdiği bir hibrit yöntem önermiştir. Bu yöntemin, kolon kanseri veri kümesi için GA-DVM ve PSO-DVM yöntemlerine göre daha başarılı olduğunu göstermiştir.

Tong ve Schierz (2011), GA ve İleri Beslemeli Yapay Sinir Ağı (İBYSA) birleştirerek hibrit bir yöntem ileri sürmüşlerdir. Lösemi ve SRBCT veri kümeleri için diğer yöntemlere göre en yüksek doğruluk oranı sonuçlarını elde etmişlerdir.

Gunavathi ve Premalatha (2014), T-test, SNR ve F-test filtre yöntemleriyle GA-DVM veya GA-KNN yöntemlerini birleştirerek oluşturdukları hibrit yöntemlerin mikrodizi veri kümelerini sınıflandırma performanslarını karşılaştırmışlardır. En başarılı sonuçların, SNR ve GA-DVM birleşimi hibrit yöntemle elde edildiğini göstermişlerdir.

Nikumbh ve Jayaraman (2012), Biyocoğrafya Tabanlı Optimizasyon (Biogeography Based Optimization-BBO) yöntemiyle RF ve DVM sınıflandırma yöntemlerini birleştirdikleri BBO-RF ve BBO-DVM yöntemlerini önermişlerdir. Çalışmalarında öncelikle genleri bilgi kazancına dayanan derecelendirme yöntemiyle ön işlemden geçirmişler. Daha sonra bu genleri kullanarak BBO-RF ve BBO-DVM yöntemlerinin başarılarını karşılaştırmışlardır. Sonuç olarak, IG + BBO-DVM hibrit yönteminin daha başarılı olduğunu göstermişlerdir.

Tabakhi ve diğ. (2015), gen seçimi için filtre yöntemin uygunluk fonksiyonu olarak kullanıldığı yeni bir ACO yöntemi önermişlerdir. Bu yöntemin başarısını Laplas skoru, karşılıklı korelasyon, MRMR gibi filtre yöntemlerle karşılaştırmışlardır. MRMR ile benzer, diğerlerinden daha başarılı sonuçlar elde edilmiştir.

Jain ve diğ. (2018), korelasyon tabanlı öznitelik seçimi, geliştirilmiş ikili PSO ve NB yöntemlerini birleştirdikleri hibrit bir yöntem ileri sürmüşlerdir. Bu yöntemle PSO-KA, DVM, RF gibi yöntemlere göre daha başarılı sonuçlar elde edildiğini göstermişlerdir.

Sahu ve Mishra (2012), k-ortalamalar kümeleme yöntemi, SNR, PSO ile DVM, KNN, Olasılıksal Sinir Ağları (OSA) gibi sınıflandırma yöntemlerini birleştirdikleri hibrit bir yöntem önermişlerdir. Önerilen yöntemde, öncelikle k-ortalamalar kümeleme algoritmasıyla gen alt kümesi oluşturulmaktadır. Bu alt kümedeki genler, SNR yöntemiyle derecelendirilmektedir. Belirli bir eşik değerin üstünde kalan genler seçilerek PSO için girdi olarak kullanılmaktadır. PSO ile optimum alt küme belirlenmektedir. Bu alt kümedeki genler kullanılarak KNN, DVM ve OSA yöntemleriyle sınıflandırma gerçekleştirilmektedir. Sonuç olarak, en başarılı sonuçlar DVM'nin kullanıldığı hibrit yöntemle elde edilmiştir. Ayrıca, PSO kullanılmadan yapılan sınıflandırma işlemlerinin başarılarının daha düşük olduğu gösterilmiştir.

Lu ve diğ. (2017), Karşılıklı Bilgiyi En Büyütme (Mutual Information Maximization-MIM) yöntemiyle AGA yöntemini birleştirmiş ve DVM, ELM, İBYSA sınıflandırma yöntemleri için başarılarını karşılaştırmışlardır. Ayrıca, Relief-F + ELM, MIM + ELM, SFS + ELM hibrit yöntemlerinin başarılarını da incelemişlerdir.

Alshamlan ve diğ. (2015), gen seçimi için Yapay Arı Kolonisi (Artificial Bee Colony-ABC) ve genetik algoritmayı birleştirdikleri Genetik Arı Kolonisi (Genetic Bee Colony-GBC) algoritmasını önermişlerdir. Önerilen yöntemin başarısını MRMR + ABC, MRMR + GA, MRMR + PSO algoritmalarıyla karşılaştırmışlardır. GBC yönteminin gen sayısını azaltma ve doğruluk oranını arttırma açısından daha üstün olduğunu göstermişlerdir.

3. MALZEME VE YÖNTEM

3.1. ÖZNİTELİK SEÇİMİ YÖNTEMLERİ

Öznitelik seçimi, gereksiz (tekrarlı) ve ilgisiz özniteliklerin çıkarılması, veriyi en iyi yansıtan ve doğruluk oranı gibi ölçütleri arttıran öznitelik alt kümesinin seçilmesi işlemidir. Öznitelik seçme yöntemleri, performans ölçütlerinin iyileştirilmesine, hesaplama zamanının azaltılmasına ve verinin daha iyi anlaşılmasına yardımcı olmaktadır. Özellikle, yüksek boyutlu veri kümelerinde sınıflandırma veya kümeleme işlemini gerçekleştirmeden önce uygulanmaktadırlar. DNA mikrodizisi gibi öznitelik (gen) sayısı 2000-60000, örnek sayısı 40-100 olan veri kümelerinde aşırı öğrenmeyi engelleme ve modelin genelleme yeteneğini arttırarak sınıflandırma başarısını iyileştirmek için öznitelik seçme yöntemleri kullanılmaktadır. Seçilecek ideal alt kümedeki genlerin sınıfla ilgisi yüksek, birbirleriyle ilgisi düşük olmalıdır. Amaç, sınıfı belirlemedeki en etkin, kendi içinde en az tekrarlı gen alt kümesini seçmektir. Öznitelik seçme yöntemleri; filtre yöntemler, sarmal yöntemler, gömülü ve hibrit yöntemler olmak üzere 4 kategoride incelenmektedir.

3.1.1. Filtre Yöntemler

Filtre yöntemler, sınıflandırma işleminden önce uygulanan, öğrenme içermeyen, verinin istatistiksel değerlendirmesini içeren yöntemlerdir. Genellikle özniteliklerin (genlerin) sınıflarla olan ilişkilerinin derecelendirilmesiyle alt küme seçme işlemi gerçekleştirilir. Sınıfla ilgili genler belirlenmeye çalışılır. Sonuca hızlı ulaşılan, uygulaması pratik olan yöntemlerdir. Buna karşın, genlerin birbirleriyle olan ilişkisi çoğunlukla dikkate alınmaz. Benzer, gereksiz genler elenmeyebilir. Birlikte daha çok anlam ifade eden genler kaybedilebilir.

Filtre yöntemler, tek değişkenli ve çok değişkenli olmak üzere ikiye ayrılır. Tek değişkenli yöntemler her özniteliği diğerlerinden bağımsız olarak, çok değişkenli yöntemler öznitelikleri diğerleriyle ilişkisi ve bağlantısıyla birlikte değerlendirir.

Başlıca filtre yöntemler; Pearson Korelasyonu gibi korelasyon tabanlı yöntemler, Karşılıklı Bilgi (Lu ve diğ., 2017; Vanitha ve diğ., 2015; Liu ve diğ., 2010; Liu ve diğ., 2005), IG, MRMR (Ding ve Peng, 2005; Babu ve Sarkar, 2017; Passi ve diğ., 2017), Relief-F (Kononenko, 1994), Fisher skor, istatistiksel testler olan F-test ve T-test olarak sıralanabilir.

3.1.1.1. Pearson Korelasyonu

Pearson Korelasyonu ile öznitelik (gen) ve hedef değer (sınıf) arasındaki doğrusal bağımlılığın derecesi ölçülmektedir. Bu yöntemle, genlerin sınıfla olan ilişkisinin derecelendirilmesi sağlanmaktadır. Böylece, derecesi düşük olan genler elenerek sınıflandırmada kullanılacak genler azaltılabilmektedir. Pearson Korelasyonu, Denklem 3.1'deki gibi hesaplanmaktadır. X^i i. özniteliği, Y hedef değerleri, N örnek sayısını göstermektedir.

$$r = \frac{kov(X^i,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{j=1}^{N}(x_j-\overline{x})(y_j-\overline{y})}{\sqrt{\sum_{j=1}^{N}(x_j-\overline{x})^2(y_j-\overline{y})^2}} \tag{3.1}$$

Ayrıca, Pearson Korelasyonu kullanılarak bir genin diğer genlerle arasındaki ilişkinin derecesi de hesaplanarak benzer genler belirlenebilir.

3.1.1.2. Karşılıklı Bilgi

Etkin genlerin belirlenebilmesi için genlerin ilgisizlik ve gereksizlik derecelerinin her ikisinin de ölçülmesi gerekmektedir. Karşılıklı Bilgi (Mutual Information-MI), bu iki ölçümün gerçekleştirilebilmesi için kullanılan yöntemlerden biridir.

Shannon ve Weaver (1949) entropi tanımına göre bir rastgele değişkenin entropisi Denklem 3.2'de gösterildiği gibi hesaplanmaktadır. Bu denklemde, X değişkeninin olasılık yoğunluğu $P(x_i)$ ile gösterilmektedir. Entropi rastgele değişkenin belirsizliğini, düzensizliğini ölçer. H(X), yaklaşık olarak rastgele değişken X'in bir örneğinden ortalama ne kadar bilgi edinileceğine karşılık gelir. Entropi 0-1 aralığında değer alır. Örnekler sınıflar arasında eşit dağılıyorsa 1 değerini, tek sınıfta toplanıyorsa 0 değerini alır. 0-1 aralığındaki değerler, düzensizliğin, belirsizliğin derecesini gösterir, 1'e yaklaştıkça düzensizlik artar.

$$H(X) = -\sum_i P(x_i)\, log P(x_i) \tag{3.2}$$

İki rastgele değişkenin birbirine bağımlılığını ölçmek için MI kullanılmaktadır. X ve Y rastgele değişkenler olmak üzere Karşılıklı Bilgi I(X,Y) Denklem 3.3 ve 3.5'deki gibi hesaplanır. X değişkeninin entropisi H(X), Y değişkeninin entropisi H(Y), X ve Y değişkenlerinin birleşik entropisi H(X,Y) ile gösterilmektedir.

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.3}$$

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

$$I(X,Y) = \sum_{y\in Y} \sum_{x\in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{3.5}$$

X ve Y bağımsız değişkenlerse, MI 0 değerini alır. Birbirine bağımlı değişkenlerse, 0'dan büyük bir değer alır. Eğer değişkenler birbirine bağımlıysa, birinden elde edilecek bilgi diğerini belirlemede kullanılabilir. Bu değişkenlerin genler olduğu düşünüldüğünde, birbirine bağımlılık derecesi yüksek olan, tekrarlı bilgi içeren genler bu yöntemle belirlenebilir. Birbiriyle en çok ilişkili genlerden, sınıfla ilgisi en yüksek olan dışındakilerin öznitelik alt kümesinden çıkarılması, başarı oranını fazla etkilemezken gen sayısının azaltılmasını sağlamaktadır.

Gen ifade seviyelerini içeren veri kümeleri için sınıfla genler arasındaki toplam Karşılıklı Bilginin en yüksek olduğu, genler arasındaki Karşılıklı Bilginin en düşük olduğu gen alt kümelerinin belirlenmesi amaçlanmaktadır.

3.1.1.3. Minimum Tekrar Maksimum İlgi

Ding ve Peng (2005) çalışmalarında genler arasındaki tekrarın (ilişkinin) en az olduğu, genlerle hedef değer arasındaki ilginin en fazla olduğu altkümenin belirlenmesine yönelik bir yöntem önermişlerdir. Bu yöntem, Minimum Tekrar Maksimum İlgi (Minimum Redundancy Maximum Relevance-MRMR) olarak adlandırılmaktadır. Öznitelik alt kümesindeki genlerin birbirinden bağımsız veya birbiriyle en az ilişkili genlerden seçilmesi, hedef değeri (sınıfı) benzer özelliklerle etkileyen değişkenlerin alt kümede en az sayıda olmasını sağlamaktadır. Bu yaklaşımla, alt kümedeki genlerin arasındaki tekrarın en aza indirilmesi sağlandığından daha verimli bir alt küme elde edilebilmektedir. Ayrıca, hedefi daha farklı açılardan etkileyen genler alt kümeye dahil edilerek alt kümenin genelleme yeteneği arttırılabilmektedir.

MRMR yönteminde, ayrık (kategorik) değişkenler için hesaplamalarda MI kullanılmaktadır. Minimum tekrar hesaplamasında temel düşünce, Karşılıklı Bilgi değeri en az olan genleri seçmektir. Buna göre A öznitelik altkümesini, |A| altkümenin eleman sayısını, $I(g_i, g_j)$ i. ve j. genler için Karşılıklı Bilgiyi göstermek üzere minimum tekrar koşulu Denklem 3.6 ile elde edilmektedir.

$$\min(T),\ T = \frac{1}{|A|^2}\sum_{g_i,g_j \in A} I(g_i, g_j) \tag{3.6}$$

Maksimum ilgi koşulu; s sınıfı, $I(s, g_i)$ i. genle sınıf arasındaki ilgiyi göstermek üzere Denklem 3.7'deki gibi alt kümedeki tüm genlerin toplam ilgisinin en büyütülmesiyle hesaplanmaktadır.

$$\text{maks}(L),\ L = \frac{1}{|A|}\sum_{g_i \in A} I(s, g_i), \qquad (s = s_1, s_2, \ldots . s_K) \tag{3.7}$$

MRMR öznitelik alt kümesinin belirlenmesinde Denklem 3.6 ve 3.7'nin eş zamanlı optimizasyonunu sağlamak için bu iki koşul Denklem 3.8 ve 3.9'daki şekliyle birlikte değerlendirilmektedir. Denklem 3.8 farkın, Denklem 3.9 oranın optimizasyonunu göstermekte ve sırayla Karşılıklı Bilgi Farkı (KBF) ve Karşılıklı Bilgi Oranı (KBO) kriterleri olarak adlandırılmaktadır.

$$\text{maks}\ (L - T) \tag{3.8}$$

$$\text{maks}\ (L/\ T) \tag{3.9}$$

MRMR alt küme seçme algoritmasına göre ilk olarak Denklem 3.7'ye göre $I(s, g_i)$ değeri en yüksek olan gen alt kümeye eklenmektedir. Geri kalan genlerden alt kümeye eklenecek olanı bulmak için $I(s, g_i)$ değeri en yüksek olan ve Denklem 3.10'da gösterildiği gibi alt kümedeki genlerle ilgisi en az olan gen belirlenmektedir. Y geriye kalan genleri göstermektedir.

$$min_{g_i \in Y}\ \ \frac{1}{|A|}\sum_{g_j \in A} I(g_i, g_j) \tag{3.10}$$

KBF ve KBO kriterleri sırasıyla Denklem 3.11 ve 3.12'deki gibi hesaplanmaktadır. Bu iki değer, alt kümeye eklenecek genleri belirlemedeki ölçütlerdir.

$$maks_{g_i \in Y}\left[I(s, g_i) - \frac{1}{|A|}\sum_{g_j \in A} I(g_i, g_j)\right] \tag{3.11}$$

$$maks_{g_i \in Y}\left[I(s, g_i) / \left(\frac{1}{|A|}\sum_{g_j \in A} I(g_i, g_j)\right)\right] \tag{3.12}$$

MRMR yönteminin sürekli değişkenler için hesaplanmasında sınıfla genler arası maksimum ilgiyi belirlemede F-test değeri, genler arasındaki benzerliği belirlemede Öklid uzaklığı veya Pearson korelasyon katsayısı kullanılmaktadır. Gen ve sınıf arasındaki F-test değeri Denklem 3.13'deki gibi tanımlanmaktadır (Ding ve Peng, 2005; Ding, 2002).

$$F(s, g_i) = [\sum_k n_k(\bar{g}_k - \bar{g})^2/(K-1)]/\sigma^2 \quad (3.13)$$

$$\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2]/(n-K) \quad (3.14)$$

Denklem 3.13 ve 3.14'de K sınıf sayısını, n örnek sayısını, n_k k. sınıftaki örnek sayısını, $\bar{g}$ i. genin tüm örnekler için ortalama gen ifadesini, $\bar{g}_k$ i. genin k. sınıf için ortalama gen ifadesini, σ^2 bileşik varyans değerini, σ_k^2 k. sınıf için varyans değerini göstermektedir.

Sürekli değişkenler için MRMR yöntemi seçim kriterleri F-test Korelasyon Farkı (FKF) ve F-test Korelasyon Oranı (FKO) sırasıyla Denklem 3.15 ve 3.16'daki gibi hesaplanmaktadır. $F(s, g_i)$ s sınıfı ve i. gen için F-test değerine, $r(g_i, g_j)$ i. ve j. genler için Pearson Korelasyon Katsayısına karşılık gelmektedir.

$$maks_{g_i \in Y}\left[F(s, g_i) - \frac{1}{|A|}\sum_{g_j \in A} r(g_i, g_j)\right] \quad (3.15)$$

$$maks_{g_i \in Y}\left[F(s, g_i)/\left(\frac{1}{|A|}\sum_{g_j \in A} r(g_i, g_j)\right)\right] \quad (3.16)$$

3.1.1.4. Bilgi Kazancı

Bilgi kazancı (IG), entropiye dayalı bir ölçüt olarak entropideki değişimi, azalmayı gösterir. Bir özniteliğin veri kümesine sağladığı bilginin ölçümüne karşılık gelmektedir. Öznitelik seçimi için IG hesaplamasının kullanılmasındaki temel hedef, sınıflar hakkında en çok bilgiyi ortaya çıkaran, ayırt edici öznitelikleri (genleri) seçmektir. Öznitelik seçimi için her özniteliğin bilgi kazancı ayrı ayrı hesaplanır, belirli bir eşik değeri geçen öznitelikler seçilir. S veri kümesini, X özniteliği, $H(S)$ veri kümesinin toplam entropisini göstermek üzere IG Denklem 3.17'deki gibi tanımlanmaktadır.

$$G(S, X) = H(S) - H_X(S_i) \quad (3.17)$$

$$H(S) = -\sum_{i=1}^{k} P(C_i, S) log P(C_i, S) \quad (3.18)$$

$$H_X(S_i) = -\sum_{i=1}^{v} \frac{|S_i|}{|S|} H(S_i) \quad (3.19)$$

3.1.1.5. Relief-F

Relief-F (Kononenko, 1994), farklı sınıfları en iyi ayırt eden öznitelikleri seçmeyi amaçlayan çok değişkenli bir filtre öznitelik seçme yöntemidir. Relief algoritmasının çok sınıflı veri kümeleri için geliştirilmiş, gürültülü ve eksik verilerde daha iyi sonuç veren bir versiyonudur (Rismiyati ve Mantau, 2018). Bu yöntemde temel amaç, en yakın örnekleri ayırt etme yeteneğine bağlı olarak her bir özniteliği derecelendirmektir. Algoritma, rastgele bir x örneği seçer ve x ile aynı sınıftan en yakın k örneğini (a) ve farklı sınıftan en yakın k örneğini (b) belirler. Seçilen rastgele x örneği ile aynı sınıftan en yakın k örneği arasındaki uzaklık ve x ile farklı sınıftan en yakın k örneği arasındaki uzaklık hesaplanır. Buna göre, x örneğinin aynı sınıftan örneklerle uzaklığı arttıkça özniteliğin derecesi azalır, farklı sınıftan örneklerle uzaklığı arttıkça derecesi artar.

Relief-F yöntemine göre f özniteliğinin derecesi $R(f)$, Denklem 3.20'de gösterildiği gibi tanımlanmaktadır.

$$R(f) = R(f) - R(f)_a + R(f)_b \tag{3.20}$$

Denklem 3.21'deki $R(f)_a$, f özniteliğinden seçilen rastgele x örneği için aynı sınıftan en yakın komşularıyla arasındaki uzaklıkların toplamının ortalamasını göstermektedir. Rastgele örnek seçme işlemi n defa tekrarlanmaktadır.

$$R(f)_a = \frac{1}{n.k}\sum_{i=1}^{k} diff(f, x, a_i) \tag{3.21}$$

Denklem 3.22'deki $R(f)_b$, farklı sınıftan en yakın komşularıyla arasındaki uzaklıkların toplamının ortalamasına karşılık gelmekte, x örneğinin dışında kalan tüm sınıflar (s) için birlikte hesaplanmaktadır. $P(s_x)$, x örneğinin sınıfının olasılığını, $P(s)$ s sınıfının olasılığını göstermektedir.

$$R(f)_b = \sum_{s \neq s_x} \frac{P(s)}{[1-P(s_x)]n.k}\sum_{i=1}^{k} diff(f, x, b_i(s)) \tag{3.22}$$

3.1.1.6. Diğer Filtre Yöntemler

Gen ifade seviyelerini içeren veri kümelerinde gen seçimi için kullanılan diğer filtre yöntemler; SNR, T-testi, F-testi, Fisher skor ve Ki-kare olarak sıralanabilir.

İki sınıflı veri kümesi için SNR, Denklem 3.23'deki gibi hesaplanır. μ_1 ve μ_2 sırasıyla 1. ve 2. sınıftaki karşılık gelen özniteliğe ait değerlerin ortalamasını, σ_1 ve ve σ_2 standart sapmalarını göstermektedir.

$$SNR = \frac{\mu_1 - \mu_2}{\sigma_1 - \sigma_2} \tag{3.23}$$

T-testi, bir özniteliğin iki sınıf arasındaki farkının istatistiksel önemini ölçmektedir. T-testi değeri, Denklem 3.24'deki gibi hesaplanmaktadır. 1. ve 2. sınıf için örnek sayıları sırasıyla n_1 ve n_2'dir. Değerlendirilen özniteliğin karşılık gelen sınıf için varyansları σ_1^2 ve σ_2^2 olarak gösterilmektedir.

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{3.24}$$

F-testi değeri her bir özniteliğin iki sınıf için varyanslarının oranı olarak Denklem 3.25'deki gibi hesaplanır.

$$F = \frac{\sigma_1^2}{\sigma_2^2} \tag{3.25}$$

Fischer skor (FS), Denklem 3.26'daki gibi hesaplanır. μ_i i. sınıfa ait ortalama değeri, μ tüm veri kümesinin ortalamasını, n_i i. sınıfın eleman sayısını, σ_i^2 i. sınıfın varyansını göstermektedir.

$$FS = \frac{\sum_{i=1}^{s} n_i(\mu_i - \mu)^2}{\sum_{i=1}^{s} n_i \sigma_i^2} \tag{3.26}$$

Ki-kare değeri, Denklem 3.27'deki gibi hesaplanır. Bu denklemde, obs_i değerlendirilen genin i. örneği için gözlemlenen değerini, exp_i beklenen değerini göstermektedir.

$$X^2 = \sum_{i=1}^{n} \frac{(obs_i - exp_i)^2}{exp_i} \tag{3.27}$$

3.1.2. Sarmal Yöntemler

Sarmal yöntemler, sınıflandırma yöntemi ve arama veya optimizasyon algoritmasından oluşmaktadır. Arama algoritmasının seçtiği özniteliklerin sınıflandırma başarıları değerlendirilerek en başarılı olan alt küme belirlenmektedir. Bu seçim ileri yönde (tek elemanlı kümeden başlayıp ekleyerek) veya geri yönde (elemanları çıkararak) gerçekleştirilebilir. Bu

yöntemde hedef, seçilen alt kümeler için sınıflandırıcıdan elde edilen başarının veya buna bağlı bir ölçütün optimizasyonudur.

N sayıda özniteliği bulunan yüksek boyutlu bir veri kümesi için (2^N-1) alt kümenin değerlendirilmesi oldukça fazla hesaplama yükü getireceğinden alt küme seçiminde sezgisel veya evrimsel algoritmalar kullanılmaktadır. Bu amaçla, en sık kullanılan algoritmalar genetik algoritma ve parçacık sürü optimizasyonudur.

Sarmal yöntemlerle, genlerin öngörü yeteneği araştırılırken genler arası ilişkiler, bağımlılıklar da dikkate alınmaktadır. Hesaplama yükü filtre yöntemlere göre daha fazla olmasına karşın verinin içeriğini koruma ve sınıflandırma başarıları yüksektir (Salema ve diğ., 2017; Chandrashekar ve Sahin, 2014). Ancak birbirine benzeyen genleri elemek için yeterli ölçütleri içermemektedir.

3.1.3. Gömülü Yöntemler

Gömülü yöntemler, öznitelikleri derecelendirmek için bir kriter oluşturmak üzere sınıflandırıcıyı kullanan yöntemlerdir. Öznitelik seçimi sınıflandırıcı eğitiminin bir parçası olarak gerçekleştirildiği için sonuçlar sınıflandırıcıya bağımlıdır. Gömülü yöntemlerin hesaplama yükü sarmal yöntemlere göre daha az olmasına karşın sınıflandırıcıya bağımlı seçilmiş öznitelikler farklı sınıflandırıcılarda aynı başarıyı gösteremeyebilir (Hira ve Gillies, 2015).

3.1.4. Hibrit Yöntemler

Filtre, sarmal ve gömülü yöntemlerin birkaçının birlikte kullanılmasıyla hibrit yöntemler oluşturulmaktadır. Filtre ve sarmal yöntemlerin birlikte kullanıldığı çoğu hibrit yöntemde filtre yöntemler genellikle ön işlem olarak kullanılmaktadır. Amaç, veri kümesindeki öznitelik sayısını filtre yöntemle 500-1000 seviyesine indirmektir. Sonrasında sarmal yöntemle öznitelik seçimi gerçekleştirilmektedir. Böylece, daha yüksek doğruluk oranı ve sonuca hızlı ulaşılması hedeflenmektedir.

3.2. BOYUT İNDİRGEME YÖNTEMLERİ

3.2.1. Temel Bileşen Analizi

Temel bileşen analizi (TBA), doğrusal dönüşümlerle varyansı maksimize eden yeni ilişkisiz değişkenler oluşturarak boyut indirgemeyi gerçekleştirmektedir (Jolliffe ve Caligma, 2016).

TBA, veri kümesinin kovaryans matrisinin hesaplanması, bu matrisin özvektörleri ve özdeğerleri bulunarak temel bileşenlerin belirlenmesi temeline dayanan bir yöntemdir. N elemanlı A veri kümesinin kovaryans matrisi M, $x_i \in$ A, μ veri kümesinin ortalaması olmak üzere Denklem 3.28'deki gibi tanımlanmaktadır.

$$M = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)\,(x_i - \mu)^T \tag{3.28}$$

M matrisinin özvektör ve özdeğerleri Denklem 3.29 ile belirlenmektedir. Varyansı en büyütmek için en büyük özdeğere karşılık gelen özvektörler seçilmektedir. M matrisinin k. en büyük özdeğeri λ_k ve buna karşılık gelen özvektörü v_k olmak üzere Xv_k veri kümesinin temel bileşenleri olarak adlandırılmaktadır.

$$Mv = \lambda v \tag{3.29}$$

M matrisinin k adet en büyük özdeğerine karşılık gelen özvektörlerinden Z dönüşüm matrisi oluşturulmaktadır. Denklem 3.30'daki dönüşüm ile W matrisi elde edilmektedir.

$$W = Z^T X \tag{3.30}$$

3.2.2. Lineer Diskriminant Analizi

Lineer diskriminant analiz (LDA), veriyi sınıfların daha iyi ayrıldığı bir boyuta taşıyarak boyut indirgemektedir (Sharma ve Paliwal, 2015). İki sınıflı, N elemanlı, p değişkenli bir veri kümesi X matrisiyle temsil edildiğinde bu veri kümesinin C_k sınıfı için ortalaması μ_k Denklem 3.31'deki gibi, tüm sınıflar için ortalaması μ Denklem 3.32'deki gibi hesaplanmaktadır.

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} x_i \qquad x_i \in C_k \tag{3.31}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.32}$$

Veriyi tanımlayan sınıflar arası saçılım matrisi S_A ve sınıf içi saçılım matrisi S_W sırasıyla Denklem 3.33 ve Denklem 3.34'deki gibi hesaplanmaktadır.

$$S_A = \sum_k N_k(\mu_k - \mu)\,(\mu_k - \mu)^T \tag{3.33}$$

$$S_W = \sum_k \sum_{x_i \in C_k} (x_i - \mu_k)\,(x_i - \mu_k)^T \tag{3.34}$$

$$J(w) = \frac{w^T S_A w}{w^T S_W w} \tag{3.35}$$

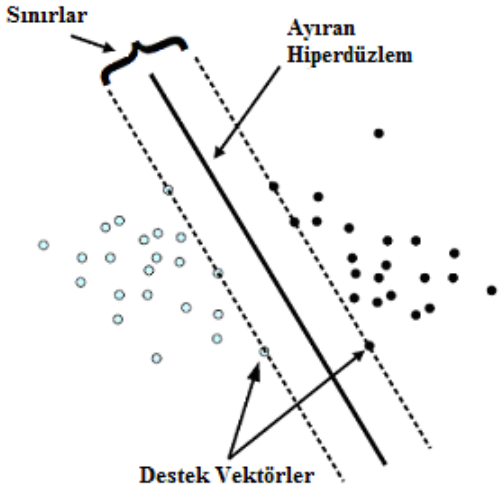
LDA, Denklem 3.35'daki $J(w)$ kriterini en büyüten w vektörünün bulunmasını amaçlamaktadır. LDA, bu vektör ile veriyi sınıfların daha iyi ayrıldığı bir boyuta taşıyarak boyut indirgeme ve sınıflandırma işlevini gerçekleştirmektedir.

3.3. SINIFLANDIRMA YÖNTEMLERİ

3.3.1. Destek vektör makinesi

Destek vektör makinesi (DVM), sınıflandırma ve regresyon işlemleri için kullanılan danışmanlı bir öğrenme yöntemidir. Bu yöntem, birçok sınıflandırma uygulamasında başarılı sonuçlar vermektedir. DVM, ikili sınıflandırma veya çok sınıflı sınıflandırma işlemleri için kullanılabilir. Bu yöntemde amaç, veriyi bir hiper düzlem yardımıyla sınıflara ayırmaktır. Bunun için farklı sınıfların destek vektörleri arasındaki uzaklığı en büyüten ayırma hiper düzleminin bulunması amaçlanmaktadır (Cortes ve Vapnik, 1995). DVM, doğrusal ve doğrusal olmayan DVM olarak ikiye ayrılır. Doğrusal DVM, verinin doğrusal olarak ayrılabildiği durumlarda, veriyi en iyi ayıran doğrusal hiper düzlemin bulunması işlemidir. Doğrusal olmayan destek vektör makinelerinde, veriler çekirdek fonksiyonları kullanılarak verilerin doğrusal olarak ayrılabildiği yüksek boyutlu bir uzaya taşınırlar.

Şekil 3.1'de iki sınıflı verinin doğrusal DVM ile sınıflandırılması örneği için sınırlar, ayıran hiper düzlem ve destek vektörler gösterilmektedir.



Şekil 3.1: İki sınıflı veri için doğrusal DVM örneği (Burges, 1998).

3.3.2. K-en yakın komşu

K en yakın komşu (KNN) sınıflandırma yöntemi, sınıfı bilinmeyen örneklerin uzaklık veya benzerlik ölçüsüne göre bilinen örneklerle ilişkilendirilmesiyle tanımlanabileceği varsayımına dayanır. Aynı örnek uzayında konumlandırılan örneklerden birbirine çok uzak olan örneklerin birbirine yakın örneklere göre aynı sınıfa ait olma olasılığı daha azdır. KNN ile sınıflandırma temel olarak, örneklerin aynı uzayda konumlandırılması ve sınıfı bilinmeyen örneğin en yakın k komşusunun çoğunluğunun ait olduğu sınıfa atanması ile gerçekleştirilir. Uzaklık hesaplamalarında genellikle Öklid uzaklığı kullanılmaktadır. Manhattan, Minkowski, Hamming uzaklıkları da kullanılan diğer ölçütlerdir. KNN sınıflandırıcının yerel konumları kullanan yaklaşımı, sınıflandırıcının eğitim kümesindeki gürültüye oldukça duyarlı olmasına neden olur. K değeri büyütüldükçe bu duyarlılık azalır.

3.3.3. Naive Bayes

Naive bayes (NB), bayes teoremine dayanarak sınıflandırma ve kestirim işlemini gerçekleştiren yöntemdir. Bu yöntemde, bütün özniteliklerin birbirinden koşullu bağımsız olduğu varsayılmaktadır. Buna göre X (x_1, x_2, ..., x_n) öznitelik vektörünü göstermek üzere, X'in i sınıfına ait olma olasılığı aşağıdaki gibi hesaplanmaktadır.

$$P(S_i|X) = P(X|S_i)P(S_i), \qquad P(X|S_i) = \prod_{k=1}^{n} P(X_k|S_i) \tag{3.36}$$

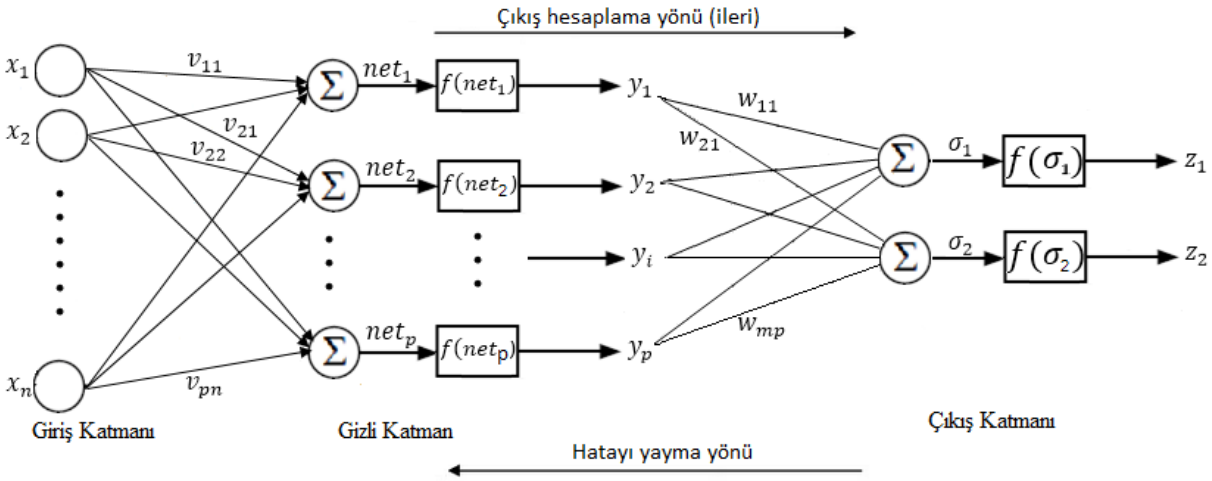
Sürekli veriler içeren öznitelikler için olasılık yoğunluk fonksiyonu kullanılarak sınıfa ait olma olasılığı hesaplanabilmektedir.

3.3.4. İleri Beslemeli Yapay Sinir Ağları (İBYSA)

Yapay sinir ağı (YSA); eksik, gürültülü verileri tolere edebilen, girdileri işleyerek beklenen değerleri elde etmek için eğitilebilen bir sistem olduğu için verilerin sınıflandırmasında sıklıkla kullanılmaktadır. YSA, yapısına göre ileri beslemeli ve geri beslemeli olmak üzere iki sınıfa ayrılır. İleri beslemeli yapay sinir ağları, girişten çıkışa doğru düzgün katmanlar halindedir. Bir katmandaki hücrenin çıkışının sadece bir sonraki katmandaki hücreler için giriş olarak kullanılabildiği yapılardır.

İleri beslemeli ağların eğitiminde çoğunlukla geri yayılım algoritmaları kullanılmaktadır. Bu algoritmalar, ağın ürettiği çıktılar ile beklenen çıktılar arasındaki hatanın en aza indirilmesi temeline dayanır. Öğrenme sürecinde, girdilere uygun çıktılar oluşturulur ve bu çıktılarla beklenen değerler karşılaştırılarak aradaki hata hesaplanır. Bu hata, kullanılan algoritmaya göre farklı yöntemlerle geriye doğru dağıtılır ve ağırlık katsayıları değiştirilir. Bu işlem belirli bir durdurma kriterine kadar devam eder. Bu kriter, giriş değerleri için elde edilen hata değerinin kabul edilebilir hata değerinden düşük olması veya belirli bir iterasyon sayısı (tekrar sayısı) olarak tanımlanabilir.

Gen ifade veri kümeleri için sıklıkla giriş, çıkış ve gizli katmadan oluşan çok katmanlı ileri beslemeli yapay sinir ağları kullanılmaktadır. Giriş katmanında, veri kümesindeki öznitelik sayısına bağlı olarak nöron sayısı değişirken, çıkış katmanında genellikle 2 nöron bulunmaktadır. Çok katmanlı ileri beslemeli YSA yapısı Şekil 3.1'de gösterilmektedir. Buna göre girdi sayısı n, gizli katmandaki nöron sayısı p, çıkış katmanındaki nöron sayısı m olmak üzere, x_1, x_2, ..., x_n giriş değerlerini, y_1, y_2, ..., y_p gizli katmanın çıkış değerlerini, z_1, z_2 çıkış katmanının çıkış değerlerini göstermektedir.



Şekil 3.1: Çok katmanlı ileri beslemeli YSA yapısı.

Gizli katmanın ağırlık katsayıları V ve çıkış katmanına ait ağırlık katsayıları W aşağıdaki gibi ifade edilmektedir.

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pn} \end{bmatrix} \quad W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{m1} & w_{m2} & \cdots & w_{mp} \end{bmatrix} \tag{3.37}$$

Şekil 3.1'deki YSA'da çıkış değerleri ileri yönde hesaplanmaktadır. Buna göre i= 1, 2, 3, ..., p ve j=1, 2, 3, ..., n olmak üzere gizli katmanın i. nöronu için net girdi değeri (net_i) ve çıkış değeri (y_i) aşağıdaki gibi hesaplanmaktadır.

$$net_i = \sum_{j=1}^{n} v_{ij} x_j, \quad y_i = f(net_i) \tag{3.38}$$

Çıkış katmanının değeri benzer şekilde aşağıdaki gibi hesaplanır (i= 1,2 ve j=1, 2, 3, ..., p). Gizli katmanın aktivasyon fonksiyonu olarak genellikle tanjant sigmoid fonksiyonu tercih edilmektedir.

$$\sigma_i = \sum_{j=1}^{p} w_{ij} y_j, \quad z_i = f(\sigma_i) \tag{3.39}$$

3.4. OPTİMİZASYON YÖNTEMLERİ

3.4.1. Genetik Algoritma

Genetik algoritmanın temelleri Holland (1975) tarafından ileri sürülmüştür. Standart genetik algoritma optimizasyon problemlerinin çözümünde kullanılabilecek evrimsel bir hesaplama yaklaşımıdır (Goldberg, 1989). GA, potansiyel çözümleri kromozom benzeri veri yapısıyla kodlar ve anlamlı bilgiyi korumak için çaprazlama ve mutasyon işlemlerini kromozomlara uygular. Bu algoritma, probleme çoklu çözümler üreterek problemin birçok açıdan değerlendirilmesine yardımcı olur.

GA temel bileşenleri ve parametreleri; popülasyonu oluşturan kromozomlar, popülasyon büyüklüğü, uygunluk fonksiyonu, çaprazlama olasılığı, mutasyon olasılığıdır. Standart genetik algoritmanın temel aşamaları aşağıdaki şekildedir:

1. N kromozomlu rastgele bir popülasyon oluşturulur. (Kromozom: potansiyel çözümler)
2. Popülasyondaki her kromozomun uygunluk fonksiyonu hesaplanır.
3. Yeni kromozom popülasyonu oluşana kadar 4-7 aşamaları tekrar edilir.
4. Uygunluk fonksiyonu değerine göre en iyi kromozomlar (ebeveynler) seçilir.
5. Çaprazlama olasılığına göre ebeveyn kromozomlar çaprazlanarak yeni bireyler (offspring) oluşturulur.
6. Mutasyon olasılığına göre yeni bireyler mutasyona uğratılır.
7. Yeni bireyler popülasyona eklenir.
8. Yeni oluşturulan popülasyon sonraki algoritma tekrarları için kullanılır.
9. Durdurma kriterine ulaşıldığında algoritma sonlandırılır. Mevcut popülasyondaki en iyi çözüm sonuç olarak döndürülür.
10. Durdurma kriterine ulaşılmadıysa 2. aşamadan itibaren algoritma tekrarlanır.

Gen ifade seviyelerini içeren veri kümeleri için GA ile öznitelik seçimde, kromozomlar genellikle ikili kodlamayla kodlanarak 0 veya 1 değerini almaktadır. Bu durumda, 1 değeri o özniteliğin (genin) seçildiğini, 0 değeri ise seçilmediğini göstermektedir. Diğer bir yaklaşım, kromozomların [0-1] aralığında değer almasıdır. Bu yöntemde, belirli bir eşik değerin üstünde kalan değerler için öznitelikler seçilmektedir. Kromozomdaki gen sayısı incelenecek öznitelik sayısına karşılık gelmektedir. Genetik algoritmanın uygunluk fonksiyonu olarak sınıflandırıcının performansı kullanılmaktadır. Her kromozomun uygunluk değeri

sınıflandırma yönteminden elde edilen doğruluk oranına göre hesaplanmaktadır. Bu amaçla, DVM ve KNN gibi sınıflandırma yöntemleri GA ile birlikte sıklıkla kullanılan yöntemlerdir.

3.4.2. Parçacık Sürü Optimizasyonu

Parçacık sürü optimizasyonu (PSO), kuş ve balık sürülerinin hareketlerinden esinlenerek Kennedy ve Eberhart (1995) tarafından geliştirilmiş bir optimizasyon yöntemidir. Algoritmanın temeli, bu sürülerin sosyolojik davranışlarına dayanmaktadır. PSO, sürülerin içlerindeki bilgi paylaşımı ile yiyecek bulabilme, hayatta kalma becerilerini geliştirdiği gözlemine dayanarak oluşturulmuştur.

PSO yönteminde her birey bir parçacık olarak tanımlanır, her parçacık potansiyel bir çözümü temsil eder. Her parçacığın pozisyon ve hız vektörü bulunmaktadır. Algoritmada, her parçacığın en iyi pozisyonu (lokal en iyi pozisyon), sürünün en iyi pozisyonu (global en iyi pozisyon), uygunluk fonksiyonu, parçacıkların pozisyon ve hız vektörleri kullanılarak optimum veya optimuma yakın bir çözüm üretilmeye çalışılmaktadır.

Algoritmanın temel aşamaları aşağıdaki gibidir:

1. Rastgele başlangıç değerleri verilerek sürü (parçacıklar: hız ve pozisyon vektörleri) oluşturulur.
2. Her parçacık için uygunluk fonksiyonu değeri hesaplanır.

   Lokal ve global en iyi pozisyon değerleri hesaplanır.

3. Tüm parçacıklar için hız ve pozisyon vektörleri sırasıyla Denklem 3.40 ve 3.41'deki gibi güncellenir.
4. Durdurma kriterine ulaşıncaya kadar 2. ve 3. aşamalar tekrarlanır.

$$V_{i,j}(t+1) = V_{i,j}(t) + c_1 r_1\left(Pbest_{i,j} - X_{i,j}(t)\right) + c_2 r_2\left(gbest_j - X_{i,j}(t)\right) \quad (3.40)$$

$$X_{i,j}(t+1) = X_{i,j}(t) + V_{i,j}(t+1) \quad (3.41)$$

Hız vektörü $V_{i,j} \in \{V_{i,1}, V_{i,2}, \dots, V_{i,n}\}$ i. parçacığın j. boyutu için hızları, pozisyon vektörü $X_{i,j} \in \{X_{i,1}, X_{i,2}, \dots, X_{i,n}\}$ i. parçacığın j. boyutu için pozisyonları, c_1, c_2 ivmelenme sabitlerini, r_1, r_2 0-1 aralığında tekdüze dağılım gösteren rastgele sayıları göstermektedir. Lokal en iyi pozisyonu gösteren $Pbest_{i,j} \in \{Pbest_{i,1}, Pbest_{i,2}, \dots, Pbest_{i,n}\}$ i. parçacığın j. boyutu için o

ana kadarki en iyi pozisyondur ve Denklem 3.42'de tanımlanmaktadır. Global en iyi pozisyonu gösteren $gbest_j \in \{Pbest_1, Pbest_2, \dots, Pbest_n\}$ j. boyut için global en iyi pozisyonu göstermektedir.

$$Pbest_{i,j}(t+1) = \begin{cases} Pbest_{i,j}(t), & f\left(X_{i,j}(t+1)\right) \geq f\left(Pbest_{i,j}(t)\right) \\ X_{i,j}(t+1), & f\left(X_{i,j}(t+1)\right) < f\left(Pbest_{i,j}(t)\right) \end{cases} \tag{3.42}$$

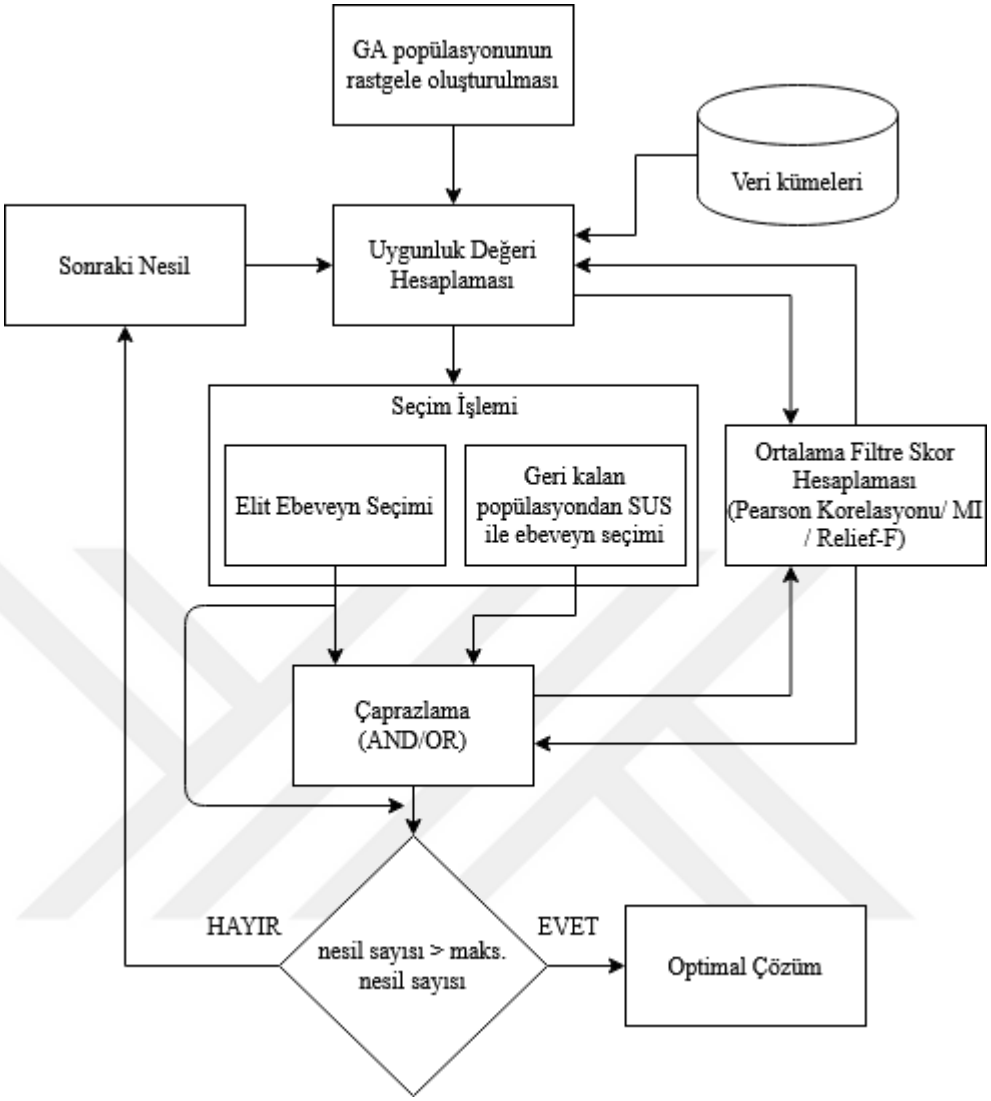
$$gbest_j = \min\{f(Pbest_1(t)), f(Pbest_2(t)), \dots, f(Pbest_n(t))\} \tag{3.43}$$

Öznitelik seçiminde, her parçacık genellikle ikili vektör olarak tanımlanmaktadır. Bu gösterimde GA yöntemindeki kromozomlara benzer olarak vektörün her bir hücresindeki 1 değeri karşılık gelen özniteliğin seçildiğini, 0 değeri seçilmediğini göstermektedir. Vektörün boyutu öznitelik sayısına karşılık gelmektedir. Uygunluk fonksiyonu sınıflandırıcının doğruluk oranı kullanılarak hesaplanmaktadır.

3.5. ÖNERİLEN HİBRİT YÖNTEMLER

Bu tez çalışmasında, GA-KNN, GA-DVM, PSO-KNN ve PSO-DVM sarmal yöntemleri filtre öznitelik seçimi yöntemleriyle geliştirilerek yeni hibrit yöntemler önerilmiştir. Bu hibrit yöntemlerde Pearson Korelasyonu, MI ve Relief-F yöntemleri kullanılmıştır.

GA tabanlı yöntemlerde kromozomlar ikili olarak kodlanmakta, kromozomların boyutu veri kümesinin gen sayısına eşit olacak şekilde belirlenmektedir. Önerilen GA tabanlı yöntemlerin akış şeması Şekil 3.2'de gösterilmiştir. İlk popülasyon rastgele olacak şekilde oluşturularak bireylerin uygunluk fonksiyonu değerleri hesaplanmaktadır. Uygunluk fonksiyonu, sınıflandırıcıdan elde edilen doğruluk oranının, sınıflandırmada kullanılan gen sayısının ve kromozomların filtre yöntem skorunun bir fonksiyonu olarak belirlenmiştir. Amaç, doğruluk oranını ve kromozomların filtre yöntem skorunu en büyütmek, sınıflandırmada kullanılacak gen sayısını en küçültmektir.



Şekil 3.2: GA tabanlı önerilen yöntemlerin akış şeması.

Denklem 3.44'de uygunluk fonksiyonu gösterilmektedir. $A(x)$ x bireyi için sınıflandırıcıdan elde edilen doğruluk oranını, $s(x)$ x bireyinde 1 olan gen sayısını, t toplam gen sayısını, α ve β [0,1] aralığında değer alan katsayıları göstermektedir. $OrtSkor(x)$ x bireyinin karşılık gelen filtre yöntem için ortalama skorudur. Bu değer, Denklem 3.45'deki gibi x bireyindeki genlerin $g(i)$ karşılık geldikleri filtre yöntem skoruyla çarpımlarının toplamının ortalaması olarak hesaplanmaktadır. Böylece, doğruluk oranı ve gen sayısının yanı sıra filtre yöntemler de GA'nın seçim işleminde etkili olmaktadır.

$$F(x) = \beta\left(\alpha\big(1 - A(x)\big) + (1-\alpha)\frac{s(x)}{t}\right) + \ (1-\beta)\big(1 - OrtSkor(x)\big) \tag{3.44}$$

$$OrtSkor(x) = \frac{\sum_{i=1}^{t} g(i)Skor(i)}{s(x)} \quad (3.45)$$

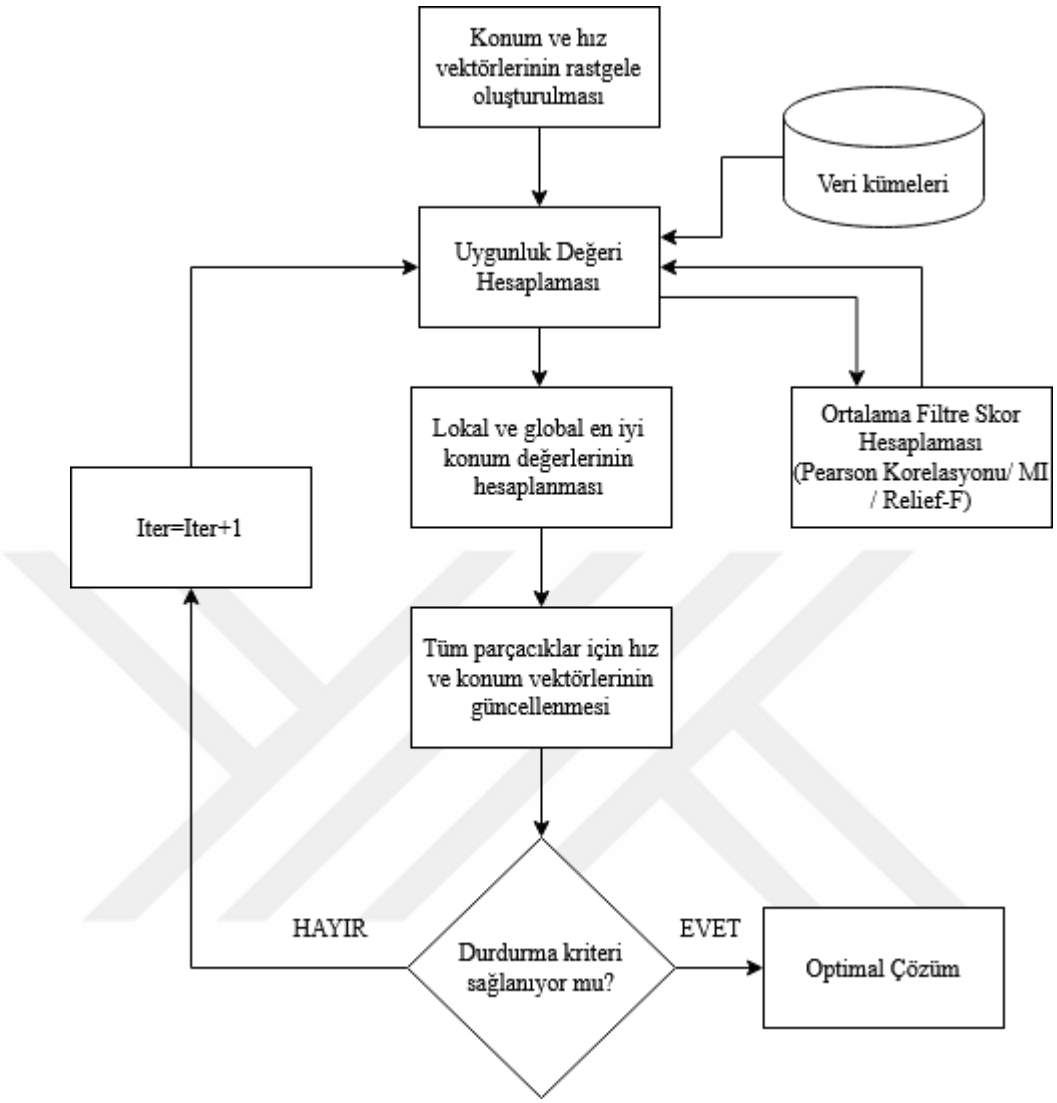
Seçim işleminde, popülasyonun %20'si elit popülasyon olarak seçilmektedir. Elit popülasyon hem ebeveyn olarak seçilmekte hem de bir sonraki nesle aktarılmaktadır. Diğer ebeveynler, uygunluk fonksiyonu değerlerine göre Stokastik Evrensel Seçim (Stochastic Universal Selection-SUS) ile geriye kalan popülasyondan belirlenmektedir.

GA çaprazlama işlemi mantıksal AND ve mantıksal OR işlemleri kullanılarak iki şekilde gerçekleştirilmektedir. AND işlemiyle her iki ebeveynde 1 olan genlerin sonraki nesle aktarılması sağlanmaktadır. OR işlemiyle çeşitlilik sağlanmaktadır. OR işlemiyle çaprazlama işleminden önce, en düşük filtre yöntem skoruna sahip olan belirli bir yüzdedeki genlerden SUS ile seçim yapılarak bu genler her iki ebeveynde de sıfırlanmaktadır. Böylece, sonuca daha hızlı ulaşılmakta ve belirli bir seviyede çeşitlilik sağlanmaktadır. OR işlemiyle çeşitlilik sağlandığı için önerilen yöntemlerde mutasyon işlemi bulunmamaktadır. Önerilen yöntemler için GA parametreleri Tablo 3.1'de gösterilmektedir. Bu değerler, farklı veri kümeleri ve önerilen yöntemler için birçok deneysel sonucun değerlendirilmesi sonrası uygunluk değeri ve doğruluk oranı açısından en iyi sonuçların elde edildiği parametrelerdir.

Tablo 3.1: Önerilen yöntemler için GA parametreleri.

Parametre	**Değeri**
Popülasyon büyüklüğü	100
Çaprazlama	AND (%50), OR (%50)
Nesil sayısı	200
Seçim kuralı	Stokastik Evrensel Seçim
Elit oranı	%20

PSO tabanlı önerilen yöntemlerde öznitelik seçimi için ikili PSO algoritması kullanılmaktadır. PSO'daki her parçacığın boyutu veri kümesinin gen sayısına eşit olacak şekilde rastgele olarak oluşturulmaktadır. Konum vektörünün her bir hücresindeki 1 değeri karşılık gelen özniteliğin seçildiğini, 0 değeri seçilmediğini göstermektedir. Şekil 3.3'de PSO tabanlı yöntemlerin akış şeması gösterilmektedir.



Şekil 3.3: PSO tabanlı önerilen yöntemlerin akış şeması.

İkili PSO algoritmasında hız vektörü $V_{i,j}(t+1)$ ve konum vektörü $X_{i,j}(t+1)$ Denklem 3.47 ve 3.48'deki gibi hesaplanmaktadır. Lokal en iyi konum $Pbest_{i,j}(t+1)$ ve global en iyi konum $gbest_j$ konum vektörü değerine göre 0 veya 1 değerini almaktadır.

$$V_{i,j}(t+1) = w(V_{i,j}(t)) + c_1 r_1\big(Pbest_{i,j} - X_{i,j}(t)\big) + c_2 r_2\big(gbest_j - X_{i,j}(t)\big) \tag{3.46}$$

$$V_{i,j}(t+1) = \text{maks}\big(\min\left(V_{maks}, V_{i,j}(t+1)\right), V_{min}\big) \tag{3.47}$$

$$X_{i,j}(t+1) = \begin{Bmatrix} 1, & \frac{1}{1+e^{-V_{i,j}}} > rand \\ 0, & \frac{1}{1+e^{-V_{i,j}}} \leq rand \end{Bmatrix} \tag{3.48}$$

Uygunluk fonksiyonu Denklem 3.44'deki gibi hesaplanmaktadır. PSO parametreleri Tablo 3.2'de gösterilmektedir.

Tablo 3.2: Önerilen yöntemler için PSO parametreleri.

Parametre	**Değeri**
Popülasyon büyüklüğü	100
Maks. Iterasyon	100
V_{maks}	6
V_{min}	-6
c_1	2
c_2	2
w	1

3.6. VERİ KÜMELERİ

Mikrodizi ve RNA dizileme yöntemleriyle elde edilen gen ifade seviyesi verileriyle oluşturulmuş herkese açık veri kümeleri bulunmaktadır. Bu veri kümeleri makine öğrenmesi yöntemleriyle gen ifade analizleri, kanser teşhisi ve sınıflandırması araştırmalarında kullanılmaktadır.

Tablo 3.3'de yöntemlerin karşılaştırılmasında kullanılan veri kümeleri gösterilmektedir. Gen ifade seviyelerini içeren bu veri kümeleri literatürde bilgisayar destekli kanser teşhisi ve sınıflandırması çalışmalarında sıklıkla kullanılan başlıca veri kümeleri arasında yer almaktadır.

Bu veri kümeleri iki sınıftan oluşan ve genellikle sağlıklı ve kanserli hastalardan alınan örneklerle oluşturulmuş veri kümeleridir. Kanserli-sağlıklı doku ve kanser alt tür sınıflandırması gibi uygulamalarda kullanılmaktadırlar. Ayrıca, sağ kalım tahmini için kullanılan veri kümeleri de bulunmaktadır.

BRCA ve COAD veri kümeleri, RNA dizileme yöntemiyle elde edilen gen ifade seviyelerini, diğer veri kümeleri mikrodizi teknolojisiyle elde edilen gen ifade seviyelerini içermektedir.

Tablo 3.3: Veri kümeleri.

Veri Kümesi	İçerik	Gen Sayısı	Örnek Sayısı	Sınıf Sayısı	Sınıf Dağılımı
Lösemi (Golub ve diğ., 1999)	Akut miyeloid lösemi (AML) Akut lenfoblastik lösemi (ALL)	7129	72	2	AML:25 ALL:47
Kolon Kanseri (Alon ve diğ., 1999)	Normal örnekler Kanserli örnekler	2000	62	2	Normal:22 Kanser:40
Prostat Kanseri (Singh ve diğ., 2002)	Normal örnekler Kanserli örnekler	12600	136	2	Normal:59 Kanser:77
Over Kanseri (Petricoin ve diğ., 2002)	Normal örnekler Kanserli örnekler	15154	253	2	Normal:91 Kanser:162
DLBCL (Shipp ve diğ., 2002)	Yaygın Büyük B-hücreli lenfoma (DLBCL) Foliküler lenfoma (FL)	7129	77	2	DLBCL:58 FL:19
SMK (Spira ve diğ, 2007)	Normal örnekler Akciğer kanserli örnekler	19994	187	2	Normal:90 Kanser:97
Meme Kanseri (Naderi ve diğ. 2006)	Lüminal tip Lüminal olmayan tip	47293	128	2	L:84 NL:44
Meme Kanseri 2 (Veer ve diğ., 2002)	5 yıl içinde uzak metastaz (S1), ilk teşhisten sonra en az 5 yıl sağlıklı kalan hastalar (S0)	24481	97	2	S1:46 S0: 51
CNS (Pomeroy ve diğ., 2002)	Hayatta kalan (Sınıf 1) Kaybedilen (Sınıf 0)	7129	60	2	Sınıf 0:39 Sınıf 1:21
BRCA (GDCGD, 2020)	Normal örnekler Kanserli örnekler	60483	1211	2	Normal:113 Kanser:1098
COAD (GDCGD, 2020)	Normal örnekler Kanserli örnekler	60483	430	2	Normal:40 Kanser:390

Lösemi veri kümesi, AML ve ALL kanser alt türlerini birbirinden ayırmak için kullanılan, lösemili hastalardan alınmış toplam 72 kemik iliği ve periferik kan örneğinden elde edilmiştir. Bunlardan 25'i AML, 47'si ALL örneğidir. Gen ifade düzeyleri 7129 gen için yüksek yoğunluklu oligonükleotit mikrodizileri kullanılarak oluşturulmuştur (Golub ve diğ., 1999).

Kolon kanseri veri kümesi, Alon ve diğ. (1999) tarafından 40 tümör ve 22 normal kolon doku örneği için oligonükleotit mikrodizileri kullanılarak, 6500'ün üzerinde genin ifade düzeylerinin

belirlenmesiyle oluşturulmuştur. Bunlardan yüksek yoğunluklu 2000 adet gen kolon kanseri veri kümesinde kullanılmaktadır.

Prostat kanseri veri kümesi; 59 normal, 77 tümörlü doku olmak üzere toplam 136 örnek için 12600 genin ifade seviyelerini içermektedir.

Over kanseri veri kümesi; 91 normal, 162 tümörlü doku olmak üzere toplam 253 örnek için 15154 genin ifade düzeylerinden oluşmaktadır.

Yaygın büyük B-hücreli lenfoma ve foliküler lenfomanın ayrımı için oluşturulan DLBCL lenfoma veri kümesi toplam 77 örnekten oluşmaktadır. Her bir örnek için 7129 genin ifade düzeyi bulunmaktadır.

SMK veri kümesi sigara içeren kişilerde solunum yolu epitel hücrelerinin akciğer kanseri teşhisindeki etkisinin araştırılması için oluşturulmuştur. 90 normal, 97 kanserli doku olmak üzere toplam 187 örnek için 19994 genin ifade seviyelerini içermektedir.

Meme kanseri veri kümesi, lüminal ve lüminal olmayan tip meme kanseri ayrımı için kullanılmakta, toplam 128 örnekten ve her örnek için 47293 gen ifade seviyesinden oluşmaktadır.

Meme kanseri 2 veri kümesi, meme kanseri teşhisinden sonra ilk 5 yıl içinde metastaz olup olmamasının ayrımı için kullanılmaktadır.

CNS veri kümesi, merkezi sinir sistemi kanseri olan hastaların hayatta kalma ve kaybedilme ayrımının araştırılması için oluşturulmuştur. Toplam 60 örnek ve her örnek için 7129 genin ifade seviyelerini içermektedir.

BRCA veri kümesi, meme kanseri teşhisi için 113 normal, 1098 kanserli doku olmak üzere toplam 1211 örnek için 60483 genin ifade seviyelerini içermektedir.

COAD veri kümesi, kolon kanseri teşhisi için kullanılmakta ve 40 normal, 390 kanserli doku örneği için 60483 genin ifade seviyelerinden oluşmaktadır.

3.7. DEĞERLENDİRME ÖLÇÜTLERİ

Yöntemlerin sınıflandırma sonuçlarının değerlendirilmesinde; doğruluk oranı, kesinlik, duyarlılık, özgüllük, F1 ölçütü ve negatif kestirim değeri (NKD) kullanılmıştır. Tablo 3.4'de 2x2 karışıklık matrisi gösterilmektedir.

Tablo 3.4: Sınıflandırmanın karışıklık matrisi.

Gerçek Değer	**Öngörülen Değer**	
	Kanser (1)	**Normal (0)**
Kanser (1)	DP	YN
Normal (0)	YP	DN

3.7.1. Doğruluk Oranı

Doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır ve Denklem 3.49'daki gibi hesaplanır.

$$Doğruluk\ oranı\ (Accuracy) = \frac{DP+DN}{DP+DN+YP+YN} \tag{3.49}$$

3.7.2. Kesinlik

Gerçek sınıfı kanser olan ve kanser olarak sınıflandırılmış (DP) örnek sayısının kanser olarak sınıflandırılmış toplam örnek sayısına oranıdır.

$$Kesinlik\ (Precision) = \frac{DP}{DP+YP} \tag{3.50}$$

3.7.3. Duyarlılık

Gerçek sınıfı kanser olan ve kanser olarak sınıflandırılmış (DP) örnek sayısının gerçekte kanser olan toplam örnek sayısına oranıdır.

$$Duyarlılık\ (Sensitivity) = \frac{DP}{DP+YN} \tag{3.51}$$

3.7.4. Özgüllük

Gerçek sınıfı normal olan ve normal olarak sınıflandırılmış (DN) örnek sayısının gerçekte normal olan toplam örnek sayısına oranıdır.

$$\text{Özgüllük } (Specificity) = \frac{DN}{DN+YP} \tag{3.52}$$

3.7.5. F1-Ölçütü

Kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır.

$$F1 = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{3.53}$$

3.7.6. Negatif Kestirim Değeri

Gerçek sınıfı normal olan ve normal olarak sınıflandırılmış (DN) örnek sayısının normal olarak sınıflandırılmış toplam örnek sayısına oranıdır.

$$NKD = \frac{DN}{DN+YN} \tag{3.54}$$

4. BULGULAR

Bu çalışmada, Tablo 3.3'de gösterilen gen ifade seviyelerini içeren veri kümeleri için başlıca filtre yöntemlerle, sarmal yöntemlerle ve önerilen yöntemlerle öznitelik seçimi ve seçilen bu özniteliklerle sınıflandırma gerçekleştirilmiştir. Ayrıca, öznitelik seçimi yapılmadan başlıca sınıflandırma yöntemlerinin bu veri kümelerini sınıflandırma sonuçları ve başlıca boyut indirgeme yöntemleri sonrası sınıflandırma sonuçları elde edilmiştir. Bu bölümde, tüm değerlendirme ölçütleri için elde edilen sınıflandırma sonuçları gösterilmektedir. Sınıflandırmada kullanılan değerlendirme ölçütleri Bölüm 3.7'de anlatılmaktadır.

Bu bölüm, 5 alt bölüme ayrılmaktadır. Birinci alt bölümde, öznitelik seçme işlemi uygulanmadan sınıflandırma işlemi gerçekleştirildiğinde elde edilen sonuçlar gösterilmektedir. Böylece, başlıca sınıflandırma yöntemlerinin bu veri kümelerini sınıflandırma başarıları karşılaştırılmıştır. İkinci alt bölümde, 11 farklı filtre yöntem ile öznitelik seçimi gerçekleştirilmiştir. Filtre yöntemlerle seçilen özniteliklerin sınıflandırma başarısına olan etkisi incelenmiştir. Üçüncü alt bölümde, TBA ve LDA yöntemleriyle boyut indirgeme işlemi gerçekleştirilmiş ve sınıflandırma sonuçları karşılaştırılmıştır. Dördüncü alt bölümde, GA-DVM, GA-KNN, PSO-DVM, PSO-KNN sarmal yöntemlerinin sınıflandırma sonuçları sunulmaktadır. Beşinci alt bölümde, önerilen yöntemlerle elde edilen sınıflandırma sonuçları gösterilmektedir.

4.1. ÖZNİTELİK SEÇİMİ UYGULANMADAN SINIFLANDIRMA SONUÇLARI

Bu bölümde, veri kümelerinden öznitelik seçme işlemi gerçekleştirilmeden başlıca sınıflandırma yöntemleri için sınıflandırma sonuçları karşılaştırılmaktadır. Bu karşılaştırmada; doğrusal DVM (DDVM), polinomal DVM (PDVM), KNN, NB, KA ve İBYSA yöntemleri kullanılmıştır. KNN yönteminde k=7, uzaklık ölçütü olarak Öklid uzaklığı kullanılmıştır. Uygulamada tasarlanan YSA, 3 katmanlı ileri beslemeli bir yapay sinir ağıdır. Gizli katmandaki nöron sayısı 10'dur. Eğitimde, ölçekli eşlenik gradyan algoritması kullanılmıştır. YSA ayrıntıları Bölüm 3.3'de anlatılmaktadır.

Sınıflandırma işlemi için eğitim ve test verilerinin belirlenmesinde üç farklı yöntem kullanılmıştır. İlk yöntem, veri kümesinin %70 eğitim seti, %30 test seti olarak parçalara ayrılması (hold-out) yöntemidir. İkinci yöntem, veri kümesinin k sayıda eşit boyutlu parçalara

ayrıldığı K-katlamalı Çapraz Doğrulama (K-fold Cross Validation-CV) yöntemidir. Üçüncü yöntem ise çapraz doğrulamanın özel bir formu olan k sayısının örnek sayısına eşitlendiği Biri Hariç Çapraz Doğrulama (Leave One Out Cross Validation-LOOCV) yöntemidir.

4.1.1. Lösemi Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.1'de başlıca sınıflandırma yöntemlerinin lösemi veri kümesini sınıflandırma sonuçları hold-out değerlendirme yöntemi için gösterilmektedir. Buna göre, doğruluk oranı en yüksek olan yöntemlerin DDVM ve İBYSA olduğu görülmektedir.

Tablo 4.1: Lösemi veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	100,00	100,00	100,00	100,00	1,0000	100,00
PDVM	66,67	0,00	0,00	100,00	0,0000	66,67
KNN	90,48	100,00	71,43	100,00	0,8333	87,50
NB	95,24	87,50	100	92,86	0,9333	100
KA	85,71	70,00	100,00	78,57	0,8235	100
İBYSA	100,00	100,00	100,00	100,00	1,000	100,00

Tablo 4.2'de sınıflandırma sonuçları CV değerlendirme yöntemi kullanılarak gösterilmektedir. Buna göre, tüm değerlendirme ölçütlerine göre en yüksek değerler DDVM, NB ve İBYSA yöntemleriyle elde edilmiştir.

Tablo 4.2: Lösemi veri kümesi için sınıflandırma sonuçları (CV).

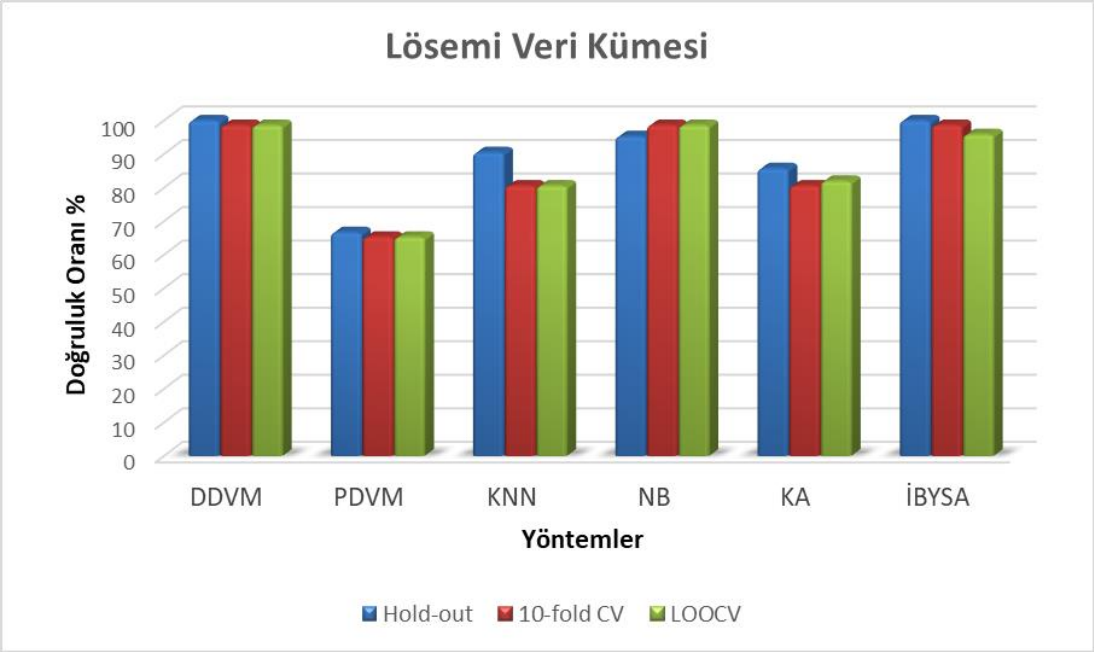
Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	98,61	100,00	96,00	100,00	0,9796	97,92
PDVM	65,28	0,00	0,00	100,00	0,0000	65,28
KNN	80,56	92,31	48,00	97,87	0,6316	77,97
NB	98,61	100,00	96,00	100,00	0,9796	97,92
KA	80,56	70,37	76,00	82,98	0,7308	86,67
İBYSA	98,61	100,00	96,00	100,00	0,9796	97,92

Tablo 4.3'de lösemi veri kümesinin sınıflandırma sonuçları LOOCV değerlendirme yöntemi kullanılarak gösterilmektedir. Buna göre, doğruluk oranı en yüksek olan yöntemlerin DDVM ve NB olduğu görülmektedir.

Tablo 4.3: Lösemi veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	98,61	100,00	96,00	100,00	0,9796	97,92
PDVM	65,28	0,00	0,00	100,00	0,0000	65,28
KNN	80,56	92,31	48,00	97,87	0,6316	77,97
NB	98,61	100,00	96,00	100,00	0,9796	97,92
KA	81,94	75,00	72,00	87,23	0,7347	85,42
İBYSA	95,83	100,00	88,00	100,00	0,9362	94,00

Şekil 4.1'de lösemi veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. En yüksek sonuç yaklaşık %98 doğruluk oranıyla NB yöntemiyle (LOOCV) elde edilmiştir.



Şekil 4.1: Lösemi veri kümesi için sınıflandırma sonuçları.

4.1.2. Kolon Kanseri Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.4, Tablo 4.5 ve Tablo 4.6'da başlıca sınıflandırma yöntemlerinin kolon kanseri veri kümesini sınıflandırma sonuçları görülmektedir. Tablo 4.4'deki sonuçlara göre, doğruluk oranı en yüksek olan yöntemler DDVM ve İBYSA'dır.

Tablo 4.5'deki sınıflandırma sonuçlarına göre, tüm değerlendirme ölçütleri açısından en başarılı sonuç İBYSA yöntemi ile elde edilmiştir. Doğruluk oranı en yüksek olan ilk iki yöntem İBYSA (%83,87) ve DDVM (%80,65) yöntemleridir.

Tablo 4.4: Kolon kanseri veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	94,44	100,00	83,33	100,00	0,9091	92,31
PDVM	66,67	0,00	0,00	100,00	0,0000	66,67
KNN	88,89	100,00	66,67	100,00	0,8000	85,71
NB	50,00	33,33	50,00	50,00	0,4000	66,67
KA	83,33	80,00	66,67	91,67	0,7273	84,62
İBYSA	94,44	100,00	83,33	100,00	0,9091	92,31

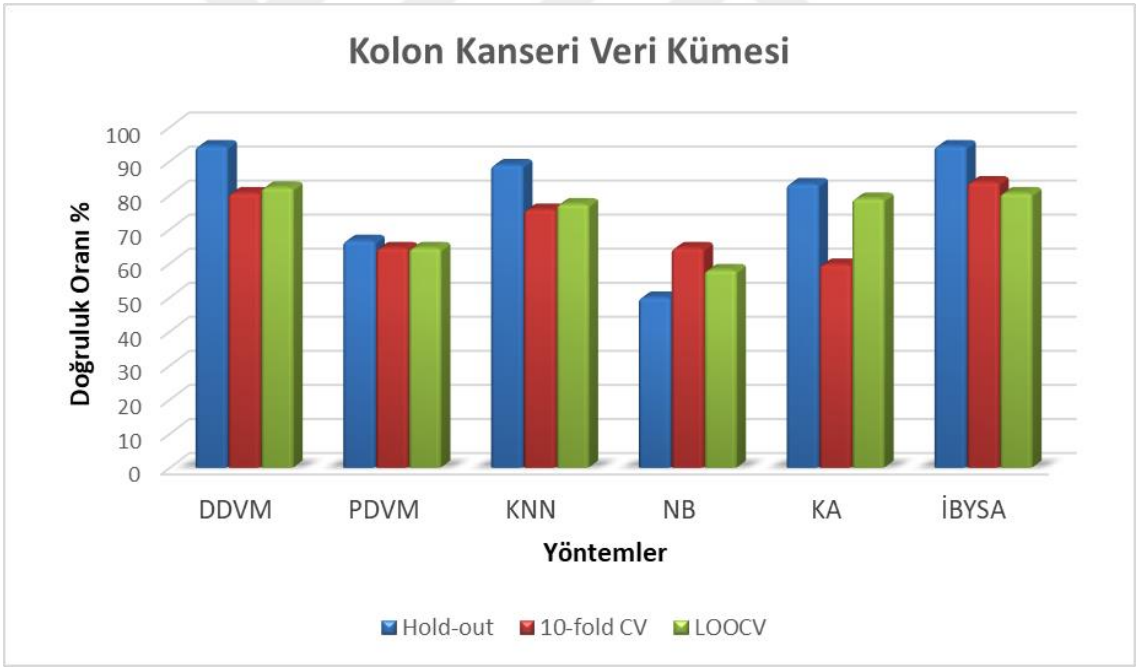
Tablo 4.5: Kolon kanseri veri kümesi için sınıflandırma sonuçları (CV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	80,65	72,73	72,73	85,00	0,7273	85,00
PDVM	64,52	0,00	0,00	100,00	0,0000	100,00
KNN	75,81	81,82	40,91	95,00	0,5455	74,50
NB	64,52	50,00	77,27	57,50	0,6071	82,14
KA	59,68	45,16	63,64	57,50	0,5283	74,19
İBYSA	83,87	77,27	77,27	87,50	0,7727	87,50

Tablo 4.6'daki sınıflandırma sonuçlarına göre, doğruluk oranı en yüksek olan yöntem %82,26 ile DDVM yöntemidir. Kesinlik ve duyarlılık ölçütleri açısından DDVM ve İBYSA aynı başarı oranını gösterirken İBYSA yönteminin doğruluk oranının %80,65 olduğu görülmektedir.

Tablo 4.6: Kolon kanseri veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	82,26	76,19	72,73	87,50	0,7442	85,37
PDVM	64,52	0,00	0,00	100,00	0,0000	64,52
KNN	77,42	83,33	45,45	95,00	0,5882	76,00
NB	58,06	44,44	72,73	50,00	0,5517	76,92
KA	79,03	71,43	68,18	85,00	0,6977	82,93
İBYSA	80,65	72,73	72,73	85,00	0,7273	85,00



Şekil 4.2: Kolon kanseri veri kümesi için sınıflandırma sonuçları.

Şekil 4.2'de kolon kanseri veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. Yaklaşık %94 doğruluk oranıyla en yüksek sonuç, DDVM yöntemiyle (hold-out) elde edilmiştir.

4.1.3. Prostat Kanseri Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.7, Tablo 4.8 ve Tablo 4.9'da prostat kanseri veri kümesi için sınıflandırma sonuçları gösterilmektedir. Tablo 4.7'deki sonuçlara göre, doğruluk oranı en yüksek olan yöntemler DDVM ve İBYSA yöntemleridir.

Tablo 4.7: Prostat kanseri veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	90,00	100,00	80,00	100,00	0,8889	83,33
PDVM	50,00	50,00	100,00	0,00	0,6667	0,00
KNN	70,00	66,67	80,00	60,00	0,7273	75,00
NB	56,67	53,85	93,33	20,00	0,6829	75,00
KA	86,67	86,67	86,67	86,67	0,8667	86,67
İBYSA	90,00	100,00	80,00	100,00	0,8889	83,33

Tablo 4.8'deki sınıflandırma sonuçlarına göre, tüm değerlendirme ölçütleri açısından en başarılı sonuç DDVM yöntemi ile elde edilmiştir. Doğruluk oranı en yüksek olan yöntemler %91,18 ile DDVM ve %90,20 ile İBYSA yöntemleridir.

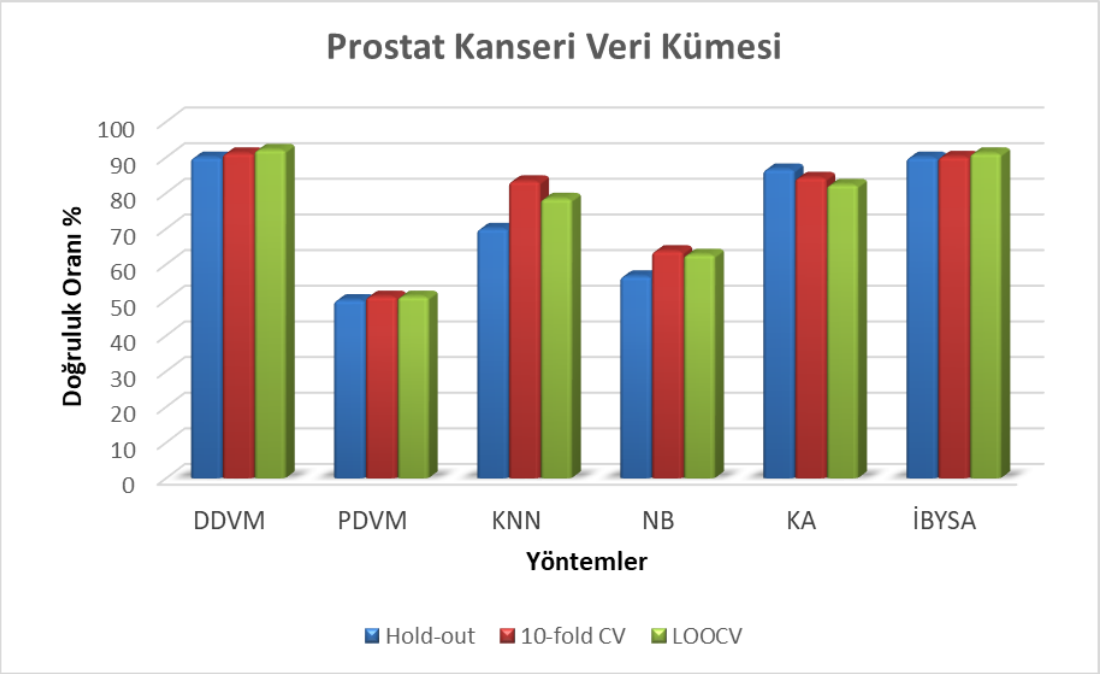
Tablo 4.8: Prostat kanseri veri kümesi için sınıflandırma sonuçları (CV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	91,18	93,88	88,46	94,00	0,9109	88,68
PDVM	50,98	50,98	100,00	0,00	0,6753	0,00
KNN	83,33	83,02	84,62	82,00	0,8381	83,67
NB	63,73	59,74	88,46	38,00	0,7132	76,00
KA	84,31	87,50	80,77	88,00	0,8400	81,48
İBYSA	90,20	92,27	88,46	92,00	0,9018	88,69

Tablo 4.9'daki sonuçlara göre, doğruluk oranı en yüksek olan yöntem %92,16 ile DDVM yöntemidir. Kesinlik, duyarlılık ve özgüllük ölçütleri açısından İBYSA en başarılı sonuçları oluşturmuştur. Doğruluk oranı %90,20'dir.

Tablo 4.9: Prostat kanseri veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	92,16	94,00	90,38	94,00	0,9216	90,38
PDVM	50,98	50,98	100,00	0,00	0,6753	0,00
KNN	78,43	82,61	73,08	84,00	0,7755	75,00
NB	62,75	58,97	88,46	36,00	0,7077	75,00
KA	82,35	81,48	84,62	80,00	0,8302	83,33
İBYSA	91,18	93,88	88,46	94,00	0,9109	88,68



Şekil 4.3: Prostat kanseri veri kümesi için sınıflandırma sonuçları.

Şekil 4.3'de prostat kanseri veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. Buna göre en yüksek doğruluk oranına (%92,16) DDVM yöntemiyle ulaşılmıştır.

4.1.4. Over Kanseri Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.10, Tablo 4.11 ve Tablo 4.12'de over kanseri veri kümesi için sınıflandırma sonuçları gösterilmektedir. Tablo 4.10'daki sonuçlara göre, %100 doğruluk oranıyla başarısı en yüksek olan yöntem İBYSA'dır.

Tablo 4.10: Over kanseri veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	98,67	100,00	97,92	100,00	0,9895	96,43
PDVM	64,00	64,00	100,00	0,00	0,7805	0,00
KNN	93,33	92,16	97,92	85,19	0,9495	95,83
KA	93,33	92,16	97,92	85,19	0,9495	95,83
İBYSA	100,00	100,00	100,00	100,00	1,0000	100,00

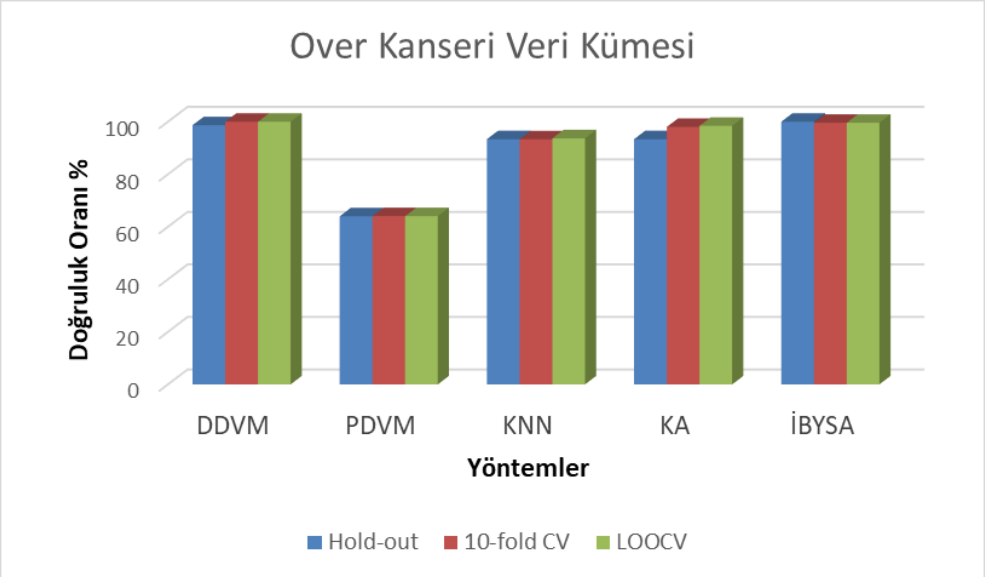
Tablo 4.11: Over kanseri veri kümesi için sınıflandırma sonuçları (CV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	100,00	100,00	100,00	100,00	1,0000	100,00
PDVM	64,03	64,03	100,00	0,00	0,7807	0,00
KNN	93,28	91,43	98,77	83,52	0,9496	97,44
KA	98,02	98,76	98,15	97,80	0,9845	96,74
İBYSA	99,60	99,39	100,00	98,90	0,9969	100,00

Tablo 4.11 ve Tablo 4.12'deki sınıflandırma sonuçlarına göre, tüm değerlendirme ölçütleri açısından en başarılı sonuç DDVM yöntemi ile elde edilmiştir. Tüm değerlendirme ölçütleri sonucu %100'dür.

Tablo 4.12: Over kanseri veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	100,00	100,00	100,00	100,00	1,0000	100,00
PDVM	64,03	64,03	100,00	0,00	0,7807	0,00
KNN	93,68	91,95	98,77	84,62	0,9524	97,47
KA	98,42	99,38	98,15	98,90	0,9876	96,77
İBYSA	99,60	99,39	100,00	98,90	0,9969	100,00



Şekil 4.4: Over kanseri veri kümesi için sınıflandırma sonuçları.

Şekil 4.4'de over kanseri veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. Yöntemlerin çoğunluğu için %100'e yakın sonuçlara ulaşılmıştır.

4.1.5. DLBCL Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.13'deki DLBCL kanseri veri kümesi için sınıflandırma sonuçlarına göre, doğruluk oranı en yüksek olan yöntemlerin DDVM ve İBYSA olduğu görülmektedir.

Tablo 4.13: DLBCL veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	95,65	94,44	100,00	83,33	0,9714	100,00
PDVM	73,91	73,91	100,00	0,00	0,8500	0,00
KNN	95,65	100,00	94,12	100,00	0,9697	85,71
NB	91,30	89,47	100,00	66,67	0,9444	100,00
KA	86,96	85,00	100,00	50,00	0,9189	100,00
İBYSA	95,65	94,44	100,00	83,33	0,9714	100,00

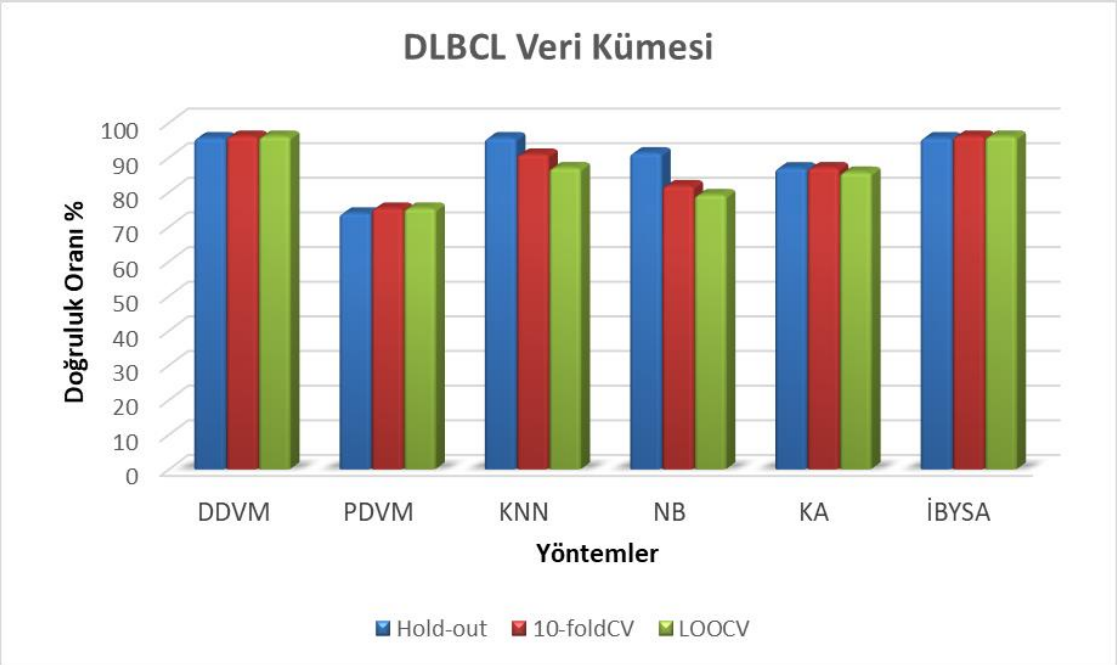
Tablo 4.14: DLBCL veri kümesi için sınıflandırma sonuçları (CV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	96,10	96,61	98,28	89,47	0,9744	94,44
PDVM	75,32	75,32	100,00	0,00	0,8593	0,00
KNN	90,91	100,00	87,93	100,00	0,9358	73,08
NB	81,82	85,48	91,38	52,63	0,8833	66,67
KA	87,01	88,71	94,83	63,16	0,9167	80,00
İBYSA	96,10	96,61	98,28	89,47	0,9744	94,44

Tablo 4.14 ve 4.15'deki sınıflandırma sonuçlarına göre, tüm değerlendirme ölçütleri açısından en başarılı sonuç DDVM ve İBYSA yöntemleriyle elde edilmiştir. Doğruluk oranı en yüksek olan yöntemler %96,10 ile DDVM ve İBYSA yöntemleridir.

Tablo 4.15: DLBCL veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	96,10	96,61	98,28	89,47	0,9744	94,44
PDVM	75,32	75,32	100,00	0,00	0,8593	0,00
KNN	87,01	100,00	82,76	100,00	0,9057	65,52
NB	79,22	85,00	87,93	52,63	0,8644	58,82
KA	85,71	87,30	94,83	57,89	0,9091	78,57
İBYSA	96,10	96,61	98,28	89,47	0,9744	94,44



Şekil 4.5: DLBCL kanseri veri kümesi için sınıflandırma sonuçları.

Şekil 4.5'de DLBCL kanseri veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. DDVM ve İBYSA yöntemleriyle en yüksek doğruluk oranına ulaşılmıştır.

4.1.6. SMK Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.16'daki sonuçlara göre, doğruluk oranı en yüksek olan yöntem %73,21 ile DDVM'dir. DDVM yöntemiyle elde edilen özgüllük oranı %59,26, duyarlılık %86,21'dir. Sağlıklı hastaları doğru sınıflandırma oranı düşük, kanserli hastaları doğru sınıflandırma oranının yüksek olduğu anlaşılmaktadır. Diğer yöntemlerle de benzer sonuçlar elde edildiği görülmektedir. Veri kümesinin sınıf dağılımının dengeli olduğu söylenebileceği için bu durumun veri kümesi dengesizliğinden kaynaklanmadığı düşünülmektedir. Hem bu durum hem de doğruluk oranı sonuçları göz önüne alındığında bu veri kümesi için sınıflandırma sonuçlarının iyileştirilmesi gerektiği anlaşılmaktadır.

Tablo 4.16: SMK veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	73,21	69,44	86,21	59,26	0,7692	80,00
PDVM	48,21	0,00	0,00	100,00	0,0000	48,21
KNN	71,43	67,57	86,21	55,56	0,7576	78,95
NB	64,29	64,52	68,97	59,26	0,6667	64,00
KA	55,36	57,14	55,17	55,56	0,5614	53,57
İBYSA	71,43	66,67	89,66	51,85	0,7647	82,35

Tablo 4.17'deki sonuçlarda da en başarılı yöntemlerin duyarlılık değerleri özgüllük değerlerinden yüksektir.

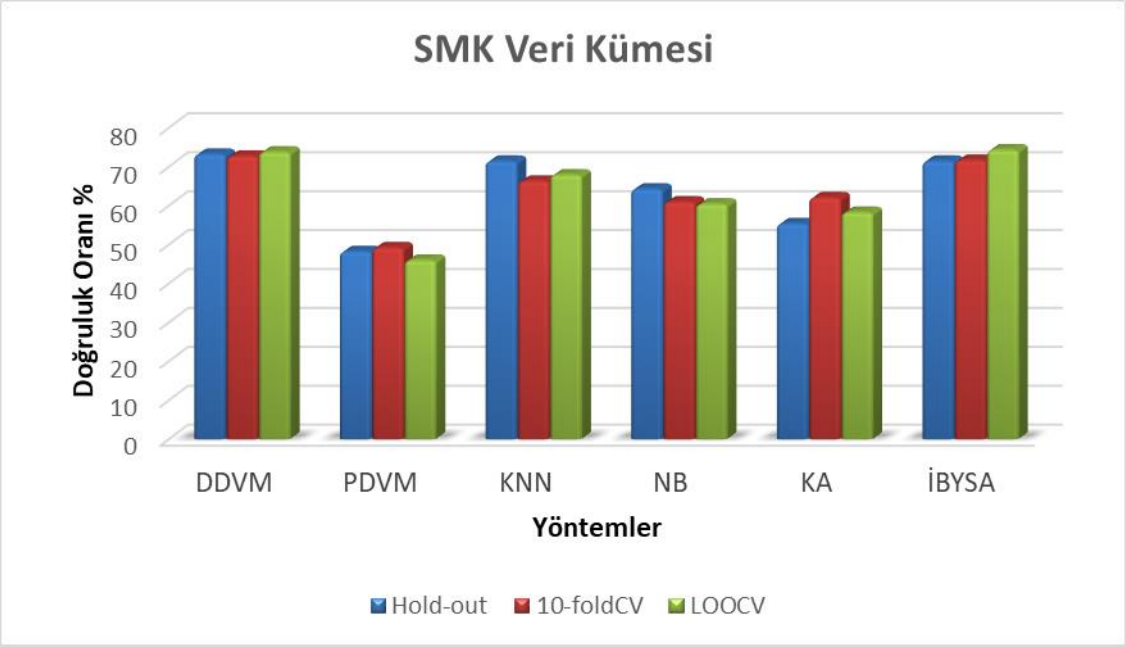
Tablo 4.17: SMK veri kümesi için sınıflandırma sonuçları (CV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	72,73	71,70	78,35	66,67	0,7488	74,07
PDVM	49,20	52,94	18,56	82,22	0,2748	48,37
KNN	66,31	64,66	77,32	54,44	0,7042	69,01
NB	60,96	63,95	56,70	65,56	0,6011	58,42
KA	62,03	64,77	58,76	65,56	0,6162	59,60
İBYSA	71,66	70,00	79,38	63,33	0,7440	74,03

Tablo 4.18'deki sınıflandırma sonuçlarına göre, doğruluk oranı en yüksek olan yöntem %74,33 ile İBYSA yöntemidir. Bu yöntemde de duyarlılık oranının özgüllük oranından daha yüksek olduğu görülmektedir.

Tablo 4.18: SMK veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	73,80	72,22	80,41	66,67	0,7610	75,95
PDVM	45,99	0,00	0,00	95,56	0,0000	46,99
KNN	67,91	66,37	77,32	57,78	0,7143	70,27
NB	60,43	63,22	56,70	64,44	0,5978	58,00
KA	58,29	59,41	61,86	54,44	0,6061	56,98
İBYSA	74,33	72,90	80,41	67,78	0,7647	76,25



Şekil 4.6: SMK veri kümesi için sınıflandırma sonuçları.

Şekil 4.6'daki SMK veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı sonuçlarına göre, İBYSA yöntemiyle maksimum doğruluk oranı (%74,33) elde edilmiştir.

4.1.7. Meme Kanseri Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.19'daki sonuçlara göre, tüm değerlendirme ölçütleri için en yüksek değerler DDVM ve İBYSA yöntemleriyle elde edilmiştir. Bu yöntemler için doğruluk oranı %84,21'dir.

Tablo 4.19: Meme kanseri veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	84,21	85,19	92,00	69,23	0,8846	81,82
PDVM	65,79	65,79	100,00	0,00	0,7937	0,00
KNN	81,58	78,13	100,00	46,15	0,8772	100,00
NB	81,58	87,50	84,00	76,92	0,8571	71,43
KA	76,32	86,36	76,00	76,92	0,8085	62,50
İBYSA	84,21	85,19	92,00	69,23	0,8846	81,82

Tablo 4.20'deki sınıflandırma sonuçlarına göre en yüksek doğruluk oranı %85,16 ile DDVM yöntemiyle elde edilmiştir. SMK veri kümesi sonuçlarına benzer olarak özgüllük oranı değerleri, duyarlılık oranı değerlerine göre daha düşüktür.

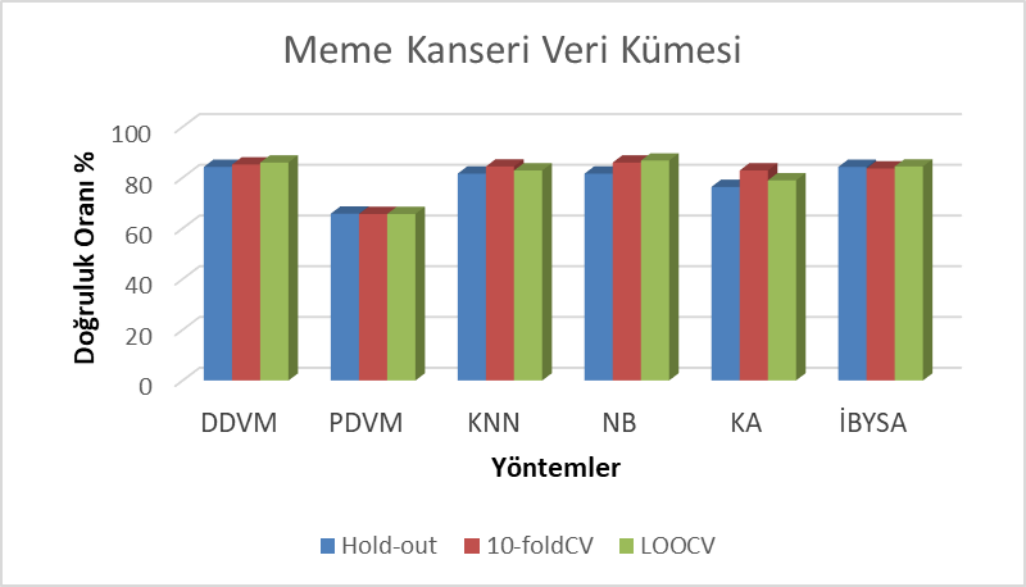
Tablo 4.20: Meme kanseri veri kümesi için sınıflandırma sonuçları (CV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	85,16	84,21	95,24	65,91	0,8939	87,88
PDVM	65,63	65,63	100,00	0,00	0,7925	0,00
KNN	84,38	81,37	98,81	56,82	0,8925	96,15
NB	85,94	87,50	91,67	75,00	0,8953	82,50
KA	82,81	82,98	92,86	63,64	0,8764	82,35
İBYSA	83,59	83,16	94,05	63,64	0,8827	84,85

Tablo 4.21: Meme kanseri veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	85,94	85,11	95,24	68,18	0,8989	88,24
PDVM	65,63	65,63	100,00	0,00	0,7925	0,00
KNN	82,81	80,39	97,62	54,55	0,8817	92,31
NB	86,72	88,51	91,67	77,27	0,9006	82,93
KA	78,91	81,32	88,10	61,36	0,8457	72,97
İBYSA	84,38	84,04	94,05	65,91	0,8876	85,29

Tablo 4.21'deki sınıflandırma sonuçlarına göre, doğruluk oranı en yüksek olan yöntem %85,94 ile DDVM yöntemidir. Bu yöntemde de duyarlılık oranının özgüllük oranından daha yüksek olduğu görülmektedir.



Şekil 4.7: Meme kanseri veri kümesi için sınıflandırma sonuçları.

Şekil 4.7'de meme kanseri veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. Maksimum doğruluk oranı, DDVM (LOOCV) yöntemi için hesaplanan %85,94'dür.

4.1.8. Meme Kanseri 2 Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.22'deki sonuçlara göre, tüm değerlendirme ölçütleri için en yüksek değerler KA yöntemiyle elde edilmiştir. Buna karşın, tüm yöntemlerin sonuçlarının oldukça düşük olduğu görülmektedir.

KA dışındaki tüm yöntemlerin doğruluk oranı %50'nin altındadır. Diğer değerlendirme ölçütleri açısından da sonuçlar oldukça düşüktür.

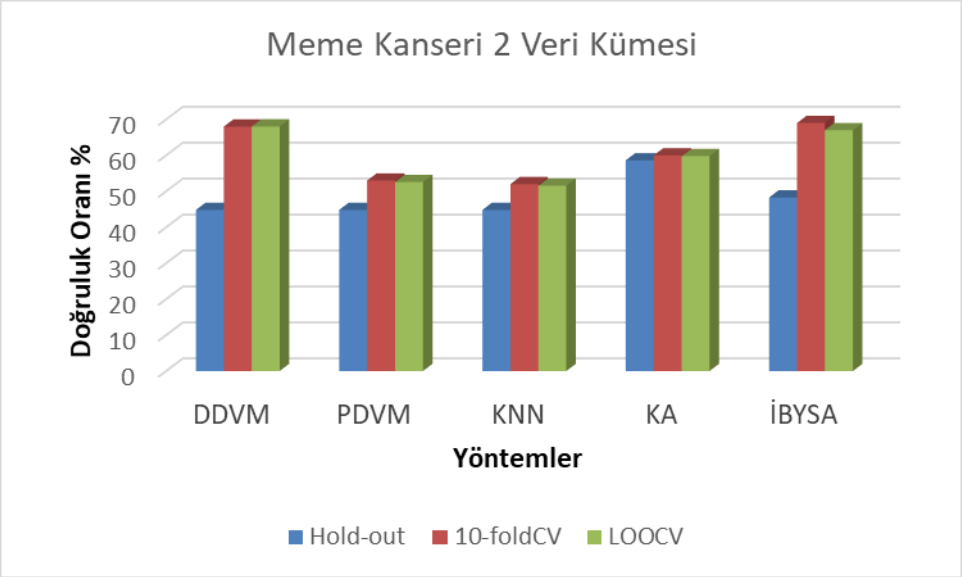
Tablo 4.22: Meme kanseri veri 2 kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	44,83	41,18	53,85	37,50	0,4667	50,00
PDVM	44,83	44,83	100,00	0,00	0,6190	0,00
KNN	44,83	33,33	23,08	62,50	0,2727	50,00
KA	58,62	53,33	61,54	56,25	0,5714	64,29
İBYSA	48,28	43,75	53,85	43,75	0,4828	53,85

Tablo 4.23'deki sınıflandırma sonuçlarına göre en yüksek doğruluk oranı %68,04 ile DDVM yöntemiyle elde edilmiştir.

Tablo 4.23: Meme kanseri 2 veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	68,04	65,96	67,39	68,63	0,6667	70,00
PDVM	52,58	0,00	0,00	100,00	0,0000	52,58
KNN	51,55	48,00	26,09	74,51	0,3380	52,78
KA	59,79	56,86	63,04	56,86	0,5979	63,04
İBYSA	67,01	66,67	60,87	72,55	0,6364	67,27



Şekil 4.8: Meme kanseri 2 veri kümesi için sınıflandırma sonuçları.

Şekil 4.8'de meme kanseri 2 veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. En yüksek doğruluk oranı DDVM yöntemi için LOOCV değerlendirme yöntemiyle elde edilmiştir. Bu veri kümesinin sınıflandırma sonuçlarının genel olarak oldukça düşük olduğu ve iyileştirilmesi gerektiği görülmektedir.

4.1.9. CNS Veri Kümesi için Sınıflandırma Sonuçları

Tablo 4.24: CNS kanseri veri kümesi için sınıflandırma sonuçları (hold-out).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	55,56	40,00	28,57	72,73	0,3333	61,54
PDVM	61,11	0,00	0,00	100,00	0,0000	61,11
KNN	55,56	40,00	28,57	72,73	0,3333	61,54
NB	66,67	66,67	28,57	90,91	0,4000	66,67
KA	66,67	55,56	71,43	63,64	0,6250	77,78
İBYSA	72,22	75,00	42,86	90,91	0,5455	71,43

Tablo 4.24, Tablo 4.25 ve Tablo 4.26'da CNS veri kümesi için sınıflandırma sonuçları gösterilmektedir. Tablo 4.24'deki sonuçlara göre, tüm değerlendirme ölçütleri açısından en başarılı sonuç İBYSA yöntemi ile elde edilmiştir.

Tablo 4.25: CNS kanseri veri kümesi için sınıflandırma sonuçları (CV).

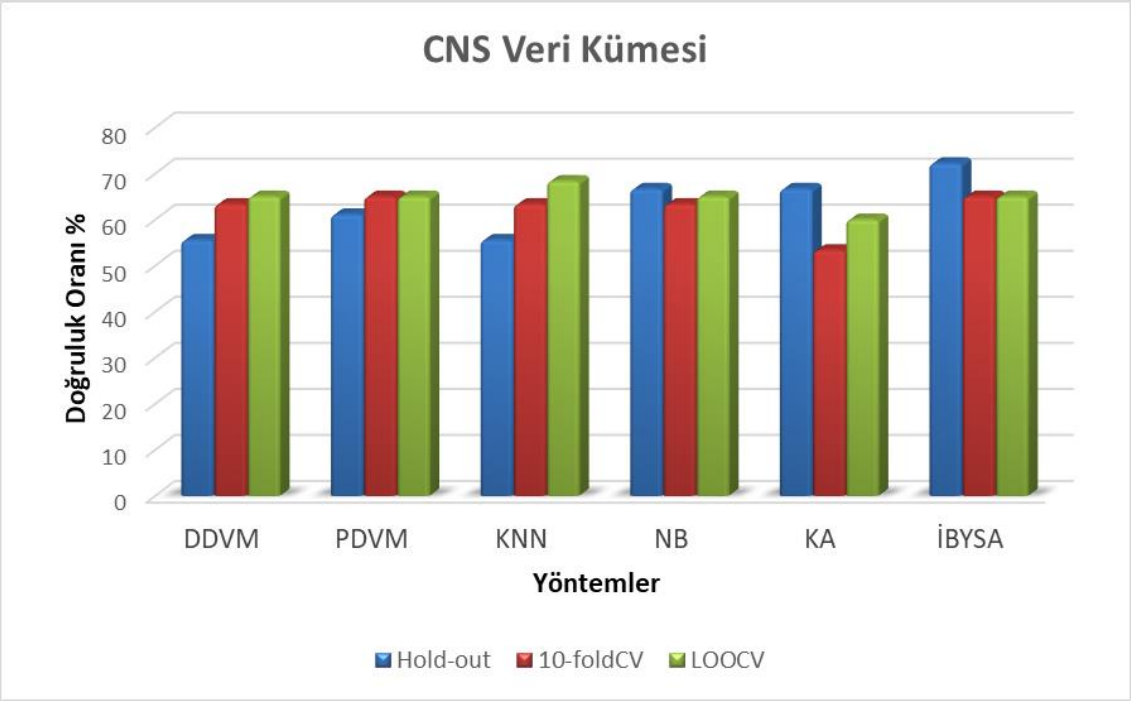
Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	63,33	47,37	42,86	74,36	0,4500	70,73
PDVM	65,00	0,00	0,00	100,00	0,0000	65,00
KNN	63,33	46,67	33,33	79,49	0,3889	68,89
NB	63,33	48,15	61,90	64,10	0,5417	75,76
KA	53,33	36,00	42,86	58,97	0,3913	65,71
İBYSA	65,00	50,00	38,10	79,49	0,4324	70,45

Tablo 4.25'deki sonuçlara göre, doğruluk oranı en yüksek olan yöntemler %65 ile polinomal DVM ve İBYSA yöntemleridir.

Tablo 4.26: CNS kanseri veri kümesi için sınıflandırma sonuçları (LOOCV).

Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
DDVM	65,00	50,00	38,10	79,49	0,4324	70,45
PDVM	65,00	0,00	0,00	100,00	0,0000	65,00
KNN	68,33	57,14	38,10	84,62	0,4571	71,74
NB	65,00	50,00	61,90	66,67	0,5532	76,47
KA	60,00	44,00	52,38	64,10	0,4783	71,43
İBYSA	65,00	50,00	38,10	79,49	0,4324	70,45

Tablo 4.26'daki sınıflandırma sonuçlarına göre, en yüksek doğruluk oranı %65'dir. Bu oran DDVM, PDVM, NB ve İBYSA yöntemleriyle elde edilmiştir.



Şekil 4.9: CNS kanseri veri kümesi için sınıflandırma sonuçları.

Şekil 4.9'da CNS kanseri veri kümesi için sınıflandırma yöntemlerinin doğruluk oranı grafiği gösterilmektedir. CNS veri kümesinin tüm sınıflandırma sonuçları incelendiğinde öznitelik seçimi olmadan doğruluk oranının oldukça düşük olduğu söylenebilir. Tüm yöntemlerin başarı oranının %72,22'nin üstüne çıkamadığı görülmektedir.

4.2. FİLTRE YÖNTEMLERLE ÖZNİTELİK SEÇİMİNİN SINIFLANDIRMA SONUÇLARINA ETKİSİNİN İNCELENMESİ

Bu bölümde, gen ifade seviyelerini içeren veri kümeleri için filtre yöntemlerle öznitelik seçme işlemi gerçekleştirilmiş ve seçilen özniteliklerin sınıflandırma başarısına olan etkisi incelenmiştir. Bu süreç, her filtre yöntem için veri kümelerindeki genlerin derecelendirilmesini ve derecelendirilen bu genlerin en yüksekten başlanarak eklenmesiyle sınıflandırma başarısının hesaplanmasını içermektedir. Sınıflandırma başarısı değerlendirmesinde hold-out yöntemi kullanılmıştır. Öznitelik seçimi uygulanmadan sınıflandırma gerçekleştirildiğinde sınıflandırma sonuçlarının yüksek olduğu lösemi, over kanseri gibi veri kümeleri bu sürece

dahil edilmemiştir. Diğer veri kümeleri için elde edilen sonuçlar bu bölümde gösterilmektedir. Veri kümelerinin tüm değerlendirme ölçütleri için sınıflandırma sonuçları EK.1'de tablolar halinde sunulmuştur.

4.2.1. Kolon Kanseri Veri Kümesi Sonuçları

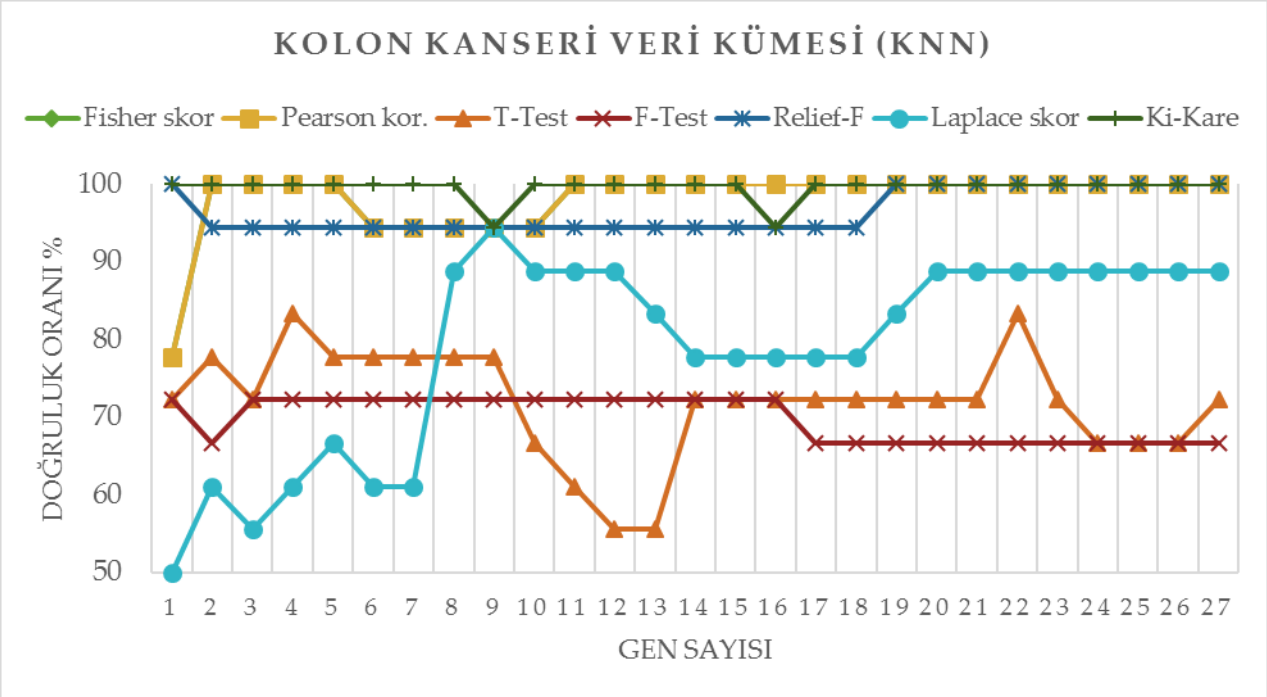
Tablo 4.27'de filtre yöntemlerle seçilen genler için maksimum doğruluk oranı ve bu sonucun elde edildiği gen sayısı gösterilmektedir.

Tablo 4.27: Kolon kanseri veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları.

Filtre Yöntemler	**Maksimum Doğruluk Oranı % (Gen sayısı)**					
	DDVM	**PDVM**	**KNN**	**NB**	**KA**	**İBYSA**
Pearson Kor.	100 (2)	100 (8)	100 (2)	100 (2)	94,44 (5)	100 (8)
Fisher skor	100 (2)	100,00 (8)	100 (2)	100 (2)	94,44 (5)	94,44 (8)
SNR	94,44 (12)	77,78 (26)	94,44 (880)	83,33 (37)	94,44 (637)	100 (33)
T-Test	100 (59)	94,44 (51)	100 (118)	83,33 (26)	88,89 (15)	94,44 (60)
F-Test	100 (22)	88,89 (16)	100 (720)	72,22 (18)	83,33 (26)	100 (175)
Relief-F	100 (1)	100 (3)	100 (1)	100 (1)	100 (31)	94,44 (3)
MRMR	94,44 (34)	88,89 (90)	94,44 (46)	94,44 (37)	94,44 (648)	100 (37)
Laplace skor	100 (13)	94,44 (39)	100 (35)	72,22 (28)	94,44 (30)	100 (13)
CFS	94,44 (219)	94,44 (11)	100 (134)	88,89 (2)	94,44 (179)	100 (10)
Karşılıklı Bilgi	94,44 (616)	83,33 (5)	100,00 (182)	72,22 (4)	83,33 (2)	83,33 (2)
Ki-kare	100,00 (1)	94,44 (7)	100,00 (1)	100,00 (1)	100,00 (13)	100,00 (25)
Tüm genler	94,44	66,67	88,89	50,00	83,33	94,44

Kolon kanseri veri kümesinin sonuçlarına göre Pearson Korelasyonu ve Fisher skor yöntemleriyle KA sınıflandırma yöntemi hariç tüm sınıflandırma yöntemleri için en fazla 8 genle %100 doğruluk oranı elde edilmiştir. Relief-F yöntemi için İBYSA yöntemi dışındaki yöntemlerin doğruluk oranı %100'dür. Bu oran Relief-F yöntemiyle seçilen 1, 3 veya 31 genin sınıflandırmada kullanılmasıyla elde edilen sonuçlardır. Bu nedenle, kolon kanseri veri kümesi için en başarılı filtre yöntemlerin Ki-kare, Relief-F, Pearson Korelasyonu ve Fisher skor olduğu görülmektedir.

Öznitelik seçimi yapılmadan tüm genler için hesaplanan sonuçlarla filtre yöntemlerin sonuçları karşılaştırıldığında tüm sınıflandırma yöntemleri için daha yüksek doğruluk oranı elde edilmiştir. Örneğin KNN yöntemi için öznitelik seçimi öncesi %88,89 olan doğruluk oranı öznitelik seçimiyle %100'e yükselmiştir.



Şekil 4.10: Filtre yöntemlerin kolon kanseri veri kümesinde KNN yöntemi için gen sayısı-doğruluk oranı grafiği.

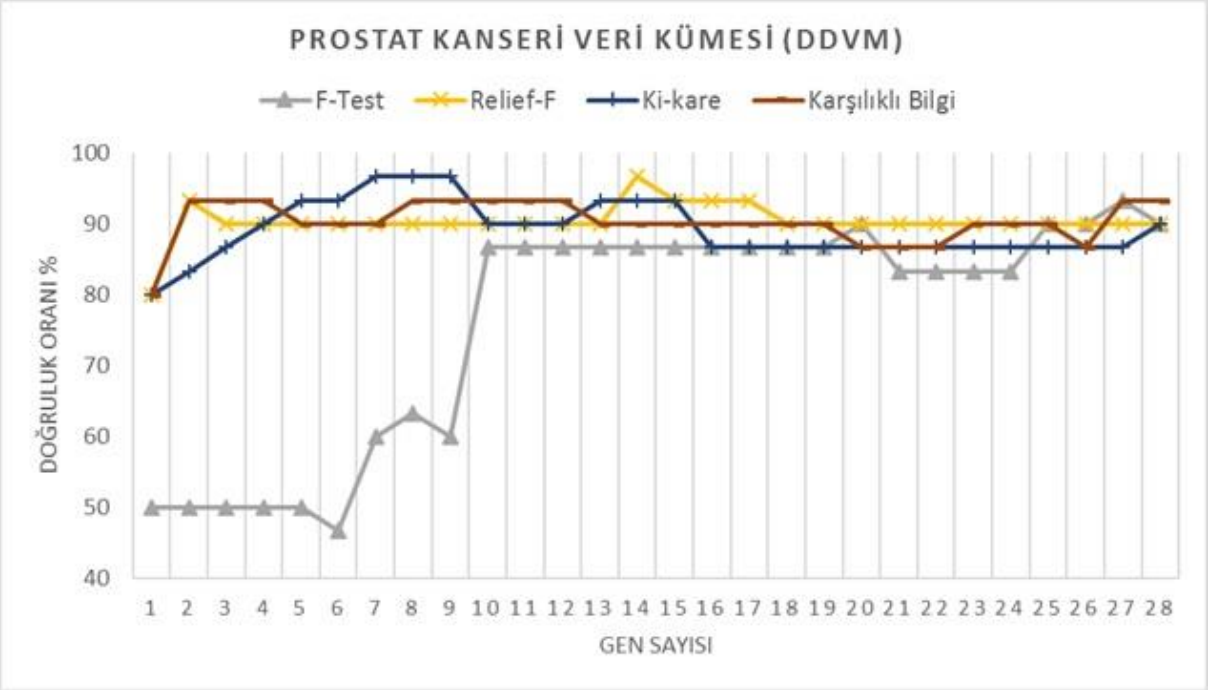
Şekil 4.10'da kolon kanseri veri kümesinin KNN yöntemi için doğruluk oranı en yüksek olan filtre yöntemlerin gen sayısı-doğruluk oranı grafiği gösterilmektedir. Pearson Korelasyonu ve Fisher skor yöntemlerinin ilk 50 gen için tüm sonuçları örtüşmektedir. Bu iki yöntemle 2 gen, Relief-F yöntemiyle 1 gen kullanılarak %100 doğruluk oranına ulaşılmıştır.

4.2.2. Prostat Kanseri Veri Kümesi Sonuçları

Tablo 4.28'deki sonuçlara göre öznitelik seçiminin, prostat kanseri veri kümesi için tüm sınıflandırma yöntemlerinin başarısını arttırdığı görülmektedir. En yüksek doğruluk oranı %96,67 ile F-test, Relief-F, Ki-kare ve Karşılıklı Bilgi yöntemleriyle seçilen genler için elde edilmiştir. DVM, KNN ve İBYSA sınıflandırma yöntemleriyle bu en yüksek doğruluk oranına ulaşıldığı görülmektedir. Karşılıklı bilgi ve KNN yöntemleri birlikte en az sayıda gen (2) ve en yüksek doğruluk oranı (%96,67) sonucunu oluşturmuştur.

Tablo 4.28: Prostat kanseri veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları.

Filtre Yöntemler	**Maksimum Doğruluk Oranı % (Gen sayısı)**					
	DDVM	**PDVM**	**KNN**	**NB**	**KA**	**İBYSA**
Pearson Kor.	93,33 (112)	93,33 (14)	93,33 (6)	93,33 (35)	83,33 (2)	93,33 (19)
Fisher skor	93,33 (112)	93,33 (14)	93,33 (6)	93,33 (35)	83,33 (2)	93,33 (5)
SNR	90,00 (119)	86,67 (55)	73,33 (243)	66,67 (2)	80,00 (360)	93,33 (92)
T-Test	93,33 (94)	93,33 (14)	93,33 (6)	93,33 (45)	83,33 (2)	93,33 (17)
F-Test	96,67 (74)	90,00 (10)	83,33 (10)	80,00 (10)	93,33 (177)	96,67 (36)
Relief-F	96,67 (14)	96,67 (14)	90,00 (2)	93,33 (2)	83,33 (691)	96,67 (9)
MRMR	90,00 (107)	73,33 (4)	90,00 (240)	90,00 (172)	86,67 (2)	93,33 (390)
Laplace skor	90,00 (70)	90,00 (728)	66,67 (224)	56,67 (1)	83,33 (2)	90,00 (203)
CFS	86,67 (644)	83,33 (47)	80,00 (46)	80,00 (3)	73,33 (54)	93,33 (702)
Karşılıklı Bilgi	93,33 (2)	96,67 (2)	96,67 (2)	93,33 (2)	90,00 (3)	93,33 (2)
Ki-kare	96,67 (7)	96,67 (29)	96,67 (8)	93,33 (5)	83,33 (2)	96,67 (14)
Tüm genler	90,00	50,00	70,00	56,67	86,67	90,00



Şekil 4.11: Filtre yöntemlerin prostat kanseri veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği.

Şekil 4.11'de DDVM doğruluk oranın en yüksek olduğu filtre yöntemler için gen sayısı-doğruluk oranı grafiği gösterilmektedir. Bu yöntemler ilk 30 gen için Relief-F, Karşılıklı Bilgi, Ki-kare ve F-test olarak sıralanabilir.

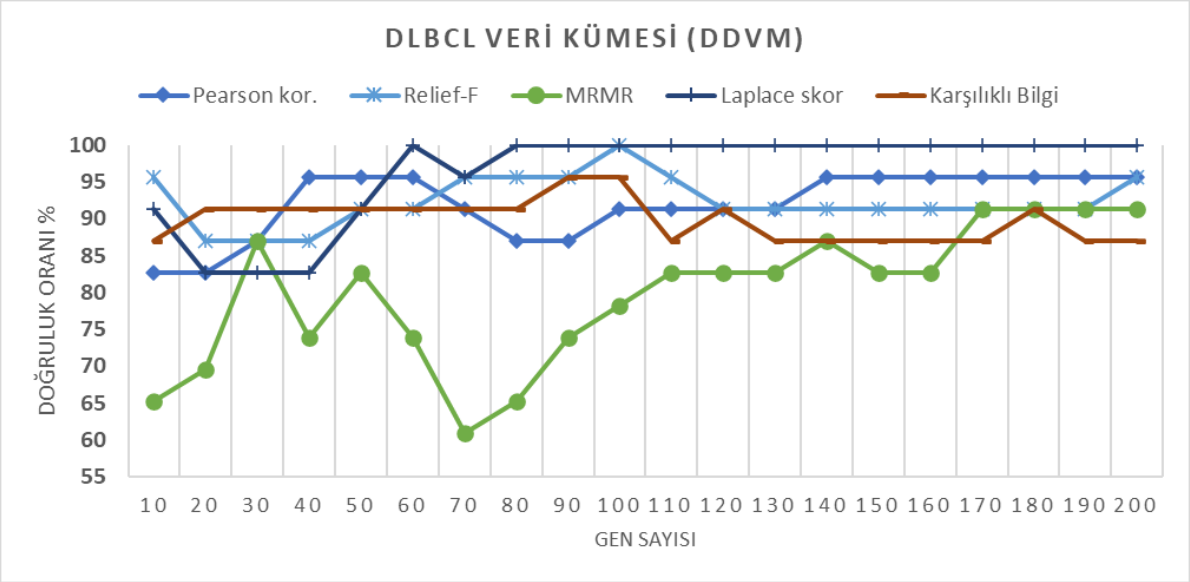
4.2.3. DLBCL Veri Kümesi Sonuçları

Tablo 4.29'daki sonuçlara göre Relief-F yöntemiyle seçilen genlerle NB, KA ve İBYSA yöntemleri için %100 doğruluk oranı elde edildiği görülmektedir. Bu sonuç, her üç yöntem için de tüm genlerin kullanıldığı sonuçlara göre daha yüksektir. Ayrıca, MRMR + KNN, Laplace skor + DDVM, Karşılıklı Bilgi + KNN, Karşılıklı Bilgi + İBYSA yöntemlerinin doğruluk oranı sonucu da %100'dür.

Tablo 4.29: DLBCL veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları.

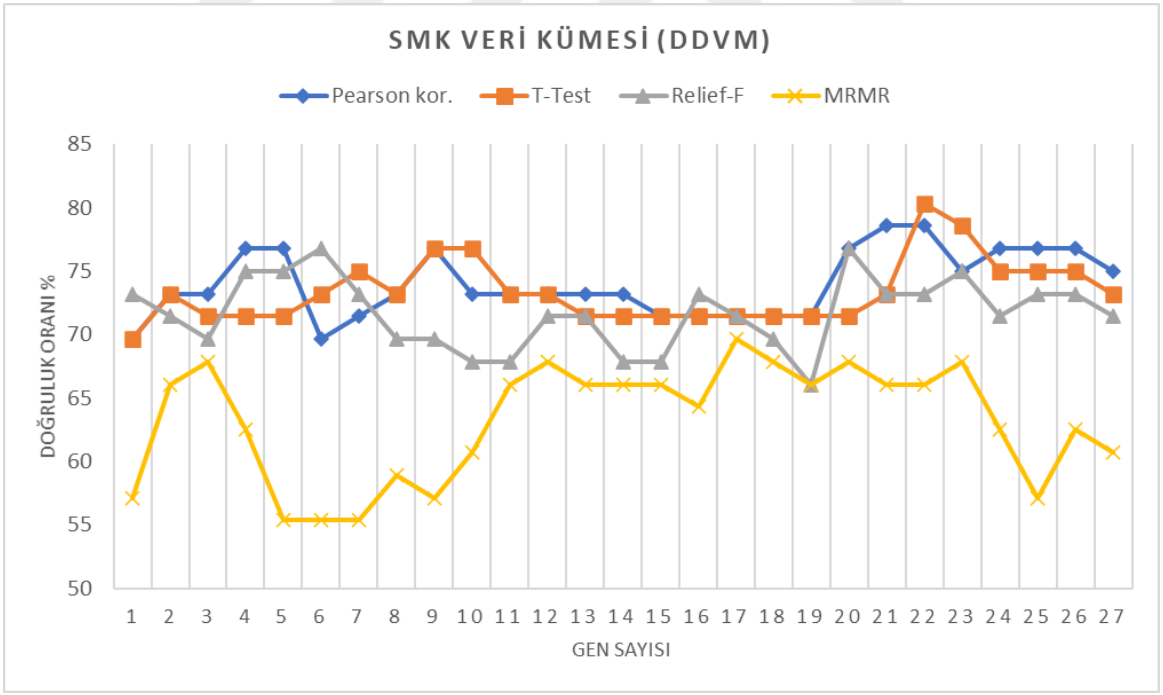
Filtre Yöntemler	Maksimum Doğruluk Oranı % (Gen sayısı)					
	DDVM	**PDVM**	**KNN**	**NB**	**KA**	**İBYSA**
Pearson Kor.	95,65 (40)	91,30 (140)	95,65 (160)	95,65 (50)	95,65 (50)	100 (200)
Fisher skor	95,65 (30)	91,30 (150)	95,65 (160)	95,65 (50)	95,65 (50)	95,65 (40)
SNR	95,65 (5460)	86,96 (80)	95,65 (350)	95,65 (130)	95,65 (130)	95,65 (130)
T-Test	95,65 (6890)	91,30 (20)	95,65 (50)	91,30 (20)	91,30 (20)	95,65 (630)
F-Test	95,65 (470)	86,96 (50)	95,65 (6890)	86,96 (80)	91,30 (6780)	95,65 (470)
Relief-F	95,65 (80)	86,96 (20)	95,65 (20)	100 (110)	100 (110)	100 (240)
MRMR	95,65 (210)	78,26 (20)	100,00 (330)	78,26 (540)	91,30 (6850)	95,65 (160)
Laplace skor	100 (60)	95,65 (30)	95,65 (140)	95,65 (70)	95,65 (70)	100 (90)
CFS	95,65 (6270)	86,96 (90)	95,65 (930)	86,96 (100)	91,30 (4220)	95,65 (560)
Karşılıklı Bilgi	95,65 (90)	91,30 (5)	100,00 (98)	91,30 (2)	86,96 (30)	100,00 (11)
Ki-kare	95,65(9)	95,65 (5)	95,65 (35)	95,65 (16)	91,30 (4)	95,65 (4)
Tüm genler	95,65	73,91	95,65	91,3	86,96	95,65

DDVM ve KNN yöntemlerinin diğer sonuçları öznitelik seçimi olmadan hesaplanan sonuçlara benzerlik göstermektedir. Karşılıklı bilgi ve İBYSA birlikte, en az sayıda genle (11) en yüksek doğruluk oranın elde edildiği yöntem olmuştur.



Şekil 4.12: Filtre yöntemlerin DLBCL veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği.

4.2.4. SMK Veri Kümesi Sonuçları



Şekil 4.13: Filtre yöntemlerin SMK veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği.

Tablo 4.30'daki sonuçlar göz önüne alındığında, Laplace skor + DDVM yöntemi hariç diğer tüm yöntemlerle, öznitelik seçimi olmadan hesaplanan sonuçlardan daha başarılı sonuçlar elde edildiği görülmektedir. En iyi sonuç, Pearson Korelasyonu + İBYSA yöntemiyle %87,5 doğruluk oranıdır.

Şekil 4.13'deki grafik; Pearson Korelasyonu, T-test, Relief-F yöntemleri için DDVM yöntemiyle elde edilen doğruluk oranı sonuçlarının ilk 30 gen için birbirine çok yakın olduğunu göstermektedir.

Tablo 4.30: SMK veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları.

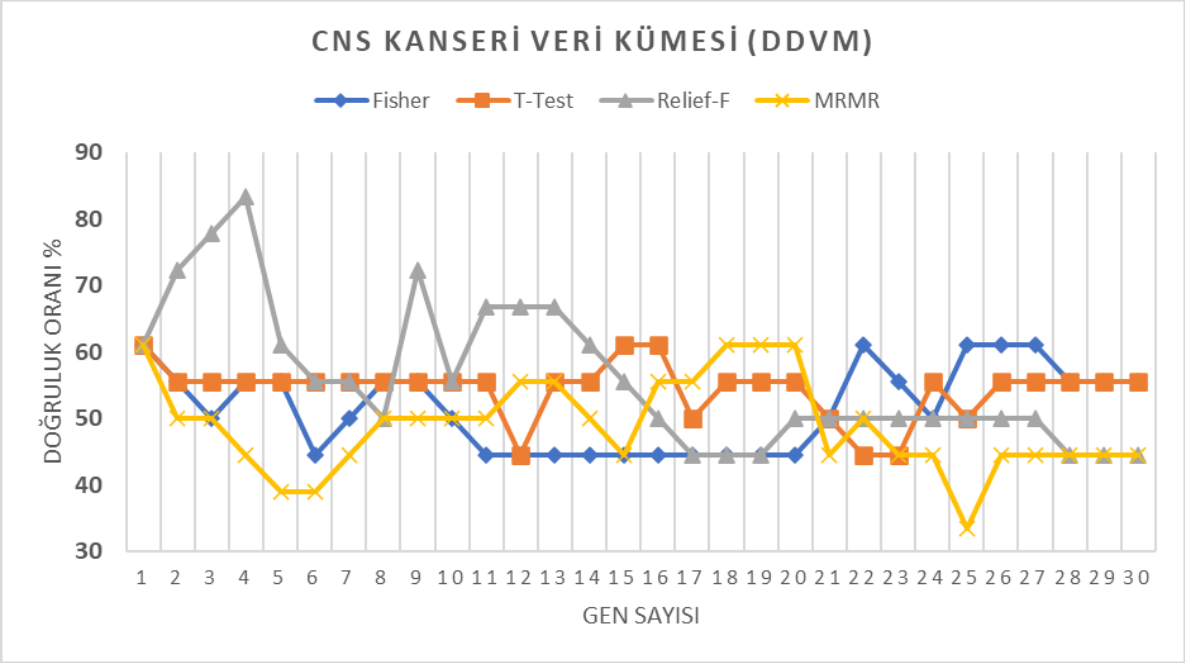
Filtre Yöntemler	**Maksimum Doğruluk Oranı % (Gen sayısı)**					
	DDVM	**PDVM**	**KNN**	**NB**	**KA**	**İBYSA**
Pearson Kor.	80,36 (331)	87,50 (130)	78,57 (215)	82,14 (24)	75,00 (3)	87,50 (38)
Fisher skor	80,36 (331)	87,50 (130)	78,57 (215)	82,14 (24)	75,00 (3)	85,71 (604)
SNR	74,30 (30)	75,00 (24)	76,79 (72)	71,43 (17)	71,43 (709)	80,36 (552)
T-Test	80,36 (22)	85,71 (35)	82,14 (349)	82,14 (10)	73,21 (22)	83,93 (27)
F-Test	75,00 (21)	73,21 (39)	73,21 (404)	75,00 (499)	71,43 (395)	78,57 (742)
Relief-F	83,93 (81)	82,20 (90)	85,71 (93)	80,36 (34)	75,00 (627)	85,71 (56)
MRMR	82,14 (611)	80,50 (350)	73,21 (274)	66,07 (413)	71,43 (4)	82,14 (497)
Laplace skor	73,21 (10)	70,00 (56)	76,79 (74)	62,50 (2)	75,00 (322)	76,79 (52)
CFS	76,79 (197)	73,30 (110)	78,57 (83)	69,64 (7)	64,29 (36)	78,57 (197)
Karşılıklı Bilgi	78,57 (146)	71,43 (339)	73,21 (3)	67,86 (9)	73,21 (569)	80,36 (151)
Ki-kare	78,57 (313)	76,79 (184)	82,14 (35)	78,57 (18)	69,64 (1)	78,57 (135)
Tüm genler	73,21	48,21	71,43	64,29	55,36	71,43

4.2.5. CNS Veri Kümesi Sonuçları

Tablo 4.31'deki CNS veri kümesi sonuçları incelediğinde, en yüksek doğruluk oranının %88,89 olduğu görülmektedir. Bu sonuç, Fisher skor, T-test, MRMR filtre yöntemleri ile İBYSA sınıflandırma yönteminin birlikte kullanılmasıyla hesaplanmıştır. En az genle en yüksek doğruluk oranı sonucunun Fisher skor yöntemiyle elde edildiği görülmektedir.

Tablo 4.31: CNS veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları.

Filtre Yöntemler	**Maksimum Doğruluk Oranı % (Gen sayısı)**					
	DDVM	**PDVM**	**KNN**	**NB**	**KA**	**İBYSA**
Pearson Kor.	77,78 (86)	77,78 (12)	77,78 (2)	66,67 (1)	72,22 (1)	83,33 (72)
Fisher skor	72,22 (36)	77,78 (12)	77,78 (2)	66,67 (1)	72,22 (1)	88,89 (77)
SNR	72,22 (25)	72,22 (551)	72,22 (156)	77,78 (11)	66,67 (3)	83,33 (19)
T-Test	72,22 (93)	72,22 (11)	77,78 (2)	61,11 (1)	72,22 (4)	88,89 (152)
F-Test	77,78 (445)	66,67 (28)	83,33 (3)	72,22 (197)	72,22 (6)	77,78 (1455)
Relief-F	83,33 (4)	72,22 (4)	66,67 (2)	61,11 (119)	61,11 (1)	83,33(5)
MRMR	77,78 (468)	77,78 (24)	77,78 (5)	77,78 (263)	83,33 (32)	88,89 (1031)
Laplace skor	77,78 (146)	72,22 (773)	72,22 (32)	66,67 (511)	77,78 (83)	83,33 (134)
CFS	72,22 (382)	72,22 (13)	61,11 (1)	66,67 (59)	72,22 (237)	72,22 (257)
Karşılıklı Bilgi	72,22 (175)	77,78 (867)	72,22 (5)	66,67 (4)	66,67 (6)	72,22 (775)
Ki-kare	77,78 (45)	77,78 (46)	83,33 (275)	66,67 (83)	77,78 (24)	83,33 (333)
Tüm genler	55,56	61,11	55,56	66,67	66,67	72,22



Şekil 4.14: Filtre yöntemlerin CNS veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği.

Şekil 4.14'deki grafikte, ilk 30 gende DDVM ile en yüksek doğruluk oranının elde edildiği yöntemin Relief-F olduğu görülmektedir. Bu yöntem için 4 gen ile %83,33 doğruluk oranına ulaşılmıştır. Diğer yöntemler için ilk 30 gende sonuç en fazla %60'dır.

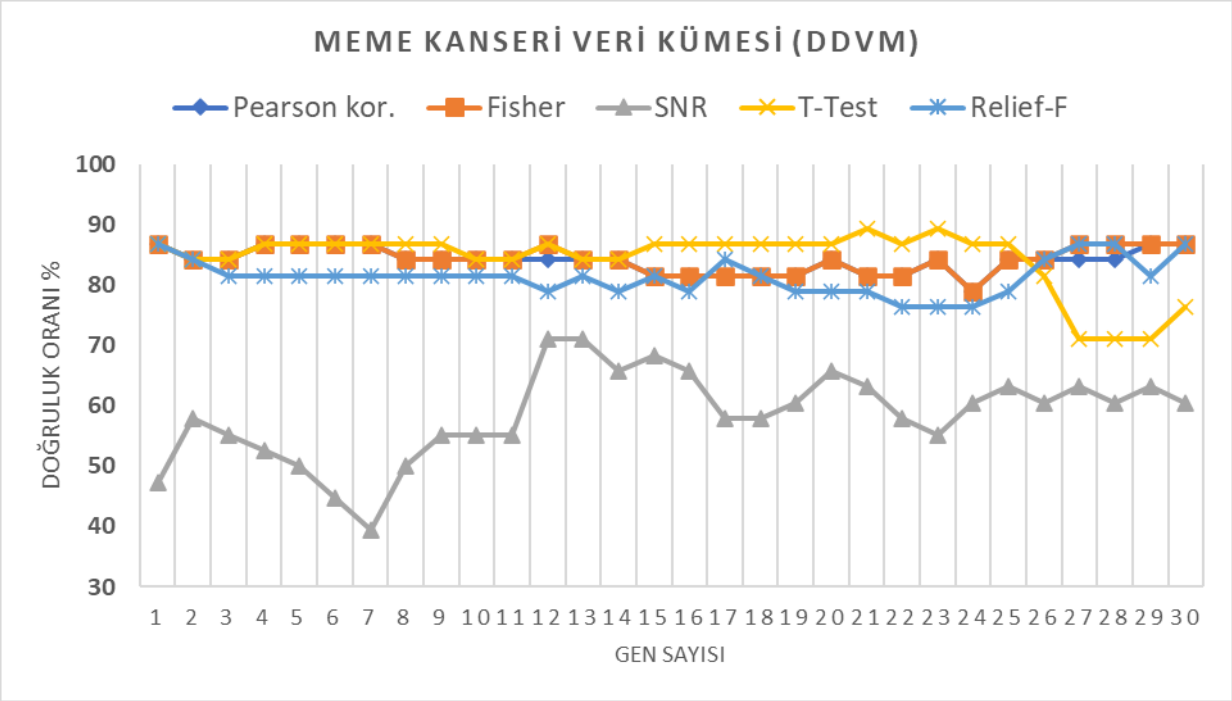
4.2.6. Meme Kanseri Veri Kümesi Sonuçları

Tablo 4.32'deki sonuçlara göre en yüksek doğruluk oranı Fisher skor + İBYSA yöntemiyle %92,11'dir. Tüm yöntemler için öznitelik seçiminin doğruluk oranını iyileştirdiği görülmektedir.

Şekil 4.15'deki meme kanseri veri kümesi sonuçlarına göre, bu veri kümesinin ilk 30 geni için T-test, Pearson Korelasyonu ve Relief-F en başarılı filtre yöntemler olmuştur. SNR %70'in üstüne çıkamamıştır.

Tablo 4.32: Meme kanseri veri kümesinden filtre yöntemlerle seçilen genler için sınıflandırma sonuçları.

Filtre Yöntemler	**Maksimum Doğruluk Oranı % (Gen sayısı)**					
	DDVM	**PDVM**	**KNN**	**NB**	**KA**	**İBYSA**
Pearson Kor.	89,47 (106)	92,11 (47)	84,21 (698)	86,84 (386)	86,84 (1)	92,11 (100)
Fisher skor	89,47 (106)	89,47 (20)	84,21 (701)	86,84 (388)	86,84 (1)	92,11 (32)
SNR	92,11 (278)	73,68 (40)	86,84 (112)	76,32 (59)	81,58 (85)	92,11 (275)
T-Test	92,11 (505)	86,84 (5)	84,21 (4)	86,84 (3)	86,84 (1)	92,11 (51)
F-Test	86,84 (93)	81,58 (5)	78,95 (58)	84,21 (1039)	84,21 (111)	89,47 (170)
Relief-F	92,11 (33)	89,47 (40)	84,21 (218)	86,84 (361)	86,84 (1)	92,11 (81)
Laplace skor	86,84 (288)	84,21 (332)	84,21 (82)	84,21 (7)	81,58 (4)	89,47 (282)
Karşılıklı Bilgi	89,47 (21)	86,84 (5)	84,21 (2)	86,84 (1)	84,21 (2)	89,47 (21)
Ki-kare	92,11 (1052)	86,84 (28)	84,21 (113)	84,21 (3)	86,84 (392)	92,11 (274)
Tüm genler	84,21	65,79	81,58	81,58	76,32	84,21



Şekil 4.15: Filtre yöntemlerin meme kanseri veri kümesinde DDVM yöntemi için gen sayısı-doğruluk oranı grafiği.

4.3. TBA VE LDA YÖNTEMLERİYLE BOYUT İNDİRGEME SONRASI SINIFLANDIRMA SONUÇLARI

Gen ifade seviyelerini içeren veri kümelerinde TBA ve LDA yöntemleriyle boyut indirgeme işlemi gerçekleştirilmiştir. Boyut indirgeme sonrasında her bir veri kümesi için KNN, DDVM, PDVM, KA, İBYSA olmak üzere beş farklı yöntemle sınıflandırma yapılmıştır.

Veri kümelerinin %70 eğitim seti, %30 test seti olarak parçalara ayrıldığı (hold-out) yöntem için elde edilen sınıflandırma sonuçları tablolar halinde sunulmaktadır.

Tablo 4.33'de TBA ve LDA ile boyut indirgeme sonrasında lösemi veri kümesi için sınıflandırma sonuçları gösterilmektedir. TBA için DDVM ve İBYSA yöntemleriyle, LDA için PDVM ve KNN yöntemleriyle diğer yöntemlere göre daha yüksek doğruluk oranı elde edildiği görülmektedir.

Tablo 4.33: Lösemi veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	100,00	100,00	100,00	100,00	1,00	100,00
	PDVM	90,48	85,71	85,71	92,86	0,86	92,86
	KNN	90,48	85,71	85,71	92,86	0,86	92,86
	KA	90,48	77,78	100,00	85,71	0,88	100,00
	İBYSA	95,24	100,00	85,71	100,00	0,92	93,33
LDA	**DDVM**	95,24	87,50	100,00	92,86	0,93	100,00
	PDVM	100,00	100,00	100,00	100,00	1,00	100,00
	KNN	95,24	100,00	85,71	100,00	0,92	93,33
	KA	90,48	100,00	71,43	100,00	0,83	87,50
	İBYSA	85,71	83,33	71,43	92,86	0,77	86,67

Boyut indirgemenin yanı sıra sınıflandırma yönteminin de doğruluk oranı sonucunu etkilediği anlaşılmaktadır.

Tablo 4.34 ve Tablo 4.35'deki kolon kanseri ve prostat kanseri veri kümeleri sınıflandırma sonuçlarına göre TBA ile yapılan boyut indirgeme işleminin daha başarılı olduğu görülmektedir. Her iki tabloda da 5 yöntemden 3'ü için TBA yöntemi sonrası daha yüksek doğruluk oranı elde edilmiştir.

Tablo 4.34: Kolon kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	100,00	100,00	100,00	100,00	1,00	100,00
	PDVM	83,33	80,00	66,67	91,67	0,73	84,62
	KNN	94,44	100,00	83,33	100,00	0,91	92,31
	KA	83,33	71,43	83,33	83,33	0,77	90,91
	İBYSA	94,44	100,00	83,33	100,00	0,91	92,31
LDA	**DDVM**	83,33	71,43	83,33	83,33	0,77	90,91
	PDVM	88,89	83,33	83,33	91,67	0,83	91,67
	KNN	88,89	83,33	83,33	91,67	0,83	91,67
	KA	83,33	80,00	66,67	91,67	0,73	84,62
	İBYSA	83,33	80,00	66,67	91,67	0,73	84,62

Tablo 4.35: Prostat kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	96,67	100,00	93,33	100,00	0,97	93,75
	PDVM	93,33	100,00	86,67	100,00	0,93	88,24
	KNN	73,33	70,59	80,00	66,67	0,75	76,92
	KA	60,00	58,82	66,67	53,33	0,63	61,54
	İBYSA	93,33	100,00	86,67	100,00	0,93	88,24
LDA	**DDVM**	86,67	86,67	86,67	86,67	0,87	86,67
	PDVM	90,00	100,00	80,00	100,00	0,89	83,33
	KNN	83,33	85,71	80,00	86,67	0,83	81,25
	KA	70,00	68,75	73,33	66,67	0,71	71,43
	İBYSA	73,33	73,33	73,33	73,33	0,73	73,33

Tablo 4.36: Over kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	100,00	100,00	100,00	100,00	1,00	100,00
	PDVM	98,67	97,96	100,00	96,30	0,99	100,00
	KNN	94,67	94,00	97,92	88,89	0,96	96,00
	KA	90,67	91,84	93,75	85,19	0,93	88,46
	İBYSA	100,00	100,00	100,00	100,00	1,00	100,00
LDA	**DDVM**	100	100	100	100	1	100
	PDVM	100	100	100	100	1	100
	KNN	100	100	100	100	1	100
	KA	100	100	100	100	1	100
	İBYSA	100	100	100	100	1	100

Tablo 4.36'daki over kanseri veri kümesi sonuçlarına göre, LDA yöntemi sonrası tüm sınıflandırma yöntemleri için %100 doğruluk oranı sonucu bulunmuştur.

Tablo 4.37: DLBCL veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	100,00	100,00	100,00	100,00	1,00	100,00
	PDVM	95,65	94,44	100,00	83,33	0,97	100,00
	KNN	95,65	100,00	94,12	100,00	0,97	85,71
	KA	82,61	84,21	94,12	50,00	0,89	75,00
	İBYSA	100,00	100,00	100,00	100,00	1,00	100,00
LDA	**DDVM**	86,96	100,00	82,35	100,00	0,90	66,67
	PDVM	86,96	85,00	100,00	50,00	0,92	100,00
	KNN	78,26	77,27	100,00	16,67	0,87	100,00
	KA	78,26	77,27	100,00	16,67	0,87	100,00
	İBYSA	73,91	73,91	100,00	0,00	0,85	0,00

Tablo 4.38: SMK veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	PDVM	76,79	72,22	89,66	62,96	0,80	85,00
	KNN	73,21	68,42	89,66	55,56	0,78	83,33
	KA	66,07	64,71	75,86	55,56	0,70	68,18
	İBYSA	76,79	72,22	89,66	62,96	0,80	85,00
LDA	PDVM	71,43	66,67	89,66	51,85	0,76	82,35
	KNN	73,21	70,59	82,76	62,96	0,76	77,27
	KA	60,71	62,07	62,07	59,26	0,62	59,26
	İBYSA	67,86	64,86	82,76	51,85	0,73	73,68

Tablo 4.37, Tablo 4.39, Tablo 4.40 ve Tablo 4.41'de TBA ile tüm sınıflandırma yöntemleri için daha yüksek doğruluk oranı sonuçları elde edilmiştir.

Tablo 4.39: Meme kanseri veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	DDVM	65,52	57,89	84,62	50,00	0,69	80,00
	PDVM	62,07	54,17	100,00	31,25	0,70	100,00
	KNN	68,97	70,00	53,85	81,25	0,61	68,42
	KA	72,41	72,73	61,54	81,25	0,67	72,22
	İBYSA	68,97	64,29	69,23	68,75	0,67	73,33
LDA	DDVM	55,17	50,00	61,54	50,00	0,55	61,54
	PDVM	44,83	40,00	46,15	43,75	0,43	50,00
	KNN	55,17	50,00	61,54	50,00	0,55	61,54
	KA	41,38	35,71	38,46	43,75	0,37	46,67
	İBYSA	37,93	33,33	38,46	37,50	0,36	42,86

Tablo 4.38’de PDVM, KA ve İBYSA yöntemleri için TBA sonrasında daha yüksek doğruluk oranı sonuçlarına ulaşılmıştır.

Tablo 4.40: Meme kanseri 2 veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	86,84	91,67	88,00	84,62	0,90	78,57
	PDVM	81,58	80,00	96,00	53,85	0,87	87,50
	KNN	86,84	83,33	100,00	61,54	0,91	100,00
	KA	81,58	84,62	88,00	69,23	0,86	75,00
	İBYSA	89,47	92,00	92,00	84,62	0,92	84,62
LDA	**DDVM**	73,68	80,00	80,00	61,54	0,80	61,54
	PDVM	78,95	81,48	88,00	61,54	0,85	72,73
	KNN	57,89	64,52	80,00	15,38	0,71	28,57
	KA	68,42	70,97	88,00	30,77	0,79	57,14
	İBYSA	71,05	79,17	76,00	61,54	0,78	57,14

Tablo 4.42 ve Tablo 4.43’deki sonuçlara göre, TBA ve LDA için elde edilen sonuçlar benzerlik göstermektedir. İki yöntem arasında doğruluk oranı açısından bir fark görülmemektedir.

Tablo 4.41: CNS veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	77,78	71,43	71,43	81,82	0,71	81,82
	PDVM	72,22	100,00	28,57	100,00	0,44	68,75
	KNN	66,67	66,67	28,57	90,91	0,40	66,67
	KA	61,11	50,00	14,29	90,91	0,22	62,50
	İBYSA	77,78	100,00	42,86	100,00	0,60	73,33
LDA	**DDVM**	55,56	42,86	42,86	63,64	0,43	63,64
	PDVM	61,11	50,00	28,57	81,82	0,36	64,29
	KNN	55,56	33,33	14,29	81,82	0,20	60,00
	KA	55,56	33,33	14,29	81,82	0,20	60,00
	İBYSA	55,56	33,33	14,29	81,82	0,20	60,00

Tablo 4.42: COAD veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	100,00	100,00	100,00	100,00	1,00	100,00
	PDVM	98,45	100,00	98,29	100,00	0,99	85,71
	KNN	98,45	99,15	99,15	91,67	0,99	91,67
	KA	97,67	99,14	98,29	91,67	0,99	84,62
	İBYSA	100,00	100,00	100,00	100,00	1,00	100,00
LDA	**DDVM**	96,90	100,00	96,58	100,00	0,98	75,00
	PDVM	99,22	100,00	99,15	100,00	1,00	92,31
	KNN	99,22	100,00	99,15	100,00	1,00	92,31
	KA	100,00	100,00	100,00	100,00	1,00	100,00
	İBYSA	100,00	100,00	100,00	100,00	1,00	100,00

Tablo 4.43: BRCA veri kümesi için boyut indirgeme sonrası sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %
TBA	**DDVM**	98,90	99,70	99,09	96,97	0,99	91,43
	KNN	96,97	98,19	98,48	81,82	0,98	84,38
	KA	96,14	97,59	98,18	75,76	0,98	80,65
	İBYSA	99,45	99,70	99,70	96,97	1,00	96,97
LDA	**DDVM**	96,69	99,69	96,67	96,97	0,98	74,42
	KNN	99,45	99,70	99,70	96,97	1,00	96,97
	KA	97,80	99,39	98,18	93,94	0,99	83,78
	İBYSA	98,62	99,69	98,79	96,97	0,99	88,89

Tüm veri kümeleri için sonuçlar incelendiğinde; kolon kanseri, prostat kanseri, SMK, meme kanseri, meme kanseri 2 ve CNS veri kümeleri için TBA yöntemiyle boyut indirgeme sonrasındaki sınıflandırma ile sınıflandırma yöntemlerinin çoğunluğu için LDA yöntemine göre daha yüksek doğruluk oranına ulaşıldığı görülmektedir. Buna göre, toplam 11 veri kümesinden altısında TBA ile daha yüksek doğruluk oranı, birinde LDA ile daha yüksek doğruluk oranı, dördünde ise benzer sonuçlar elde edildiği anlaşılmaktadır.

4.4. SARMAL YÖNTEMLERLE SINIFLANDIRMA SONUÇLARI

Bu bölümde; GA-KNN, GA-DVM, PSO-KNN, PSO-DVM sarmal öznitelik seçme yöntemlerinin gen ifade seviyelerini içeren veri kümelerini sınıflandırma sonuçları sunulmaktadır. Veri kümeleri, eğitim ve test için 3 farklı şekilde parçalara ayrılmıştır. Bunlar; veri kümelerinin %70 eğitim seti, %30 test seti olarak parçalandığı yöntem (hold-out), 10 katlamalı çapraz doğrulama (10-fold CV) ve biri hariç çapraz doğrulama (LOOCV) yöntemleridir.

4.4.1. GA İçeren Sarmal Yöntemlerin Sınıflandırma Sonuçları

Tablo 4.44: GA içeren sarmal yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	GA-KNN	100,00	100,00	100,00	100,00	1,00	100,00	171
	GA-DVM	94,44	100,00	83,33	100,00	0,91	92,31	208
10-fold CV	GA-KNN	83,87	80,00	72,73	90,00	0,7619	85,71	171
	GA-DVM	91,67	90,48	86,36	94,74	0,8837	92,31	208
LOOCV	GA-KNN	79,03	71,43	68,18	85,00	0,70	82,93	171
	GA-DVM	87,10	79,17	86,36	87,50	0,83	92,11	208

Tablo 4.44'de kolon kanseri veri kümesi için sarmal yöntemlerin sınıflandırma sonuçları gösterilmektedir.

Tablo 4.45: GA içeren sarmal yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	GA-KNN	90,00	92,86	86,67	93,33	0,90	87,50	4085
	GA-DVM	93,33	100,00	86,67	100,00	0,93	88,24	3931
10-fold CV	GA-KNN	80,00	84,44	74,51	85,71	0,7917	76,36	4064
	GA-DVM	89,00	90,00	88,24	89,80	0,8911	88,00	3931
LOOCV	GA-KNN	72,55	75,00	69,23	76,00	0,72	70,37	4085
	GA-DVM	90,20	90,38	90,38	90,00	0,90	90,00	3931

Tablo 4.46: GA içeren sarmal yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA-KNN**	100,00	100,00	100,00	100,00	1,00	100,00	1853
	GA-DVM	100,00	100,00	100,00	100,00	1,00	100,00	1814
10-fold CV	**GA-KNN**	86,25	100,00	81,67	100,00	0,8991	64,52	1853
	GA-DVM	98,75	100,00	98,33	100,00	0,9916	95,24	1814
LOOCV	**GA-KNN**	87,01	100,00	82,76	100,00	0,91	65,52	1853
	GA-DVM	97,40	98,28	98,28	94,74	0,98	94,74	1814

Tablo 4.47: GA içeren sarmal yöntemlerin SMK veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA-KNN**	76,79	72,22	89,66	62,96	0,80	85,00	7011
	GA-DVM	85,71	78,38	100,00	70,37	0,88	100,00	8850
10-fold CV	**GA-KNN**	61,50	62,63	63,92	58,89	0,6300	60,23	7011
	GA-DVM	57,73	56,76	45,65	68,63	0,5100	58,33	8850
LOOCV	**GA-KNN**	61,50	62,63	63,92	58,89	0,63	60,23	7011
	GA-DVM	47,06	0,00	0,00	97,78	0,00	47,57	8850

Tablo 4.45'de prostat kanseri veri kümesi için, Tablo 4.46'da DLBCL veri kümesi için, Tablo 4.47'de SMK veri kümesi için GA-KNN ve GA-DVM sarmal yöntemlerin sınıflandırma sonuçları gösterilmektedir.

Tablo 4.48: GA içeren sarmal yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	GA-KNN	75,86	87,50	53,85	93,75	0,67	71,43	9080
	GA-DVM	65,52	63,64	53,85	75,00	0,58	66,67	230
10-fold CV	GA-KNN	70,00	69,77	63,83	75,47	0,6667	70,18	8977
	GA-DVM	58,00	60,87	29,79	83,02	0,4000	57,14	9080
LOOCV	GA-KNN	57,73	56,76	45,65	68,63	0,51	58,33	9080
	GA-DVM	68,04	65,96	67,39	68,63	0,67	70,00	230

Tablo 4.48'de meme kanseri veri kümesi için sarmal yöntemlerin sınıflandırma sonuçları gösterilmektedir.

Tablo 4.49: GA içeren sarmal yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	GA-KNN	86,84	83,33	100,00	61,54	0,91	100,00	21338
	GA-DVM	84,21	85,19	92,00	69,23	0,88	81,82	21103
10-fold CV	GA-KNN	81,54	79,25	97,67	50,00	0,8750	91,67	21338
	GA-DVM	83,08	85,56	89,53	70,45	0,8750	77,50	21103
LOOCV	GA-KNN	78,13	81,11	86,90	61,36	0,84	71,05	21338
	GA-DVM	85,94	85,11	95,24	68,18	0,90	88,24	21103

Tablo 4.49'da meme kanseri 2 veri kümesi için sarmal yöntemlerin sınıflandırma sonuçları sunulmaktadır.

Tablo 4.50: GA içeren sarmal yöntemlerin CNS veri kümesi sınıflandırma için sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA-KNN**	83,33	75,00	85,71	81,82	0,80	90,00	1945
	GA-DVM	72,22	66,67	57,14	81,82	0,62	75,00	1807
10-fold CV	**GA-KNN**	65,00	50,00	33,33	82,05	0,4000	69,57	1945
	GA-DVM	73,33	64,71	52,38	84,62	0,5789	76,74	1807
LOOCV	**GA-KNN**	53,33	37,04	47,62	56,41	0,42	66,67	1945
	GA-DVM	71,67	61,11	52,38	82,05	0,56	76,19	1807

Tablo 4.50 ve Tablo 4.51'de sırasıyla CNS ve BRCA veri kümeleri için sarmal yöntemlerin sınıflandırma sonuçları gösterilmektedir.

Tablo 4.51: GA içeren sarmal yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA-KNN**	90,91	93,81	96,36	36,36	0,95	50,00	1892
	GA-DVM	100,00	100,00	100,00	100,00	1,00	100,00	28773
10-fold CV	**GA-KNN**	90,91	93,94	96,17	39,82	0,9505	51,72	1892
	GA-DVM	99,09	99,82	99,18	98,23	0,9950	92,50	28773
LOOCV	**GA-KNN**	89,76	96,20	92,35	64,60	0,94	46,50	1892
	GA-DVM	99,17	99,82	99,27	98,23	1,00	93,28	28773

Tablo 4.52'de COAD veri kümesi için sınıflandırma sonuçları gösterilmektedir.

Tablo 4.52: GA içeren sarmal yöntemlerin COAD veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	GA-KNN	100,00	100,00	100,00	100,00	1,00	100,00	28877
	GA-DVM	99,22	100,00	99,15	100,00	1,00	92,31	28877
10-fold CV	GA-KNN	96,51	97,47	98,72	75,00	0,9809	85,71	28901
	GA-DVM	99,07	99,74	99,23	97,50	0,9949	92,86	28877
LOOCV	GA-KNN	98,14	98,72	99,23	87,50	0,99	92,11	28877
	GA-DVM	99,30	99,74	99,49	97,50	1,00	95,12	28877

4.4.2. PSO İçeren Sarmal Yöntemlerin Sınıflandırma Sonuçları

Tablo 4.53: PSO içeren sarmal yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	PSO-KNN	100,00	100,00	100,00	100,00	1,0000	100,00	898
	PSO-DVM	83,33	80,00	66,67	91,67	0,7273	84,62	984
10-fold CV	PSO-KNN	83,33	80,00	72,73	89,47	0,7619	85,00	898
	PSO-DVM	90,00	86,36	86,36	92,11	0,8636	92,11	984
LOOCV	PSO-KNN	85,00	80,95	77,27	89,47	0,7907	87,18	898
	PSO-DVM	88,33	82,61	86,36	89,47	0,8444	91,89	984

Tablo 4.53'de kolon kanseri veri kümesi için PSO-KNN ve PSO-DVM sarmal yöntemlerin sınıflandırma sonuçları görülmektedir.

Tablo 4.54: PSO içeren sarmal yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	83,33	81,25	86,67	80,00	0,8387	85,71	6203
	PSO-DVM	83,33	85,71	80,00	86,67	0,8276	81,25	6191
10-fold CV	**PSO-KNN**	85,00	90,91	78,43	91,84	0,8421	80,36	6203
	PSO-DVM	91,00	90,38	92,16	89,80	0,9126	91,67	6191
LOOCV	**PSO-KNN**	86,27	91,30	80,77	92,00	0,8571	82,14	6203
	PSO-DVM	93,14	94,12	92,31	94,00	0,9320	92,16	6191

Prostat kanseri ve DLBCL veri kümeleri için PSO tabanlı yöntemlerin sınıflandırma sonuçları sırasıyla Tablo 4.54 ve Tablo 4.55'de gösterilmektedir.

Tablo 4.55: PSO içeren sarmal yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	95,65	94,44	100,00	83,33	0,9714	100,00	3450
	PSO-DVM	91,30	89,47	100,00	66,67	0,9444	100,00	3502
10-fold CV	**PSO-KNN**	91,25	100,00	88,33	100,00	0,9381	74,07	3450
	PSO-DVM	95,00	96,67	96,67	90,00	0,9667	90,00	3502
LOOCV	**PSO-KNN**	92,21	100,00	89,66	100,00	0,9455	76,00	3450
	PSO-DVM	97,40	96,67	100,00	89,47	0,9831	100,00	3502

Tablo 4.56: PSO içeren sarmal yöntemlerin SMK veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	64,29	62,86	75,86	51,85	0,6875	66,67	9947
	PSO-DVM	73,21	69,44	86,21	59,26	0,7692	80,00	9926
10-fold CV	**PSO-KNN**	63,85	63,01	69,70	57,81	0,6619	64,91	9947
	PSO-DVM	60,00	60,61	60,61	59,38	0,6061	59,38	9926
LOOCV	**PSO-KNN**	62,03	61,61	71,13	52,22	0,6603	62,67	9947
	PSO-DVM	74,87	73,15	81,44	67,78	0,7707	77,22	9926

SMK veri kümesi için sınıflandırma sonuçları Tablo 4.56'dadır.

Tablo 4.57: PSO içeren sarmal yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	58,62	54,55	46,15	68,75	0,5000	61,11	23666
	PSO-DVM	51,72	46,67	53,85	50,00	0,5000	57,14	23753
10-fold CV	**PSO-KNN**	57,00	56,67	36,17	75,47	0,4416	57,14	23589
	PSO-DVM	58,00	59,26	34,04	79,25	0,4324	57,53	23753
LOOCV	**PSO-KNN**	53,61	51,72	32,61	72,55	0,4000	54,41	23666
	PSO-DVM	68,04	66,67	65,22	70,59	0,6593	69,23	23753

Meme kanseri ve meme kanseri 2 veri kümeleri için sınıflandırma sonuçları sırasıyla Tablo 4.57 ve 4.58'dedir.

Tablo 4.58: PSO içeren sarmal yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	84,21	80,65	100,00	53,85	0,8929	100,00	12081
	PSO-DVM	84,21	85,19	92,00	69,23	0,8846	81,82	12161
10-fold CV	**PSO-KNN**	80,77	78,50	97,67	47,73	0,8705	91,30	12081
	PSO-DVM	84,62	88,37	88,37	77,27	0,8837	77,27	12161
LOOCV	**PSO-KNN**	81,25	79,41	96,43	52,27	0,8710	88,46	12081
	PSO-DVM	85,94	85,11	95,24	68,18	0,8989	88,24	12161

CNS veri kümesi için PSO-KNN ve PSO-DVM yöntemlerinin sınıflandırma sonuçları Tablo 4.59'dadır.

Tablo 4.59: PSO içeren sarmal yöntemlerin CNS veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	55,56	40,00	28,57	72,73	0,3333	61,54	3504
	PSO-DVM	66,67	60,00	42,86	81,82	0,5000	69,23	3537
10-fold CV	**PSO-KNN**	68,33	56,25	42,86	82,05	0,4865	72,73	3504
	PSO-DVM	70,00	60,00	42,86	84,62	0,5000	73,33	3537
LOOCV	**PSO-KNN**	70,00	58,82	47,62	82,05	0,5263	74,42	3504
	PSO-DVM	73,33	66,67	47,62	87,18	0,5556	75,56	3537

Tablo 4.60: PSO içeren sarmal yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	98,35	98,80	99,39	87,88	0,9909	93,55	30238
	PSO-DVM	100,00	100,00	100,00	100,00	1,0000	100,00	30307
10-fold CV	**PSO-KNN**	97,52	98,63	98,63	86,73	0,9863	86,73	30238
	PSO-DVM	98,93	99,72	99,09	97,35	0,9941	91,67	30307
LOOCV	**PSO-KNN**	97,61	98,64	98,72	86,73	0,9868	87,50	30238
	PSO-DVM	99,09	99,73	99,27	97,35	0,9950	93,22	30307

BRCA ve COAD veri kümeleri için sınıflandırma sonuçları sırasıyla Tablo 4.60 ve Tablo 4.61'de gösterilmektedir.

Tablo 4.61: PSO içeren sarmal yöntemlerin COAD veri kümesi için sınıflandırma sonuçları.

Yöntem		Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO-KNN**	99,22	99,15	100,00	91,67	0,9957	100,00	30181
	PSO-DVM	99,22	100,00	99,15	100,00	0,9957	92,31	30044
10-fold CV	**PSO-KNN**	99,07	98,98	100,00	90,00	0,9949	100,00	30181
	PSO-DVM	99,23	99,17	100,00	90,00	0,9959	100,00	29886
LOOCV	**PSO-KNN**	98,84	98,98	99,74	90,00	0,9936	97,30	30181
	PSO-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	30044

4.5. ÖNERİLEN HİBRİT YÖNTEMLERİN SINIFLANDIRMA SONUÇLARI

Bu bölümde, önerilen yöntemlerin sınıflandırma sonuçları tablolarda sunulmaktadır.

Tablo 4.62: GA tabanlı önerilen yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	100,00	100,00	100,00	100,00	1,00	100,0	5
	GA (Relief-F)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	9
	GA (MI)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	2
	GA (Pearson)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	7
	GA (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	6
	GA (MI)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	3
10-fold CV	**GA (Pearson)-KNN**	88,33	85,71	81,82	92,11	0,84	89,74	3
	GA (Relief-F)-KNN	86,67	88,89	72,73	94,74	0,8000	85,71	1
	GA (MI)-KNN	86,67	88,89	72,73	94,74	0,8000	85,71	1
	GA (Pearson)-DVM	95,00	91,30	95,45	94,74	0,9333	97,30	20
	GA (Relief-F)-DVM	98,33	100,00	95,45	100,00	0,9767	97,44	33
	GA (MI)-DVM	98,33	100,00	95,45	100,00	0,9767	97,44	49
LOOCV	**GA (Pearson)-KNN**	83,87	80,00	72,73	90,00	0,76	85,71	7
	GA (Relief-F)-KNN	87,10	88,89	72,73	95,00	0,8000	86,36	10
	GA (MI)-KNN	85,48	84,21	72,73	92,50	0,7805	86,05	3
	GA (Pearson)-DVM	90,32	86,36	86,36	92,50	0,86	92,50	8
	GA (Relief-F)-DVM	91,94	86,96	90,91	92,50	0,8889	94,87	5
	GA (MI)-DVM	87,10	79,17	86,36	87,50	0,8261	92,11	4

GA ve PSO tabanlı önerilen yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları sırasıyla Tablo 4.62 ve Tablo 4.63'de gösterilmektedir.

Tablo 4.63: PSO tabanlı önerilen yöntemlerin kolon kanseri veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	100,00	100,00	100,00	100,00	1,00	100,0	895
	PSO (Relief-F)-KNN	100,00	100,00	100,00	100,00	1,00	100,0	878
	PSO (MI)-KNN	100,00	100,00	100,00	100,00	1,00	100,0	905
	PSO (Pearson)-DVM	100,00	100,00	100,00	100,00	1,00	100,0	893
	PSO (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,00	100,0	918
	PSO (MI)-DVM	100,00	100,00	100,00	100,00	1,00	100,0	894
10-fold CV	**PSO (Pearson)-KNN**	85,00	80,95	77,27	89,47	0,7907	87,18	967
	PSO (Relief-F)-KNN	88,33	85,71	81,82	92,11	0,8372	89,74	991
	PSO (MI)-KNN	88,33	89,47	77,27	94,74	0,8293	87,80	925
	PSO (Pearson)-DVM	90,00	86,36	86,36	92,11	0,8636	92,11	927
	PSO (Relief-F)-DVM	88,33	85,71	81,82	92,11	0,8372	89,74	994
	PSO (MI)-DVM	93,33	90,91	90,91	94,74	0,9091	94,74	952
LOOCV	**PSO (Pearson)-KNN**	82,26	78,95	68,18	90,00	0,7317	83,72	859
	PSO (Relief-F)-KNN	87,10	88,89	72,73	95,00	0,8000	86,36	881
	PSO (MI)-KNN	88,71	89,47	77,27	95,00	0,8293	88,37	877
	PSO (Pearson)-DVM	80,65	75,00	68,18	87,50	0,7143	83,33	922
	PSO (Relief-F)-DVM	87,10	85,00	77,27	92,50	0,8095	88,10	886
	PSO (MI)-DVM	85,48	78,26	81,82	87,50	0,8000	89,74	898

Prostat kanseri veri kümesi için sonuçlar Tablo 4.64 ve Tablo 4.65'de sunulmaktadır.

Tablo 4.64: GA tabanlı önerilen yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	90,00	92,86	86,67	93,33	0,8966	87,50	408
	GA (Relief-F)-KNN	93,33	100,00	86,67	100,00	0,9286	88,24	170
	GA (MI)-KNN	93,33	100,00	86,67	100,00	0,9286	88,24	208
	GA (Pearson)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	84
	GA (Relief-F)-DVM	96,67	100,00	93,33	100,00	0,9655	93,75	918
	GA (MI)-DVM	96,67	100,00	93,33	100,00	0,9655	93,75	101
10-fold CV	**GA (Pearson)-KNN**	94,00	94,12	94,12	93,88	0,9412	93,88	8
	GA (Relief-F)-KNN	86,00	84,91	88,24	83,67	0,8654	87,23	1
	GA (MI)-KNN	86,00	84,91	88,24	83,67	0,8654	87,23	1
	GA (Pearson)-DVM	97,00	98,00	96,08	97,96	0,9703	96,00	275
	GA (Relief-F)-DVM	99,00	100,00	98,04	100,00	0,9901	98,00	210
	GA (MI)-DVM	97,00	98,00	96,08	97,96	0,9703	96,00	437
LOOCV	**GA (Pearson)-KNN**	79,41	82,98	75,00	84,00	0,7900	76,36	399
	GA (Relief-F)-KNN	88,24	95,45	80,77	96,00	0,8750	82,76	180
	GA (MI)-KNN	93,33	100,00	86,67	100,00	0,9286	88,24	210
	GA (Pearson)-DVM	89,22	87,27	92,31	86,00	0,9000	91,49	89
	GA (Relief-F)-DVM	93,14	94,12	92,31	94,00	0,9320	92,16	912
	GA (MI)-DVM	94,12	94,23	94,23	94,00	0,9423	94,00	108

Tablo 4.65: PSO tabanlı önerilen yöntemlerin prostat kanseri veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	90,00	92,86	86,67	93,33	0,90	87,50	6151
	PSO (Relief-F)-KNN	96,67	100,00	93,33	100,00	0,97	93,75	6110
	PSO (MI)-KNN	90,00	92,86	86,67	93,33	0,90	87,50	6140
	PSO (Pearson)-DVM	93,33	100,00	86,67	100,00	0,93	88,24	5961
	PSO (Relief-F)-DVM	90,00	92,86	86,67	93,33	0,90	87,50	6181
	PSO (MI)-DVM	96,67	100,00	93,33	100,00	0,97	93,75	6009
10-fold CV	**PSO (Pearson)-KNN**	87,00	89,58	84,31	89,80	0,8687	84,62	6261
	PSO (Relief-F)-KNN	84,00	85,71	82,35	85,71	0,8400	82,35	6203
	PSO (MI)-KNN	85,00	92,86	76,47	93,88	0,8387	79,31	6177
	PSO (Pearson)-DVM	94,00	94,12	94,12	93,88	0,9412	93,88	6135
	PSO (Relief-F)-DVM	95,00	97,92	92,16	97,96	0,9495	92,31	6156
	PSO (MI)-DVM	95,00	96,00	94,12	95,92	0,9505	94,00	6227
LOOCV	**PSO (Pearson)-KNN**	78,43	89,47	65,38	92,00	0,7556	71,88	6138
	PSO (Relief-F)-KNN	83,33	90,70	75,00	92,00	0,8211	77,97	6140
	PSO (MI)-KNN	83,33	90,70	75,00	92,00	0,8211	77,97	6150
	PSO (Pearson)-DVM	94,12	94,23	94,23	94,00	0,9423	94,00	6115
	PSO (Relief-F)-DVM	94,12	94,23	94,23	94,00	0,9423	94,00	6023
	PSO (MI)-DVM	91,18	92,16	90,38	92,00	0,9126	90,20	5950

Tablo 4.66: GA tabanlı önerilen yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	100,00	100,00	100,00	100,00	1,0000	100,0	69
	GA (Relief-F)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	86
	GA (MI)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	33
	GA (Pearson)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	193
	GA (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	68
	GA (MI)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	98
10-fold CV	**GA (Pearson)-KNN**	96,25	98,31	96,67	95,00	0,9748	90,48	2
	GA (Relief-F)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	82
	GA (MI)-KNN	97,50	98,33	98,33	95,00	0,9833	95,00	2
	GA (Pearson)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	101
	GA (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	214
	GA (MI)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	38
LOOCV	**GA (Pearson)-KNN**	90,91	94,74	93,10	84,21	0,9400	80,00	58
	GA (Relief-F)-KNN	93,51	96,49	94,83	89,47	0,9565	85,00	89
	GA (MI)-KNN	96,10	98,25	96,55	94,74	0,9739	90,00	38
	GA (Pearson)-DVM	97,40	96,67	100,00	89,47	0,9800	100,0	188
	GA (Relief-F)-DVM	98,70	100,00	98,28	100,00	0,9913	95,00	49
	GA (MI)-DVM	97,40	98,28	98,28	94,74	0,9828	94,74	112

DLBCL veri kümesi için sonuçlar, Tablo 4.66 ve Tablo 4.67'de gösterilmektedir.

Tablo 4.67: PSO tabanlı önerilen yöntemlerin DLBCL veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	100,00	100,00	100,00	100,00	1,00	100,0	3358
	PSO (Relief-F)-KNN	100,00	100,00	100,00	100,00	1,00	100,0	3349
	PSO (MI)-KNN	100,00	100,00	100,00	100,00	1,00	100,0	3320
	PSO (Pearson)-DVM	100,00	100,00	100,00	100,00	1,00	100,0	3350
	PSO (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,00	100,0	3360
	PSO (MI)-DVM	100,00	100,00	100,00	100,00	1,00	100,0	3323
10-fold CV	**PSO (Pearson)-KNN**	90,00	98,15	88,33	95,00	0,9298	73,08	3537
	PSO (Relief-F)-KNN	92,50	98,21	91,67	95,00	0,9483	79,17	3468
	PSO (MI)-KNN	93,75	100,00	91,67	100,00	0,9565	80,00	3499
	PSO (Pearson)-DVM	96,25	98,31	96,67	95,00	0,9748	90,48	3444
	PSO (Relief-F)-DVM	93,75	96,61	95,00	90,00	0,9580	85,71	3394
	PSO (MI)-DVM	97,50	98,33	98,33	95,00	0,9833	95,00	3468
LOOCV	**PSO (Pearson)-KNN**	88,31	98,04	86,21	94,74	0,9174	69,23	3362
	PSO (Relief-F)-KNN	90,91	100,00	87,93	100,00	0,9358	73,08	3355
	PSO (MI)-KNN	87,01	98,00	84,48	94,74	0,9074	66,67	3360
	PSO (Pearson)-DVM	97,40	98,28	98,28	94,74	0,9828	94,74	3380
	PSO (Relief-F)-DVM	98,70	98,31	100,00	94,74	0,9915	100,0	3363
	PSO (MI)-DVM	98,70	98,31	100,00	94,74	0,9915	100,0	3365

Tablo 4.68: GA tabanlı önerilen yöntemlerin SMK veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	80,36	73,68	96,55	62,96	0,8358	94,44	73
	GA (Relief-F)-KNN	83,93	76,32	100,00	66,67	0,8657	100,0	114
	GA (MI)-KNN	83,93	76,32	100,00	66,67	0,8657	100,0	221
	GA (Pearson)-DVM	87,50	80,56	100,00	74,07	0,8923	100,0	842
	GA (Relief-F)-DVM	87,50	86,67	89,66	85,19	0,8814	88,46	1375
	GA (MI)-DVM	92,86	87,88	100,00	85,19	0,9355	100,0	419
10-fold CV	**GA (Pearson)-KNN**	68,45	68,27	73,20	63,33	0,7100	68,67	73
	GA (Relief-F)-KNN	68,98	66,96	79,38	57,78	0,7264	72,22	114
	GA (MI)-KNN	70,59	69,09	78,35	62,22	0,7343	72,73	221
	GA (Pearson)-DVM	79,68	81,05	79,38	80,00	0,8000	78,26	842
	GA (Relief-F)-DVM	78,61	77,14	83,51	73,33	0,8020	80,49	1375
	GA (MI)-DVM	80,75	80,20	83,51	77,78	0,8182	81,40	419
LOOCV	**GA (Pearson)-KNN**	68,45	68,27	73,20	63,33	0,7100	68,67	73
	GA (Relief-F)-KNN	68,98	66,96	79,38	57,78	0,7264	72,22	114
	GA (MI)-KNN	70,59	69,09	78,35	62,22	0,7343	72,73	221
	GA (Pearson)-DVM	79,68	81,05	79,38	80,00	0,8000	78,26	842
	GA (Relief-F)-DVM	78,61	77,14	83,51	73,33	0,8020	80,49	1375
	GA (MI)-DVM	80,75	80,20	83,51	77,78	0,8182	81,40	419

Önerilen yöntemlerin SMK veri kümesi için sonuçları Tablo 4.68 ve Tablo 4.69'da gösterilmektedir.

Tablo 4.69: PSO tabanlı önerilen yöntemlerin SMK veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	76,79	71,05	93,10	59,26	0,8059	88,89	9853
	PSO (Relief-F)-KNN	76,79	72,22	89,66	62,96	0,8000	85,00	9860
	PSO (MI)-KNN	76,79	71,05	93,10	59,26	0,8059	88,89	9731
	PSO (Pearson)-DVM	82,14	78,79	89,66	74,07	0,8387	86,96	9892
	PSO (Relief-F)-DVM	83,93	79,41	93,10	74,07	0,8571	90,91	9850
	PSO (MI)-DVM	82,14	78,79	89,66	74,07	0,8387	86,96	9820
10-fold CV	**PSO (Pearson)-KNN**	63,85	63,01	69,70	57,81	0,6619	64,91	10002
	PSO (Relief-F)-KNN	70,77	68,42	78,79	62,50	0,7324	74,07	9981
	PSO (MI)-KNN	66,92	65,33	74,24	59,38	0,6950	69,09	9987
	PSO (Pearson)-DVM	67,69	66,67	72,73	62,50	0,6957	68,97	9981
	PSO (Relief-F)-DVM	68,46	67,61	72,73	64,06	0,7007	69,49	9891
	PSO (MI)-DVM	70,00	69,01	74,24	65,63	0,7153	71,19	10020
LOOCV	**PSO (Pearson)-KNN**	64,71	63,96	73,20	55,56	0,6827	65,79	9855
	PSO (Relief-F)-KNN	63,64	62,83	73,20	53,33	0,6762	64,86	9868
	PSO (MI)-KNN	64,17	64,15	70,10	57,78	0,6700	64,20	9901
	PSO (Pearson)-DVM	74,33	72,90	80,41	67,78	0,7647	76,25	9899
	PSO (Relief-F)-DVM	77,01	75,00	83,51	70,00	0,7902	79,75	9863
	PSO (MI)-DVM	73,80	72,22	80,41	66,67	0,7610	75,95	9828

Tablo 4.70: GA tabanlı önerilen yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	75,86	80,00	61,54	87,50	0,6957	73,68	110
	GA (Relief-F)-KNN	86,21	84,62	84,62	87,50	0,8462	87,50	7
	GA (MI)-KNN	93,10	100,00	84,62	100,00	0,9167	88,89	377
	GA (Pearson)-DVM	91,25	93,44	95,00	80,00	0,9421	84,21	12
	GA (Relief-F)-DVM	93,75	98,25	93,33	95,00	0,9573	82,61	529
	GA (MI)-DVM	96,25	98,31	96,67	95,00	0,9748	90,48	339
10-fold CV	**GA (Pearson)-KNN**	77,00	77,27	72,34	81,13	0,7473	76,79	22
	GA (Relief-F)-KNN	80,00	82,98	76,47	83,67	0,7959	77,36	16
	GA (MI)-KNN	82,00	82,35	82,35	81,63	0,8236	81,63	12
	GA (Pearson)-DVM	87,00	93,18	80,39	93,88	0,8632	82,14	32
	GA (Relief-F)-DVM	88,00	84,21	94,12	81,63	0,8889	93,02	256
	GA (MI)-DVM	89,23	91,86	91,86	84,09	0,9186	84,09	318
LOOCV	**GA (Pearson)-KNN**	75,00	79,55	68,63	81,63	0,7368	71,43	102
	GA (Relief-F)-KNN	77,00	76,92	78,43	75,51	0,7767	77,08	9
	GA (MI)-KNN	79,00	80,00	78,43	79,59	0,7921	78,00	287
	GA (Pearson)-DVM	86,92	88,76	91,86	77,27	0,9029	82,93	17
	GA (Relief-F)-DVM	87,00	91,30	82,35	91,84	0,8660	83,33	523
	GA (MI)-DVM	88,00	95,35	80,39	95,92	0,8723	82,46	432

Meme kanseri veri kümesi için sonuçlar Tablo 4.70 ve 4.71'de sunulmaktadır.

Tablo 4.71: PSO tabanlı önerilen yöntemlerin meme kanseri veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	85,00	86,00	84,31	85,71	0,8515	84,00	23138
	PSO (Relief-F)-KNN	85,00	89,13	80,39	89,80	0,8454	81,48	23003
	PSO (MI)-KNN	89,00	90,00	88,24	89,80	0,8911	88,00	23023
	PSO (Pearson)-DVM	87,00	85,19	90,20	83,67	0,8762	89,13	22745
	PSO (Relief-F)-DVM	87,00	86,54	88,24	85,71	0,8738	87,50	22715
	PSO (MI)-DVM	86,00	86,27	86,27	85,71	0,8627	85,71	22851
10-fold CV	**PSO (Pearson)-KNN**	75,00	77,08	72,55	77,55	0,7475	73,08	23314
	PSO (Relief-F)-KNN	77,00	76,92	78,43	75,51	0,7767	77,08	23286
	PSO (MI)-KNN	78,00	77,36	80,39	75,51	0,7885	78,72	23512
	PSO (Pearson)-DVM	86,00	87,76	84,31	87,76	0,8600	84,31	23423
	PSO (Relief-F)-DVM	86,00	81,36	94,12	77,55	0,8727	92,68	23346
	PSO (MI)-DVM	87,00	88,00	86,27	87,76	0,8713	86,00	23816
LOOCV	**PSO (Pearson)-KNN**	72,41	85,71	46,15	93,75	0,6000	68,18	23223
	PSO (Relief-F)-KNN	72,41	85,71	46,15	93,75	0,6000	68,18	23012
	PSO (MI)-KNN	75,86	87,50	53,85	93,75	0,6700	71,43	23117
	PSO (Pearson)-DVM	82,00	85,11	78,43	85,71	0,8163	79,25	22965
	PSO (Relief-F)-DVM	84,00	84,31	84,31	83,67	0,8431	83,67	22864
	PSO (MI)-DVM	85,00	89,13	80,39	89,80	0,8454	81,48	22915

Tablo 4.72: GA tabanlı önerilen yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	86,84	85,71	96,00	69,23	0,9057	90,00	183
	GA (Relief-F)-KNN	92,11	92,31	96,00	84,62	0,9412	91,67	24
	GA (MI)-KNN	89,47	86,21	100,00	69,23	0,9259	100,0	274
	GA (Pearson)-DVM	94,74	96,00	96,00	92,31	0,9600	92,31	216
	GA (Relief-F)-DVM	94,74	92,59	100,00	84,62	0,9615	100,0	227
	GA (MI)-DVM	94,74	92,59	100,00	84,62	0,9615	100,0	396
10-fold CV	**GA (Pearson)-KNN**	90,00	91,01	94,19	81,82	0,9257	87,80	304
	GA (Relief-F)-KNN	85,38	86,02	93,02	70,45	0,8939	83,78	24
	GA (MI)-KNN	83,85	84,95	91,86	68,18	0,8827	81,08	274
	GA (Pearson)-DVM	89,23	90,00	94,19	79,55	0,9205	87,50	216
	GA (Relief-F)-DVM	90,77	91,11	95,35	81,82	0,9318	90,00	227
	GA (MI)-DVM	90,00	90,11	95,35	79,55	0,9266	89,74	396
LOOCV	**GA (Pearson)-KNN**	86,72	91,36	88,10	84,09	0,9000	78,72	183
	GA (Relief-F)-KNN	84,38	85,56	91,67	70,45	0,8851	81,58	24
	GA (MI)-KNN	82,81	82,29	94,05	61,36	0,8778	84,38	274
	GA (Pearson)-DVM	87,50	89,53	91,67	79,55	0,9100	83,33	216
	GA (Relief-F)-DVM	88,28	89,66	92,86	79,55	0,9123	85,37	227
	GA (MI)-DVM	88,28	88,76	94,05	77,27	0,9133	87,18	396

Meme kanseri 2 veri kümesi için sonuçlar Tablo 4.72 ve 4.73'de gösterilmektedir.

Tablo 4.73: PSO tabanlı önerilen yöntemlerin meme kanseri 2 veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	86,84	83,33	100,00	61,54	0,9091	100,0	11988
	PSO (Relief-F)-KNN	86,84	83,33	100,00	61,54	0,9091	100,0	12095
	PSO (MI)-KNN	86,84	83,33	100,00	61,54	0,9091	100,0	12086
	PSO (Pearson)-DVM	84,21	85,19	92,00	69,23	0,8846	81,82	12249
	PSO (Relief-F)-DVM	84,21	85,19	92,00	69,23	0,8846	81,82	12041
	PSO (MI)-DVM	84,21	85,19	92,00	69,23	0,8846	81,82	12085
10-fold CV	**PSO (Pearson)-KNN**	85,38	82,52	98,84	59,09	0,8995	96,30	12303
	PSO (Relief-F)-KNN	83,85	81,55	97,67	56,82	0,8889	92,59	12101
	PSO (MI)-KNN	84,62	81,73	98,84	56,82	0,8947	96,15	12192
	PSO (Pearson)-DVM	82,31	85,39	88,37	70,45	0,8686	75,61	12284
	PSO (Relief-F)-DVM	85,38	85,26	94,19	68,18	0,8950	85,71	12242
	PSO (MI)-DVM	86,15	86,17	94,19	70,45	0,9000	86,11	12214
LOOCV	**PSO (Pearson)-KNN**	82,03	80,20	96,43	54,55	0,8757	88,88	12205
	PSO (Relief-F)-KNN	82,03	79,61	97,62	52,27	0,8770	92,00	12195
	PSO (MI)-KNN	82,81	80,39	97,62	54,55	0,8817	92,30	12095
	PSO (Pearson)-DVM	86,72	86,02	95,24	70,46	0,9040	88,57	12183
	PSO (Relief-F)-DVM	85,94	85,11	95,24	68,18	0,8989	88,23	12262
	PSO (MI)-DVM	86,72	86,02	95,24	70,46	0,9040	88,57	12046

Tablo 4.74: GA tabanlı önerilen yöntemlerin CNS veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	88,89	100,00	71,43	100,00	0,8333	84,62	36
	GA (Relief-F)-KNN	94,44	87,50	100,00	90,91	0,9333	100,0	23
	GA (MI)-KNN	94,44	100,00	85,71	100,00	0,9231	91,67	34
	GA (Pearson)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	334
	GA (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	47
	GA (MI)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	25
10-fold CV	**GA (Pearson)-KNN**	76,67	70,59	57,14	87,18	0,6316	79,07	3
	GA (Relief-F)-KNN	81,67	77,78	66,67	89,74	0,7179	83,33	196
	GA (MI)-KNN	73,33	66,67	47,62	87,18	0,5556	75,56	1
	GA (Pearson)-DVM	93,33	94,74	85,71	97,44	0,9000	92,68	28
	GA (Relief-F)-DVM	95,00	100,00	85,71	100,00	0,9231	92,86	100
	GA (MI)-DVM	90,00	89,47	80,95	94,87	0,8500	90,24	119
LOOCV	**GA (Pearson)-KNN**	71,67	59,09	61,90	76,92	0,6000	78,95	36
	GA (Relief-F)-KNN	76,67	66,67	66,67	82,05	0,6667	82,05	23
	GA (MI)-KNN	73,33	63,16	57,14	82,05	0,6000	78,05	34
	GA (Pearson)-DVM	85,00	87,50	66,67	94,87	0,7600	84,09	334
	GA (Relief-F)-DVM	83,33	78,95	71,43	89,74	0,7500	85,37	47
	GA (MI)-DVM	95,00	90,91	95,24	94,87	0,9302	97,37	25

CNS veri kümesi için sonuçlar Tablo 4.74 ve 4.75’de sunulmaktadır.

Tablo 4.75: PSO tabanlı önerilen yöntemlerin CNS veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	77.78	80.00	57.14	90.91	0.67	76.92	3466
	PSO (Relief-F)-KNN	77.78	71.43	71.43	81.82	0.71	81.82	3510
	PSO (MI)-KNN	77.78	71.43	71.43	81.82	0.71	81.82	3488
	PSO (Pearson)-DVM	77.78	71.43	71.43	81.82	0.71	81.82	3434
	PSO (Relief-F)-DVM	72.22	66.67	57.14	81.82	0.62	75.00	3318
	PSO (MI)-DVM	77.78	71.43	71.43	81.82	0.71	81.82	3462
10-fold CV	**PSO (Pearson)-KNN**	73,33	66,67	47,62	87,18	0,5556	75,56	3502
	PSO (Relief-F)-KNN	65,00	50,00	47,62	74,36	0,4878	72,50	3501
	PSO (MI)-KNN	70,00	58,82	47,62	82,05	0,5263	74,42	3539
	PSO (Pearson)-DVM	71,67	62,50	47,62	84,62	0,5405	75,00	3493
	PSO (Relief-F)-DVM	71,67	61,11	52,38	82,05	0,5641	76,19	3565
	PSO (MI)-DVM	75,00	66,67	57,14	84,62	0,6154	78,57	3622
LOOCV	**PSO (Pearson)-KNN**	63,33	47,06	38,10	76,92	0,4211	69,77	3462
	PSO (Relief-F)-KNN	63,33	46,15	28,57	82,05	0,3529	68,09	3514
	PSO (MI)-KNN	61,67	44,44	38,10	74,36	0,4103	69,05	3483
	PSO (Pearson)-DVM	65,00	50,00	47,62	74,36	0,4878	72,50	3430
	PSO (Relief-F)-DVM	63,33	47,37	42,86	74,36	0,4500	70,73	3312
	PSO (MI)-DVM	70,00	60,00	42,86	84,62	0,5000	73,33	3460

Tablo 4.76: GA tabanlı önerilen yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	98,07	98,50	99,39	84,85	0,9894	93,33	8
	GA (Relief-F)-KNN	98,62	99,09	99,39	90,91	0,9924	93,75	30
	GA (MI)-KNN	99,72	99,70	100,00	96,97	0,9985	100,0	433
	GA (Pearson)-DVM	99,72	99,70	100,00	96,97	0,9985	100,0	529
	GA (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	625
	GA (MI)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	629
10-fold CV	**GA (Pearson)-KNN**	97,44	98,28	98,91	83,19	0,9859	88,68	8
	GA (Relief-F)-KNN	98,18	99,09	98,91	91,15	0,9900	89,57	30
	GA (MI)-KNN	98,68	99,27	99,27	92,92	0,9927	92,92	433
	GA (Pearson)-DVM	99,01	99,73	99,18	97,35	0,9945	92,44	529
	GA (Relief-F)-DVM	99,34	100,00	99,27	100,00	0,9963	93,39	625
	GA (MI)-DVM	99,01	99,63	99,27	96,46	0,9945	93,16	629
LOOCV	**GA (Pearson)-KNN**	97,44	98,19	99,00	82,30	0,9900	89,42	8
	GA (Relief-F)-KNN	98,02	99,09	98,72	91,15	0,9891	88,03	30
	GA (MI)-KNN	98,27	99,18	98,91	92,04	0,9904	89,66	433
	GA (Pearson)-DVM	99,01	99,63	99,27	96,46	0,9900	93,16	529
	GA (Relief-F)-DVM	99,17	100,00	99,09	100,00	0,9954	91,87	625
	GA (MI)-DVM	98,84	99,63	99,09	96,46	0,9936	91,60	629

BRCA veri kümesi için sonuçlar Tablo 4.76 ve 4.77'de gösterilmektedir.

Tablo 4.77: PSO tabanlı önerilen yöntemlerin BRCA veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	98.62	99.39	99.09	93.94	0.99	91.18	30307
	PSO (Relief-F)-KNN	98.90	99.10	99.70	90.91	0.99	96.77	30276
	PSO (MI)-KNN	98.35	99.09	99.09	90.91	0.99	90.91	30081
	PSO (Pearson)-DVM	100.00	100.00	100.00	100.00	1.00	100.0	29843
	PSO (Relief-F)-DVM	100.00	100.00	100.00	100.00	1.00	100.0	29622
	PSO (MI)-DVM	100.00	100.00	100.00	100.00	1.00	100.0	29772
10-fold CV	**PSO (Pearson)-KNN**	97,77	98,72	98,81	87,61	0,9877	88,39	30135
	PSO (Relief-F)-KNN	97,44	99,08	98,09	91,15	0,9858	83,06	30180
	PSO (MI)-KNN	97,60	99,35	97,99	93,81	0,9867	82,81	30200
	PSO (Pearson)-DVM	99,42	99,91	99,45	99,12	0,9968	94,92	30052
	PSO (Relief-F)-DVM	99,17	99,82	99,27	98,23	0,9954	93,28	29941
	PSO (MI)-DVM	99,09	99,82	99,18	98,23	0,9950	92,50	29860
LOOCV	**PSO (Pearson)-KNN**	97,77	99,17	98,36	92,04	0,9877	85,25	30305
	PSO (Relief-F)-KNN	97,11	98,72	98,09	87,61	0,9840	82,50	30274
	PSO (MI)-KNN	96,78	98,80	97,63	88,50	0,9821	79,37	30083
	PSO (Pearson)-DVM	99,01	99,73	99,18	97,35	0,9945	92,44	29841
	PSO (Relief-F)-DVM	99,26	99,91	99,27	99,12	0,9959	93,33	29620
	PSO (MI)-DVM	99,09	99,82	99,18	98,23	0,9950	92,50	29770

Tablo 4.78: GA tabanlı önerilen yöntemlerin COAD veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**GA (Pearson)-KNN**	100,00	100,00	100,00	100,00	1,0000	100,0	88
	GA (Relief-F)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	38
	GA (MI)-KNN	100,00	100,00	100,00	100,00	1,0000	100,0	1
	GA (Pearson)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	219
	GA (Relief-F)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	124
	GA (MI)-DVM	100,00	100,00	100,00	100,00	1,0000	100,0	1
10-fold CV	**GA (Pearson)-KNN**	98,37	98,98	99,23	90,00	0,9910	92,31	19
	GA (Relief-F)-KNN	97,91	98,47	99,23	85,00	0,9885	91,89	2
	GA (MI)-KNN	98,14	98,97	98,97	90,00	0,9897	90,00	1
	GA (Pearson)-DVM	99,30	99,49	99,74	95,00	0,9962	97,44	219
	GA (Relief-F)-DVM	98,84	99,49	99,23	95,00	0,9936	92,68	124
	GA (MI)-DVM	97,91	98,47	99,23	85,00	0,9885	91,89	1
LOOCV	**GA (Pearson)-KNN**	98,37	99,23	98,97	92,50	0,9900	90,24	88
	GA (Relief-F)-KNN	99,30	99,49	99,74	95,00	0,9962	97,44	38
	GA (MI)-KNN	98,60	98,48	100,00	85,00	0,9924	100,0	1
	GA (Pearson)-DVM	99,30	99,49	99,74	95,00	0,9962	97,44	219
	GA (Relief-F)-DVM	98,84	99,49	99,23	95,00	0,9936	92,68	124
	GA (MI)-DVM	97,91	98,47	99,23	85,00	0,9885	91,89	1

COAD veri kümesi için sonuçlar Tablo 4.78 ve 4.79'da gösterilmektedir.

Tablo 4.79: PSO tabanlı önerilen yöntemlerin COAD veri kümesi için sınıflandırma sonuçları.

	Yöntem	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen Sayısı
Hold-out	**PSO (Pearson)-KNN**	100.00	100.00	100.00	100.00	1.00	100.0	29896
	PSO (Relief-F)-KNN	100.00	100.00	100.00	100.00	1.00	100.0	29954
	PSO (MI)-KNN	100.00	100.00	100.00	100.00	1.00	100.0	29844
	PSO (Pearson)-DVM	99.22	100.00	99.15	100.00	1.00	92.31	29766
	PSO (Relief-F)-DVM	99.22	100.00	99.15	100.00	1.00	92.31	29720
	PSO (MI)-DVM	99.22	100.00	99.15	100.00	1.00	92.31	29695
10-fold CV	**PSO (Pearson)-KNN**	99,07	98,98	100,00	90,00	0,9949	100,0	29996
	PSO (Relief-F)-KNN	98,60	98,48	100,00	85,00	0,9924	100,0	30112
	PSO (MI)-KNN	98,37	98,48	99,74	85,00	0,9911	97,14	30018
	PSO (Pearson)-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	29936
	PSO (Relief-F)-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	29784
	PSO (MI)-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	29958
LOOCV	**PSO (Pearson)-KNN**	98,14	98,97	98,97	90,00	0,9897	90,00	29892
	PSO (Relief-F)-KNN	98,60	98,48	100,00	85,00	0,9924	100,0	29950
	PSO (MI)-KNN	97,67	98,22	99,23	82,50	0,9872	91,67	29840
	PSO (Pearson)-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	29764
	PSO (Relief-F)-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	29725
	PSO (MI)-DVM	99,30	99,74	99,49	97,50	0,9961	95,12	29705

5. TARTIŞMA VE SONUÇ

Bu çalışmada, gen ifade seviyelerini içeren veri kümelerinden kanser teşhisi ve sınıflandırması için yeni hibrit yöntemler önerilmiştir. Ayrıca, başlıca sınıflandırma yöntemlerinin, öznitelik seçimi için kullanılan başlıca filtre yöntemlerin, başlıca boyut indirgeme yönetmelerinin, GA-DVM, GA-KNN, PSO-DVM ve PSO-KNN sarmal yöntemlerinin bu veri kümelerini sınıflandırma sonuçları deneysel olarak elde edilerek karşılaştırılmıştır. Önerilen hibrit yöntemlerin sınıflandırma sonuçları elde edilen bu sonuçlarla ve literatürdeki diğer çalışmalarla karşılaştırılmıştır. Buna göre, tüm veri kümeleri için tüm değerlendirme ölçütleri açısından en başarılı sonuçların önerilen hibrit yöntemlerle elde edildiği görülmüştür.

Gen ifade veri kümelerinin başlıca sınıflandırma yöntemleri ile sınıflandırma sonuçları tüm değerlendirme ölçütleri için Bulgular 4.1'de gösterilmiştir. Tablo 5.1'de tüm veri kümeleri için sınıflandırma yöntemlerinin doğruluk oranı sonuçları özetlenmektedir. Bu sonuçlar k=10 için çapraz doğrulama yapıldığında elde edilen sonuçlardır.

Tablo 5.1: Başlıca sınıflandırma yöntemleri için doğruluk oranı (%) sonuçları (CV).

Veri Kümesi	DDVM	PDVM	KNN	NB	KA	İBYSA
Lösemi	98,61	65,28	80,56	98,61	80,56	98,61
Kolon Kanseri	80,65	64,52	75,81	64,52	59,68	83,87
Prostat Kanseri	91,18	50,98	83,33	63,73	84,31	90,20
Over Kanseri	100,00	64,03	93,28	90,00	98,02	99,60
DLBCL	96,10	75,32	90,91	81,82	87,01	96,10
SMK	72,73	49,20	66,31	60,96	62,03	71,66
Meme Kanseri	85,16	65,63	84,38	85,94	82,81	83,59
Meme Kanseri 2	68,04	52,58	51,55	51,00	59,79	67,01
CNS	63,33	65,00	63,33	63,33	53,33	65,00

Over kanseri, lösemi ve DLBCL gibi birkaç veri kümesi dışındaki veri kümelerinin öznitelik seçimi öncesi sınıflandırma sonuçlarının düşük olduğu anlaşılmaktadır. En yüksek sonuçların DDVM, İBYSA ve KNN yöntemleriyle elde edildiği görülmektedir.

Filtre yöntemlerle öznitelik seçiminin sınıflandırma sonuçlarına etkisi Bulgular 4.2'de gösterilmektedir. Filtre yöntemlerle öznitelik seçimi gerçekleştirildiğinde tüm değerlendirme ölçütleri açısından iyileşme elde edildiği görülmüştür. Filtre yöntemlerin başarısı veri kümesine göre değişmekle birlikte Pearson Korelasyonu, Relief-F ve Karşılıklı Bilgi yöntemleriyle seçilen genlerden diğerlerine göre daha yüksek doğruluk oranı sonuçlarına ulaşıldığı görülmektedir. Sınıflandırma yöntemleri; DVM, KNN, İBYSA yöntemlerinin filtre yöntemlerle birlikte kullanılmasıyla diğer yöntemlere göre daha başarılı sonuçlar elde edilmiştir. Ayrıca, seçilen gen sayısına göre doğruluk oranı değişimini gösteren grafiklerden bazı genlerin birlikte kullanılmasının başarıyı düşürdüğü anlaşılmaktadır.

TBA ve LDA ile boyut indirgeme sonrası sınıflandırma sonuçları Bulgular 4.3'de gösterilmektedir. Buna göre, veri kümelerinin çoğunluğu için TBA yöntemiyle boyut indirgeme sonrası sınıflandırma ile LDA yöntemine göre daha yüksek doğruluk oranı sonuçları elde edildiği görülmektedir.

Sarmal yöntemlerin ve önerilen hibrit yöntemlerin tüm değerlendirme ölçütleri için sınıflandırma sonuçları sırasıyla Bulgular 4.4'de ve Bulgular 4.5'de gösterilmektedir. Tablo 5.2'de sarmal ve önerilen yöntemlerin doğruluk oranı sonuçları özetlenmektedir. Bu sonuçlar, k=10 için çapraz doğrulama yapılarak elde edilen sonuçlardır. Buna göre, tüm veri kümeleri için önerilen hibrit yöntemlerle en yüksek doğruluk oranı sonuçlarına ulaşıldığı görülmektedir. Ayrıca, Bulgular 4.5'de gösterildiği gibi tüm değerlendirme ölçütleri açısından en iyi sonuçlar önerilen yöntemlerle elde edilmiştir.

Önerilen GA(ReliefF)-DVM yöntemiyle kolon kanseri, prostat kanseri, DLBCL, meme kanseri 2 (M2), CNS veri kümeleri için en yüksek doğruluk oranı sonuçlarına ulaşılmıştır. GA tabanlı ve DVM içeren tüm önerilen yöntemlerle DLBCL veri kümesi için %100 doğruluk oranı elde edilmiştir. GA(ReliefF)-DVM yöntemi, bu veri kümesinde gen sayısını 82'ye, diğer iki yöntem 2'ye indirgemiştir. SMK ve meme kanseri (M1) veri kümeleri için GA(MI)-DVM yöntemiyle en yüksek doğruluk oranlarına ulaşılmıştır. BRCA ve COAD veri kümeleri için önerilen yöntemlerin sonuçları birbirine yakındır. En yüksek sonuçlar sırasıyla %99,42 ve %99,30'dur.

Tablo 5.2: Önerilen hibrit yöntemlerin ve sarmal yöntemlerin doğruluk oranı (%) sonuçları (CV).

Yöntemler	Kolon	Prostat	DLBCL	SMK	M1	M2	CNS	BRCA	COAD
GA-KNN	83,87	80,00	86,25	61,50	70,00	81,54	65,00	90,91	96,51
GA-DVM	91,67	89,00	98,75	57,73	58,00	83,08	73,33	99,09	99,07
PSO-KNN	83,33	85,00	91,25	63,85	57,00	80,77	68,33	97,52	99,07
PSO-DVM	90,00	91,00	95,00	60,00	58,00	84,62	70,00	98,93	99,23
GA(Pearson)-KNN	88,33	94,00	96,25	68,45	77,00	90,00	76,67	97,44	98,37
GA(Relief-F)-KNN	86,67	86,00	100,0	68,98	80,00	85,38	81,67	98,18	97,91
GA(MI)-KNN	86,67	86,00	97,50	70,59	82,00	83,85	73,33	98,68	98,14
GA(Pearson)-DVM	95,00	97,00	**100,0**	79,68	87,00	89,23	93,33	99,01	**99,30**
GA(Relief-F)-DVM	**98,33**	**99,00**	**100,0**	78,61	88,00	**90,77**	**95,00**	99,34	98,84
GA(MI)-DVM	**98,33**	97,00	**100,0**	**80,75**	**89,23**	90,00	90,00	99,01	97,91
PSO(Pearson)-KNN	85,00	87,00	90,00	63,85	75,00	85,38	73,33	97,77	99,07
PSO(Relief-F)-KNN	88,33	84,00	92,50	70,77	77,00	83,85	65,00	97,44	98,60
PSO(MI)-KNN	88,33	85,00	93,75	66,92	78,00	84,62	70,00	97,60	98,37
PSO(Pearson)-DVM	90,00	94,00	96,25	67,69	86,00	82,31	71,67	**99,42**	**99,30**
PSO(Relief-F)-DVM	88,33	95,00	93,75	68,46	86,00	85,38	71,67	99,17	**99,30**
PSO(MI)-DVM	93,33	95,00	97,50	70,00	87,00	86,15	75,00	99,09	**99,30**

DVM içeren önerilen yöntemlerle diğer yöntemlere göre genellikle daha yüksek sonuçlar elde edildiği görülmektedir. GA tabanlı ve DVM içeren yöntemlerle 9 veri kümesinin 7'si için en yüksek doğruluk oranı sonuçlarına ulaşılmıştır. COAD veri kümesi için PSO tabanlı yöntemlerin ve GA(Pearson)-DVM yönteminin sonucu %99,30'dur. Ancak GA(Pearson)-DVM yöntemi gen sayısını 219'a indirgemiş, PSO tabanlı yöntemler ise gen sayısını 29784'e indirgeyebilmiştir.

KNN içeren GA tabanlı önerilen yöntemlerle, GA-KNN sarmal yöntemine göre tüm veri kümeleri için, GA-DVM sarmal yöntemine göre 6 veri kümesi için daha yüksek sonuçlar elde edilmiştir.

Tablo 5.3: Yöntemlerin seçtiği sınıflandırmada kullanılan gen sayıları.

Yöntemler	Kolon	Prostat	DLBCL	SMK	M1	M2	CNS	BRCA	COAD
GA-KNN	171	4064	1853	7011	21338	8977	1945	1892	28901
GA-DVM	208	3931	1814	8850	21103	9080	1807	28773	28877
PSO-KNN	898	6203	3450	9947	23589	12081	3504	30238	30181
PSO-DVM	984	6191	3502	9926	23753	12161	3537	30307	29886
GA(Pearson)-KNN	3	8	**2**	73	22	304	3	8	19
GA(Relief-F)-KNN	**1**	**1**	**82**	114	16	**24**	**196**	30	2
GA(MI)-KNN	**1**	1	**2**	**221**	12	274	1	433	1
GA(Pearson)-DVM	20	275	101	842	32	216	28	529	**219**
GA(Relief-F)-DVM	33	210	214	1375	256	227	100	**625**	124
GA(MI)-DVM	49	437	38	419	**318**	396	119	629	1
PSO(Pearson)-KNN	967	6261	3537	10002	23314	12303	3502	30135	29996
PSO(Relief-F)-KNN	991	6203	3468	9981	23286	12101	3501	30180	30112
PSO(MI)-KNN	925	6177	3499	9987	23512	12192	3539	30200	30018
PSO(Pearson)-DVM	927	6135	3444	9981	23423	12284	3493	**30052**	**29936**
PSO(Relief-F)-DVM	994	6156	3394	9891	23346	12242	3565	29941	**29784**
PSO(MI)-DVM	952	6227	3468	10020	23816	12214	3622	29860	**29958**

GA-KNN sarmal yöntemiyle M1 veri kümesi için GA-DVM sarmal yöntemine göre daha yüksek doğruluk oranı sonucu elde edilmesine karşın önerilen yöntemlerde bu veri kümesi için

en yüksek doğruluk oranı GA(MI)-DVM yöntemiyle elde edilmiştir. Bu da önerilen yöntemlerin verideki değişimlere karşı daha dayanıklı (robust) olduğunu göstermektedir.

Tablo 5.3'de yöntemlerin seçtiği ve sınıflandırmada kullanılan gen sayıları gösterilmektedir. En yüksek doğruluk oranı sonuçlarının elde edildiği sınıflandırma sonuçları için gen sayıları koyulaştırılmış olarak gösterilmiştir. Buna göre, gen sayısını indirgemede en başarılı yöntemlerin önerilen GA tabanlı yöntemler olduğu anlaşılmaktadır. Bu yöntemlerdeki çaprazlama işleminde AND ve OR kullanılması ve genlerin filtre yöntem skorları kullanılarak çaprazlamanın özelleştirilmesinin gen sayısını azaltmada etkili olduğu düşünülmektedir. GA tabanlı önerilen yöntemler, seçilen gen sayısını sarmal yöntemlere göre ortalama %95 oranında azaltmıştır. Bununla birlikte örneğin GA(Relief-F)-DVM yöntemiyle seçilen genler arasında veri kümesinin karşılık geldiği kanser türünde tıbbi olarak önemli olduğu bilinen genler, diğer kanser türleri ve hastalıklarla ilişkili genler de bulunmaktadır.

Tablo 5.4'de literatürdeki diğer bazı çalışmaların ve önerilen yöntemlerin doğruluk oranı (seçilen gen sayıları) sonuçları gösterilmektedir. Buna göre, önerilen yöntemlerle daha yüksek doğruluk oranlarına ulaşıldığı görülmektedir. Kolon kanseri, prostat kanseri, M2 ve CNS veri kümeleri için GA(Relief-F)-DVM yöntemiyle sırasıyla %98,33, %99,00, %90,77, %95,00 olan en yüksek doğruluk oranı sonuçlarına ulaşılmıştır. DLBCL veri kümesi için birçok çalışmada %100 doğruluk oranı elde edildiği görülmektedir. Ancak GA(MI)-DVM yöntemiyle, %100 doğruluk oranına yöntemin seçtiği 38 gen ile ulaşılmıştır. Diğer çalışmalardaki seçilen gen sayısı belirtilmemiştir.

Sonuç olarak önerilen yöntemlerle diğer yöntemlere göre tüm değerlendirme ölçütleri açısından iyileşme sağlandığı gösterilmiştir. En yüksek doğruluk oranları, F1 değerleri ve en az gen sayısı önerilen yöntemlerle elde edilmiştir. En başarılı yöntem GA(Relief-F)-DVM yöntemi olarak belirlenmiştir. GA'nın filtre yöntemler kullanılarak geliştirilmesi ve çaprazlama işleminin özelleştirilmesi bu başarıyı sağlayan en önemli etkenlerdir. Ayrıca, bu çalışmada literatürdeki başlıca öznitelik seçimi yöntemlerinin gen ifade seviyelerini içeren veri kümelerini sınıflandırma başarıları karşılaştırılmıştır. Kullanılan yöntemler ve veri kümeleri açısından kapsamlı bir karşılaştırma sunulmuştur. Daha sonraki çalışmalarda ACO gibi farklı optimizasyon algoritmalarının ve farklı sınıflandırma yöntemlerinin değerlendirilmesi düşünülmektedir.

Tablo 5.4: Önerilen yöntemlerin ve diğer literatür çalışmalarının (%) doğruluk oranı (seçilen gen sayısı) sonuçları.

Yöntemler	Kolon Kanseri	Prostat Kanseri	DLBCL	M2	CNS
BM (3-fold CV) (Chen, 2006)	92,40	-	-	-	-
GA-ELM (10-fold CV) (Arunkumar ve diğ., 2017)	-	-	-	84,00 (9)	-
F-test + GA-DVM (5-fold CV) (Gunavathi ve Premalatha, 2014)	85,00	92,68	84,00	-	81,25
SNR + GA-DVM (5-fold CV) (Gunavathi ve Premalatha, 2014)	95,00	65,85	100	-	81,25
MIM + AGA-ELM (Lu ve diğ., 2017)	89,09	96,54	-	-	-
ACO-S-DVM (10-fold CV) (Li ve diğ., 2013)	81,42 (69)	-	-	-	-
Maks-Min ACO-DVM (LOOCV) (Yu ve diğ. 2009)	95,00 (10.8)	-	100	-	-
PSO-DVM (LOOCV) (Shen ve diğ., 2007)	94,00	-	-	-	-
IG + BBO-RF (10-fold CV) (Nikumbh ve diğ., 2012)	92,34 (11)	-	-	-	-
IG-DVM (10-fold CV) (Gao ve diğ., 2017)	90,32	96,08	100	-	-
(Peng ve diğ., 2006) (LOOCV)	87,00 (4) 91,95 (5)	-	-	-	88,66 (6)
FLD-NRS (Sun ve diğ., 2018)	88,0 (6)	80,00 (4)	-	-	-
MOCEPO (10-fold CV) (Baliarsingh ve diğ., 2019)	96,74	-	-	-	-
IG+micro GA -DVM (10-fold CV) (Pragadeesh ve diğ., 2019)	-	-	-	-	92,86 (29)
IG-MBKH (10-fold CV) (Zhang ve diğ., 2020)	96,47	-	-	-	90,34
MRMR+CHS+KNN (10-fold CV) (Wang ve diğ., 2021)	80,64 (72)	-	96,10 (61)	-	-
MOCS-EO (10-fold CV) (Othman ve diğ., 2020)	-	-	-	-	76,30 (952)
GA(Pearson)-KNN	88,33 (3)	94,00 (8)	96,25 (2)	90,00 (304)	76,67 (3)
GA(Relief-F)-KNN	86,67 (1)	86,00 (1)	100,0 (82)	85,38 (24)	81,67 (196)
GA(MI)-KNN	86,67 (1)	86,00 (1)	97,50 (2)	83,85 (274)	73,33 (1)
GA(Pearson)-DVM	95,00 (20)	97,00 (275)	**100,0 (101)**	89,23 (216)	93,33 (28)
GA(Relief-F)-DVM	**98,33 (33)**	**99,00 (210)**	**100,0 (214)**	**90,77 (227)**	**95,00 (100)**
GA(MI)-DVM	**98,33 (49)**	97,00(437)	**100,0 (38)**	90,00 (396)	90,00 (119)
PSO(Pearson)-KNN	85,00 (967)	87,00 (6261)	90,00 (3537)	85,38 (23421)	73,33 (3502)
PSO(Relief-F)-KNN	88,33 (991)	84,00 (6203)	92,50 (3468)	83,85 (23304)	65,00 (3501)
PSO(MI)-KNN	88,33 (925)	85,00 (6177)	93,75 (3499)	84,62 (23643)	70,00 (3539)
PSO(Pearson)-DVM	90,00 (927)	94,00 (6135)	96,25 (3444)	82,31 (23530)	71,67 (3493)
PSO(Relief-F)-DVM	88,33 (994)	95,00 (6156)	93,75 (3394)	85,38 (23469)	71,67 (3565)
PSO(MI)-DVM	93,33 (952)	95,00 (6227)	97,50 (3468)	86,15 (23907)	75,00 (3622)

KAYNAKLAR

Alba, E., Nieto, GJ., Jourdan, L., Talbi, EG., 2008, Gene Selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, *IEEE Congress on Evolutionary Computation (CEC 2007)*, 25-28 Sept. 2007, Singapore, 284-290.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A., 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci.*, 96 (12), 6745-6750.

Al-Batah, M., Zaqaibeh, B., Alomari, S.A., Alzboon, M.S., 2019, Gene microarray cancer classification using correlation based feature selection algorithm and rules classifiers, *International Journal of Online & Biomedical Engineering*, 15(8), 62-73.

Alshamlan, H.M., Badr, G.H., Alohali, Y.A., 2015, Genetic bee colony (GBC) algorithm: A new gene selection method for microarray cancer classification, Computational Biology and Chemistry, 56, 49-60.

Anaissi, A., Kennedy, P.J., 2011, Feature selection of imbalanced gene expression microarray data, *12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 6-8 July 2011 Sydney, 73-78.

Arunkumar, C., Sooraj, MP., Ramakrishnan, MPS., 2017, Finding expressed genes using Genetic Algorithm and Extreme Learning Machines, *4th International Conference on Advanced Computing and Communication Systems (ICACCS -2017)*, 6-7 Jan. 2017, Coimbatore, India, 1-4.

Babu, M., Sarkar, K., 2017, A comparative study of gene selection methods for cancer classification using microarray data, *International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 23-25 September 2016 Kolkata, 204–211.

Baliarsingh, S.K., Vipsita, S., Muhammad, K., Bakshi, S., 2019, Analysis of high-dimensional biomedical data using an evolutionary multi-objective emperor penguin optimizer, *Swarm and Evolutionary Computation*, 48, 262-273.

Barton, N.H., Briggs, D.E., Eisen, J.A., Goldstein, D.B., Patel, N.H., 2007, *Evolution*, Cold Spring Harbor Laboratory Press, New York, USA.

Burges, C.J.C., 1998, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2 (2), 121-167.

Chan, W.H., Mohamad, M.S., Deris, S., Zaki, N., Kasim, S., Omatu, S., Corchado, J.M., Ashwa, H.A., 2016, Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme, *Computers in Biology and Medicine*, 77, 102-115.

Chandrashekar, G., Sahin, F., 2014, A survey on feature selection methods, *Computers and Electrical Engineering*, 40 (1), 16-28.

Chen, XW., 2006, Margin-based wrapper methods for gene identification using microarray, *Neurocomputing*, 69 (16-18), 2236-2243.

Chuang, L.Y, Yang, C.H., Wu, K.C., Yang, C.H., 2011, A hybrid features election method for DNA microarray data, *Computers in Biology and Medicine*, 41 (4), 228-237.

Cortes, C. and Vapnik, V., 1995, Support-vector networks, *Machine Learning*, 20 (3), 273-297.

Devita, V.T., Lawrance, T.S., Rosenberg, S.A., 2011, *Cancer: principles & practice of oncology: primer of molecular biology of cancer*, Lippincott Williams & Wilkins, Philadelphia, ISBN: 978-1-4511-1897-1.

Ding, C. and Peng, H., Minimum redundancy feature selection from microarray gene expression data, 2005, *Journal of Bioinformatics and Computational Biology*, 3 (2), 185-205.

Ding, C.H.Q., 2002, Analysis of gene expression profiles: Class discovery and leaf ordering, *Proceedings of the Sixth Annual International Conference on Computational Biology* (*Proc. RECOMB)*, 18-21 April 2002, Washington, 127-136.

Fahrudin, T.M., Syarif, I., Barakbah, A.R., 2016, Ant colony algorithm for feature selection on microarray datasets, *International Electronics Symposium (IES)*, 29-30 Sept. 2016, Denpasar, Indonesia, 351-356.

Gao, L., Ye, M., Lu, X., Huang, D., 2017, Hybrid method based on information gain and support vector machine for gene selection in cancer Classification, *Genomics, Proteomics & Bioinformatics*, 15 (6), 389-395.

Goldberg, D.E., 1989, Genetic algorithms in search, *Optimization and Machine Learning*, Addison-Wesley Publishing Company Inc., London.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, 286 (5439), 531-537.

Gunavathi, C., Premalatha, K., 2014, Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification, *International Journal of Computer and Information Engineering*, 8 (8), 1490- 1497.

Guyon, I., Weston, J, Barnhill, S., 2002, Gene selection for cancer classification using support vector machines, *Machine Learning*, 46, 389-422.

Hejmadi, M., 2010, *Introduction to cancer biology*, Momna Hejmadi & Ventus Publishing APS, ISBN: 978-87-7681-476-6.

Hira, Z.M. and Gillies, D.F., 2015, A review of feature selection and feature extraction methods applied on microarray data, Advances *in Bioinformatics*, 2015, 198363, 13.

Holland, J., 1975, *Adaption in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.

Huerta, E.B., Duval, B., Hao, J.K., 2006, A hybrid GA/SVM approach for gene selection and classification of microarray data, *Applications of Evolutionary Computing*, 3907, 34-44.

Hwang, T., Sun, CH., Yun, T., Yi GS., 2010, FiGS: a filter-based gene selection workbench for microarray data, *BMC Bioinformatics*, 11 (50).

Jain, I., Jain, V.K., Jain, R., 2018, Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification, *Applied Soft Computing*, 62, 203-215.

Jolliffe, IT, Cadima, J., 2016, Principal component analysis: A review and recent developments, *Phil. Trans. R. Soc. A*, 374:20150202.

Juliusdottir, T., Keedwell, E, Corne, D., Narayanan, A., 2005, Two-phase EA/k-NN for feature selection and classification in cancer microarray datasets, *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '05)*, 15-15 Nov. 2005 La Jolla, 1-8.

Kar, S., Sharma, KD., Maitra, M., 2015, Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique, *Expert Systems with Applications*, 42 (1), 612-627.

Kennedy, J., and Eberhart, R., 1995, Particle swarm optimization, *Proceedings of the IEEE International Conference on Neural Networks*, 4-9 May 1998 Anchorage, 1942-1948.

Khadijah, Rismiyati, Mantau AJ., 2018, Multiclass classification of cancer based on microarray data using extreme learning machine, *1st International Conference on Informatics and Computational Sciences*, 15-16 Nov. Semarang, Indonesia, 159-164.

Kononenko, I., 1994, Estimating attributes: Analysis and extensions of RELIEF, *Machine Learning: ECML-94, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*, In: Bergadano F., De Raedt L. (ed.), 784, Springer, Berlin, Heidelberg, 171-182.

Lee, C.P and Leu, Y., 2011, A novel hybrid feature selection method for microarray data analysis, *Applied Soft Computing*, 11 (1), 208-213.

Lee, CP., Lin, WS., Chen, YM., Kuo, BJ., 2011, Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method, *Expert Systems with Applications*, 38 (5), 4661-4667.

Lee, M.Y., 2016, *Microarray bioprinting technology: Fundamentals and practices*, Springer, Switzerland.

Lee, Z.J., 2008, An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer, *Artificial Intelligence in Medicine*, 42 (1), 81-93.

Leenavinmalar, F., Kumarkombaiya, A., 2016, A novel cancer gene search model and classification using GPSO-BPNN, *International Journal on Engineering Technology and Sciences-IJETS*, 3 (6), 56-61.

Leung, Y. and Hung, Y., 2010, A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7 (1), 108-117.

Li, Y., Wang, G., Chen, H., Shi, L, Qin, L., 2013, An ant colony optimization based dimension reduction method for high-dimensional datasets, *Journal of Bionic Engineering*, 10, 231-241.

Liu, H., Liu, L., Zhang, H., 2010, Ensemble gene selection for cancer classification, *Pattern Recognition*, 43, 2763-2772.

Liu, X., Krishnan, A., and Mondry, A., 2005, An entropy-based gene selection method for cancer classification using microarray data, *BMC Bioinformatics*, 6 (76).

Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., Gao, Z., 2017, A hybrid feature selection algorithm for gene expression data classification, *Neurocomputing*, 256, 56-62.

Luo, K., Wang, G., Li, Q., Tao, J., 2019, An improved SVM-RFE based on F-statistic and mPDC for gene selection in cancer classification, *IEEE Access*, 7, 147617-147628.

Maldonado, S., Weber, R., Basak, J., 2011, Simultaneous feature selection and classification using kernel-penalized support vector machines, *Information Sciences*, 181, 115-128.

Meenachi, L., Ramakrishnan, S., 2021, Metaheuristic search based feature selection methods for classification of cancer, *Pattern Recognition*, 119, 108079.

Mundra, A.P. and Rajapakse, J.C., 2010, SVM-RFE with MRMR filter for gene selection, *IEEE Transactıons on Nanobioscience*, 9 (1), 31-37.

Naderi, A., Teschendorff, A.E., Barbosa-Morais, N.L., Pinder, SE, Green, A.R., Powe, D.G., Robertson, JF., Aparicio, S., Ellis, I.O., Brenton. J.D., Caldas, C., A gene-expression signature to predict survival in breast cancer across independent data sets, 2006.

Nikumbh, S., Ghosh, S. and Jayaraman, V.K., 2012, Biogeography-based informative gene selection and cancer classification using SVM and random forests, *IEEE Congress on Evolutionary Computation (CEC)*, 10-15 June 2012 Brisbane, QLD, Australia, 1-6.

Othman, M.S., Kumaran, S.R., Yusuf, L.M., 2020, Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data, *IEEE Access*, 8, 186348-186361.

Passi, K., Nour, A., Jain, C.K., 2017, Markov blanket: Efficient strategy for feature subset selection method for high dimensional microarray cancer datasets, *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 13-16 Nov. 2017 Kansas City, 1864-1871.

Peng, Y., Li, W., Liu, Y., 2006, A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification, *Cancer Informatics*, 2, 301-311.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., and Liotta, L.A., 2002, Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet*, 359, 572-577.

Pomeroy, S.L., Tamayo, P., Gaasenbeek ,M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerovak, L.C., Blackk, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmo, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R., 2002, Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, 415 (6870), 436-442.

Pragadeesh, C., Jeyaraj, R., Siranjeevi, K., Abishek, R., Jeyakumar, J., 2019, Hybrid feature selection using micro genetic algorithm on microarray gene expression data, *Journal of Intelligent & Fuzzy Systems*, 36 (3): 2241-2246.

Qaraad, M., Amjad, S., Manhrawy, I.I.M., Fathi, H., Hassan, B.A. et al., 2021, A hybrid feature selection optimization model for high dimension data classification, *IEEE Access*, 9: 42884-42895.

Rabiee, M., Rostami, A., Rabiee, N., Bagherzadeh, M., 2021, *Microarray technologies*, Biomedical applications of microfluidic devices, In: Hamblin M.R., Karimi M. (eds.), Chapter 4, Academic Press, USA, 77-98.

Raghavachari, N., Garcia-Reyero, N., 2018, *Gene expression analysis: methods and protocols*, Human Press, Springer, New York, USA.

Rismiyati, K., Mantau, A.J., 2018, Multiclass classification of cancer based on microarray data using extreme learning machine, *1st International Conference on Informatics and Computational Sciences (ICICoS)*, 15-16 Nov. 2017 Semarang, Indonesia, IEEE, 159-164.

Rueda, L., Ali, A., 2014, *Introduction to microarrays*, Microarray image and data analysis: Theory and practice, In: Rueda L. (ed.), Chapter 1, CRC press, New York, USA.

Ryu, J. and Cho, SB., 2002, Towards optimal feature and classifier for gene expression classification of cancer, *Advances in Soft Computing AFSS*, Lecture Notes in Computer Science, 2275, 310–317.

Sahu, B., Mishra, D., 2012, A novel feature selection algorithm using particle swarm optimization for cancer microarray data, *Procedia Engineering*, 38, 27-31.

Salema, H., Attiyab, G., El-Fishawy, N., 2017, Classification of human cancer diseases by gene expression profiles, *Applied Soft Computing*, 50, 124-134.

Shah, M., Marchand, M., Corbeil, J., 2011, Feature selection with conjunctions of decision stumps and learning from microarray data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (1), 174-186.

Shannon, C.E., Weaver, W., 1949, *The mathematical theory of communication*, University of Illinois Press.

Sharma, A., Paliwal, K.K., 2015, Linear discriminant analysis for the small sample size problem: an overview, *International Journal of Machine Learning and Cybernetics*, 2015 (6), 443-454.

Shen, Q., Shi, WM., Kong, W., Ye BX., 2007, Combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification, Talanta 2007; 71 (4): 1679-1683.

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

Shreem, S.S., Abdullah, S., Nazri, M.Z.A., Alzaqebah, M., 2012, Hybridizing reliefF, MRMR filters and GA wrapper approaches for gene selection, *Journal of Theoretical and Applied Information Technology*, 46 (2), 1034-1039.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R., 2002, Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, 1 (2), 203-209.

Somasundaram, K., Mungamuri, S.K. and Wajapeyee, N., 2002, DNA microarray technology and its applications in cancer biology, *Applied genomics and proteomics*, 1(4), 209-218.

Spira, A., Beane, J., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y., Calner, P., Sebastiani, P., 2007, Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, *Nat. Med.*, 13 (3), 361-366.

Sun, L., Zhang, X., Xu, J., Wang, W., Liu, R., 2018, A Gene selection approach based on the fisher linear discriminant and the neighborhood rough set, *Bioengineered*, 9 (1), 144-151.

Tabakhi, S., Najafi, A., Ranjbar, R., Moradi, P., 2015, Gene selection for microarray data classification using a novel ant colony optimization, *Neurocomputing*, 168, 1024-1036.

Tong, D.L. and Schierz, A.C., 2011, Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data, *Artificial Intelligence in Medicine*, 53 (1), 47-56.

Uriarte, R.D. and Andres, S.A., 2006, Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, 7 (3).

U.S. National Cancer Institute, *Genomic Data Commons Data Portal (GDCD)*, https://portal.gdc.cancer.gov [Ziyaret tarihi: 2 Şubat 2020].

Uzma, Halim, Z., 2021, An ensemble filter-based heuristic approach for cancerous gene expression classification, *Knowledge-Based Systems*, 234, 107560.

Vanitha, C.D.A., Devaraj, D., Venkatesulu, M., 2015, Gene expression data classification using support vector machine and mutual information-based gene selection, *Procedia Computer Science*, 47, 13-21.

Veer, L.J., Dai, H., Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., Kooy, K.D., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., and Friend, S.H., 2002, Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415 (6871), 530-536.

Wang, A., Liu, H., Chen, G., 2021, Chaotic harmony search based multi-objective feature selection for classification of gene expression profiles. *In: IEEE 9th International Conference on Bioinformatics and Computational Biology (ICBCB)*, Taiyuan, China, 107-112.

Wang, Y., and Makedon, F., 2004, Application of relief-f feature filtering algorithm to selecting informative genes for cancer classification using microarray data, *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, 19- Aug. 2004 Stanford, 497-498.

Wanga, G., Song, Q., Xu, B., Zhou, Y., 2013, Selecting feature subset for high dimensional data via the propositional FOIL rules, Pattern *Recognition*, 46, 199-214.

Yıldız, O., Tez, M., Bilge, H.Ş., Akcayol, M.A, Güler, İ., 2012, Meme kanseri sınıflandırması için veri füzyonu ve genetik algoritma tabanlı gen seçimi, *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 27 (3), 659-668.

Yu, H., Gu, G., Liu, H., Shen, J., and Zhao, J., 2009, A modified ant colony optimization algorithm for tumor marker gene selection, *Genomics, Proteomics & Bioinformatics*, 7 (4), 200-208.

Zhang, G., Hou, J., Wang, J., Yan, C., 2020, Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdisciplinary Sciences: Computational Life Sciences*, 12, 288-301.

Zhao, S., Fung-Leung, W-P., Bittner, A., Ngo, K., Liu, X., 2014, Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells, *Plos One*, 9(1), e78644.

EKLER

EK.1. FİLTRE YÖNTEMLERLE ÖZNİTELİK SEÇİMİ SONRASI SINIFLANDIRMA SONUÇLARI

Tablo Ek.1.1: Kolon kanseri veri kümesi için DDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	100,00	100,00	100,00	100,00	1,0000	100,00	2
Fisher skor	100,00	100,00	100,00	100,00	1,0000	100,00	2
SNR	94,44	100,00	83,33	100,00	0,9091	92,31	12
T-Test	100,00	100,00	100,00	100,00	1,0000	100,00	59
F-Test	100,00	100,00	100,00	100,00	1,0000	100,00	22
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	1
MRMR	94,44	100,00	83,33	100,00	0,9091	92,31	34
Laplace skor	100,00	100,00	100,00	100,00	1,0000	100,00	13
CFS	94,44	100,00	83,33	100,00	0,9091	92,31	219
Karşılıklı Bilgi	94,44	100,00	83,33	100,00	0,9091	92,31	616
Ki-kare	100,00	100,00	100,00	100,00	1,0000	100,00	1

Tablo Ek.1.2: Kolon kanseri veri kümesi için PDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	100,00	100,00	100,00	100,00	1,0000	100,00	8
Fisher skor	100,00	100,00	100,00	100,00	1,0000	100,00	8
SNR	77,78	100,00	33,33	100,00	0,5000	75,00	26
T-Test	94,44	100,00	83,33	100,00	0,9091	92,31	51
F-Test	88,89	75,00	100,00	83,33	0,8571	100,00	16
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	3
MRMR	88,89	83,33	83,33	91,67	0,8333	91,67	90
Laplace skor	94,44	85,71	100,00	91,67	0,9231	100,00	39
CFS	94,44	100,00	83,33	100,00	0,9091	92,31	11
Karşılıklı Bilgi	83,33	71,43	83,33	83,33	0,7692	90,91	5
Ki-kare	94,44	85,71	100,00	91,67	0,9231	100,00	7

Tablo Ek.1.3: Kolon kanseri veri kümesi için KNN yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	100,00	100,00	100,00	100,00	1,0000	100,00	2
Fisher skor	100,00	100,00	100,00	100,00	1,0000	100,00	2
SNR	94,44	100,00	83,33	100,00	0,9091	92,31	880
T-Test	100,00	100,00	100,00	100,00	1,0000	100,00	118
F-Test	100,00	100,00	100,00	100,00	1,0000	100,00	720
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	1
MRMR	94,44	100,00	83,33	100,00	0,9091	92,31	46
Laplace skor	100,00	100,00	100,00	100,00	1,0000	100,00	35
CFS	100,00	100,00	100,00	100,00	1,0000	100,00	134
Karşılıklı Bilgi	100,00	100,00	100,00	100,00	1,0000	100,00	182
Ki-kare	100,00	100,00	100,00	100,00	1,0000	100,00	1

Tablo Ek.1.4: Kolon kanseri veri kümesi için NB yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	100,00	100,00	100,00	100,00	1,0000	100,00	2
Fisher skor	100,00	100,00	100,00	100,00	1,0000	100,00	2
SNR	83,33	80,00	66,67	91,67	0,7273	84,62	37
T-Test	83,33	71,43	83,33	83,33	0,7692	90,91	26
F-Test	72,22	57,14	66,67	75,00	0,6154	81,82	18
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	1
MRMR	94,44	85,71	100,00	91,67	0,9231	100,00	37
Laplace skor	72,22	60,00	50,00	83,33	0,5455	76,92	28
CFS	88,89	83,33	83,33	91,67	0,8333	91,67	2
Karşılıklı Bilgi	72,22	55,56	83,33	66,67	0,6667	88,89	4
Ki-kare	100,00	100,00	100,00	100,00	1,0000	100,00	1

Tablo Ek.1.5: Kolon kanseri veri kümesi için KA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	94,44	85,71	100,00	91,67	0,9231	100,00	5
Fisher skor	94,44	85,71	100,00	91,67	0,9231	100,00	5
SNR	94,44	85,71	100,00	91,67	0,9231	100,00	637
T-Test	88,89	83,33	83,33	91,67	0,8333	91,67	15
F-Test	83,33	80,00	66,67	91,67	0,7273	84,62	26
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	31
MRMR	94,44	100,00	83,33	100,00	0,9091	92,31	648
Laplace skor	94,44	85,71	100,00	91,67	0,9231	100,00	30
CFS	94,44	85,71	100,00	91,67	0,9231	100,00	179
Karşılıklı Bilgi	83,33	100,00	50,00	100,00	0,6667	80,00	2
Ki-kare	100,00	100,00	100,00	100,00	1,0000	100,00	13

Tablo Ek.1.6: Kolon kanseri veri kümesi için İBYSA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	100,00	100,00	100,00	100,00	1,0000	100,00	8
Fisher skor	94,44	100,00	83,33	100,00	0,9091	92,31	8
SNR	100,00	100,00	100,00	100,00	1,0000	100,00	33
T-Test	94,44	100,00	83,33	100,00	0,9091	92,31	60
F-Test	100,00	100,00	100,00	100,00	1,0000	100,00	175
Relief-F	94,44	85,71	100,00	91,67	0,9231	100,00	3
MRMR	100,00	100,00	100,00	100,00	1,0000	100,00	37
Laplace skor	100,00	100,00	100,00	100,00	1,0000	100,00	13
CFS	100,00	100,00	100,00	100,00	1,0000	100,00	10
Karşılıklı Bilgi	83,33	100,00	50,00	100,00	0,6667	80,00	2
Ki-kare	100,00	100,00	100,00	100,00	1,0000	100,00	25

Tablo Ek.1.7: Prostat kanseri veri kümesi için DDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	93,33	100,00	86,67	100,00	0,9286	88,24	112
Fisher skor	93,33	100,00	86,67	100,00	0,9286	88,24	112
SNR	90,00	100,00	80,00	100,00	0,8889	83,33	119
T-Test	93,33	100,00	86,67	100,00	0,9286	88,24	94
F-Test	96,67	100,00	93,33	100,00	0,9655	93,75	74
Relief-F	96,67	100,00	93,33	100,00	0,9655	93,75	14
MRMR	90,00	92,86	86,67	93,33	0,8966	87,50	107
Laplace skor	90,00	92,86	86,67	93,33	0,8966	87,50	70
CFS	86,67	92,31	80,00	93,33	0,8571	82,35	644
Karşılıklı Bilgi	93,33	100,00	86,67	100,00	0,9286	88,24	2
Ki-kare	96,67	100,00	93,33	100,00	0,9655	93,75	7

Tablo Ek.1.8: Prostat kanseri veri kümesi için PDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	93,33	100,00	86,67	100,00	0,9286	88,24	14
Fisher skor	93,33	100,00	86,67	100,00	0,9286	88,24	14
SNR	86,67	100,00	73,33	100,00	0,8462	78,95	55
T-Test	93,33	100,00	86,67	100,00	0,9286	88,24	14
F-Test	90,00	87,50	93,33	86,67	0,9032	92,86	10
Relief-F	96,67	100,00	93,33	100,00	0,9655	93,75	14
MRMR	73,33	81,82	60,00	86,67	0,6923	68,42	4
Laplace skor	90,00	100,00	80,00	100,00	0,8889	83,33	728
CFS	83,33	81,25	86,67	80,00	0,8387	85,71	47
Karşılıklı Bilgi	96,67	100,00	93,33	100,00	0,9655	93,75	2
Ki-kare	96,67	100,00	93,33	100,00	0,9655	93,75	29

Tablo Ek.1.9: Prostat kanseri veri kümesi için KNN yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	93,33	100,00	86,67	100,00	0,9286	88,24	6
Fisher skor	93,33	100,00	86,67	100,00	0,9286	88,24	6
SNR	73,33	81,82	60,00	86,67	0,6923	68,42	243
T-Test	93,33	100,00	86,67	100,00	0,9286	88,24	6
F-Test	83,33	81,25	86,67	80,00	0,8387	85,71	10
Relief-F	90,00	92,86	86,67	93,33	0,8966	87,50	2
MRMR	90,00	87,50	93,33	86,67	0,9032	92,86	240
Laplace skor	66,67	64,71	73,33	60,00	0,6875	69,23	224
CFS	80,00	76,47	86,67	73,33	0,8125	84,62	46
Karşılıklı Bilgi	96,67	100,00	93,33	100,00	0,9655	93,75	2
Ki-kare	96,67	100,00	93,33	100,00	0,9655	93,75	8

Tablo Ek.1.10: Prostat kanseri veri kümesi için NB yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	93,33	100,00	86,67	100,00	0,9286	88,24	35
Fisher skor	93,33	100,00	86,67	100,00	0,9286	88,24	35
SNR	66,67	66,67	66,67	66,67	0,6667	66,67	2
T-Test	93,33	100,00	86,67	100,00	0,9286	88,24	45
F-Test	80,00	73,68	93,33	66,67	0,8235	90,91	10
Relief-F	93,33	100,00	86,67	100,00	0,9286	88,24	2
MRMR	90,00	92,86	86,67	93,33	0,8966	87,50	172
Laplace skor	56,67	53,85	93,33	20,00	0,6829	75,00	1
CFS	80,00	73,68	93,33	66,67	0,8235	90,91	3
Karşılıklı Bilgi	93,33	100,00	86,67	100,00	0,9286	88,24	2
Ki-kare	93,33	100,00	86,67	100,00	0,9286	88,24	5

Tablo Ek.1.11: Prostat kanseri veri kümesi için KA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	83,33	85,71	80,00	86,67	0,8276	81,25	2
Fisher skor	83,33	85,71	80,00	86,67	0,8276	81,25	2
SNR	80,00	76,47	86,67	73,33	0,8125	84,62	360
T-Test	83,33	85,71	80,00	86,67	0,8276	81,25	2
F-Test	93,33	100,00	86,67	100,00	0,9286	88,24	177
Relief-F	83,33	85,71	80,00	86,67	0,8276	81,25	691
MRMR	86,67	92,31	80,00	93,33	0,8571	82,35	2
Laplace skor	83,33	85,71	80,00	86,67	0,8276	81,25	2
CFS	73,33	73,33	73,33	73,33	0,7333	73,33	54
Karşılıklı Bilgi	90,00	100,00	80,00	100,00	0,8889	83,33	3
Ki-kare	83,33	85,71	80,00	86,67	0,8276	81,25	2

Tablo Ek.1.12: Prostat kanseri veri kümesi için İBYSA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	93,33	100,00	86,67	100,00	0,9286	88,24	19
Fisher skor	93,33	100,00	86,67	100,00	0,9286	88,24	5
SNR	93,33	100,00	86,67	100,00	0,9286	88,24	92
T-Test	93,33	100,00	86,67	100,00	0,9286	88,24	17
F-Test	96,67	100,00	93,33	100,00	0,9655	93,75	36
Relief-F	96,67	100,00	93,33	100,00	0,9655	93,75	9
MRMR	93,33	93,33	93,33	93,33	0,9333	93,33	390
Laplace skor	90,00	92,86	86,67	93,33	0,8966	87,50	203
CFS	93,33	100,00	86,67	100,00	0,9286	88,24	702
Karşılıklı Bilgi	93,33	100,00	86,67	100,00	0,9286	88,24	2
Ki-kare	96,67	100,00	93,33	100,00	0,9655	93,75	14

Tablo Ek.1.13: DLBCL veri kümesi için DDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	95,65	94,44	100,00	83,33	0,9714	100,00	40
Fisher skor	95,65	100,00	94,12	100,00	0,9697	85,71	30
SNR	95,65	94,44	100,00	83,33	0,9714	100,00	5460
T-Test	95,65	94,44	100,00	83,33	0,9714	100,00	6890
F-Test	95,65	94,44	100,00	83,33	0,9714	100,00	470
Relief-F	95,65	94,44	100,00	83,33	0,9714	100,00	80
MRMR	95,65	94,44	100,00	83,33	0,9714	100,00	210
Laplace skor	100,00	100,00	100,00	100,00	1,0000	100,00	60
CFS	95,65	94,44	100,00	83,33	0,9714	100,00	6270
Karşılıklı Bilgi	95,65	94,44	100,00	83,33	0,9714	100,00	90
Ki-kare	95,65	100,00	94,12	100,00	0,9697	85,71	9

Tablo Ek.1.14: DLBCL veri kümesi için PDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	91,30	89,47	100,00	66,67	0,9444	100,00	140
Fisher skor	91,30	89,47	100,00	66,67	0,9444	100,00	150
SNR	86,96	85,00	100,00	50,00	0,9189	100,00	80
T-Test	91,30	89,47	100,00	66,67	0,9444	100,00	20
F-Test	86,96	85,00	100,00	50,00	0,9189	100,00	50
Relief-F	86,96	85,00	100,00	50,00	0,9189	100,00	20
MRMR	78,26	87,50	82,35	66,67	0,8485	57,14	20
Laplace skor	95,65	100,00	94,12	100,00	0,9697	85,71	30
CFS	86,96	88,89	94,12	66,67	0,9143	80,00	90
Karşılıklı Bilgi	91,30	89,47	100,00	66,67	0,9444	100,00	5
Ki-kare	95,65	94,44	100,00	83,33	0,9714	100,00	5

Tablo Ek.1.15: DLBCL veri kümesi için KNN yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	95,65	100,00	94,12	100,00	0,9697	85,71	160
Fisher skor	95,65	100,00	94,12	100,00	0,9697	85,71	160
SNR	95,65	100,00	94,12	100,00	0,9697	85,71	350
T-Test	95,65	100,00	94,12	100,00	0,9697	85,71	50
F-Test	95,65	100,00	94,12	100,00	0,9697	85,71	6890
Relief-F	95,65	94,44	100,00	83,33	0,9714	100,00	20
MRMR	100,00	100,00	100,00	100,00	1,0000	100,00	330
Laplace skor	95,65	100,00	94,12	100,00	0,9697	85,71	140
CFS	95,65	94,44	100,00	83,33	0,9714	100,00	930
Karşılıklı Bilgi	100,00	100,00	100,00	100,00	1,0000	100,00	98
Ki-kare	95,65	94,44	100,00	83,33	0,9714	100,00	35

Tablo Ek.1.16: DLBCL veri kümesi için NB yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	95,65	100,00	94,12	100,00	0,9697	85,71	50
Fisher skor	95,65	100,00	94,12	100,00	0,9697	85,71	50
SNR	95,65	94,44	100,00	83,33	0,9714	100,00	130
T-Test	91,30	94,12	94,12	83,33	0,9412	83,33	20
F-Test	86,96	85,00	100,00	50,00	0,9189	100,00	80
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	110
MRMR	78,26	80,00	94,12	33,33	0,8649	66,67	540
Laplace skor	95,65	100,00	94,12	100,00	0,9697	85,71	70
CFS	86,96	85,00	100,00	50,00	0,9189	100,00	100
Karşılıklı Bilgi	91,30	94,12	94,12	83,33	0,9412	83,33	2
Ki-kare	95,65	100,00	94,12	100,00	0,9697	85,71	16

Tablo Ek.1.17: DLBCL veri kümesi için KA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	95,65	100,00	94,12	100,00	0,9697	85,71	50
Fisher skor	95,65	100,00	94,12	100,00	0,9697	85,71	50
SNR	95,65	94,44	100,00	83,33	0,9714	100,00	130
T-Test	91,30	94,12	94,12	83,33	0,9412	83,33	20
F-Test	91,30	89,47	100,00	66,67	0,9444	100,00	6780
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	110
MRMR	91,30	89,47	100,00	66,67	0,9444	100,00	6850
Laplace skor	95,65	100,00	94,12	100,00	0,9697	85,71	70
CFS	91,30	89,47	100,00	66,67	0,9444	100,00	4220
Karşılıklı Bilgi	86,96	88,89	94,12	66,67	0,9143	80,00	30
Ki-kare	91,30	94,12	94,12	83,33	0,9412	83,33	4

Tablo Ek.1.18: DLBCL veri kümesi için İBYSA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	100,00	100,00	100,00	100,00	1,0000	100,00	200
Fisher skor	95,65	94,44	100,00	83,33	0,9714	100,00	40
SNR	95,65	94,44	100,00	83,33	0,9714	100,00	130
T-Test	95,65	94,44	100,00	83,33	0,9714	100,00	630
F-Test	95,65	94,44	100,00	83,33	0,9714	100,00	470
Relief-F	100,00	100,00	100,00	100,00	1,0000	100,00	240
MRMR	95,65	94,44	100,00	83,33	0,9714	100,00	160
Laplace skor	100,00	100,00	100,00	100,00	1,0000	100,00	90
CFS	95,65	94,44	100,00	83,33	0,9714	100,00	560
Karşılıklı Bilgi	100,00	100,00	100,00	100,00	1,0000	100,00	11
Ki-kare	95,65	100,00	94,12	100,00	0,9697	85,71	4

Tablo Ek.1.19: SMK veri kümesi için DDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	80,36	78,13	86,21	74,07	0,8197	83,33	331
Fisher skor	80,36	78,13	86,21	74,07	0,8197	83,33	331
T-Test	80,36	75,00	93,10	66,67	0,8308	90,00	22
F-Test	75,00	70,27	89,66	59,26	0,7879	84,21	21
Relief-F	83,93	77,78	96,55	70,37	0,8615	95,00	81
MRMR	82,14	85,19	79,31	85,19	0,8214	79,31	611
Laplace skor	73,21	70,59	82,76	62,96	0,7619	77,27	10
CFS	76,79	73,53	86,21	66,67	0,7937	81,82	197
Karşılıklı Bilgi	78,57	75,76	86,21	70,37	0,8065	82,61	146
Ki-kare	78,57	72,97	93,10	62,96	0,8182	89,47	313

Tablo Ek.1.20: SMK veri kümesi için PDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	87,50	86,67	89,66	85,19	0,8814	88,46	130
Fisher skor	87,50	86,67	89,66	85,19	0,8814	88,46	130
SNR	75,00	72,73	82,76	66,67	0,7742	78,26	24
T-Test	85,71	81,82	93,10	77,78	0,8710	91,30	35
F-Test	73,21	66,67	96,55	48,15	0,7887	92,86	39
Karşılıklı Bilgi	71,43	84,21	55,17	88,89	0,6667	64,86	339
Ki-kare	76,79	80,77	72,41	81,48	0,7636	73,33	184

Tablo Ek.1.21: SMK veri kümesi için KNN yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	78,57	77,42	82,76	74,07	0,8000	80,00	215
Fisher skor	78,57	77,42	82,76	74,07	0,8000	80,00	215
SNR	76,79	76,67	79,31	74,07	0,7797	76,92	72
T-Test	82,14	85,19	79,31	85,19	0,8214	79,31	349
F-Test	73,21	65,91	100,00	44,44	0,7945	100,00	404
Relief-F	85,71	78,38	100,00	70,37	0,8788	100,00	93
MRMR	73,21	73,33	75,86	70,37	0,7458	73,08	274
Laplace skor	76,79	71,05	93,10	59,26	0,8060	88,89	74
CFS	78,57	75,76	86,21	70,37	0,8065	82,61	83
Karşılıklı Bilgi	73,21	69,44	86,21	59,26	0,7692	80,00	3
Ki-kare	82,14	78,79	89,66	74,07	0,8387	86,96	35

Tablo Ek.1.22: SMK veri kümesi için NB yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	82,14	77,14	93,10	70,37	0,8438	90,48	24
Fisher skor	82,14	77,14	93,10	70,37	0,8438	90,48	24
SNR	71,43	69,70	79,31	62,96	0,7419	73,91	17
T-Test	82,14	77,14	93,10	70,37	0,8438	90,48	10
F-Test	75,00	71,43	86,21	62,96	0,7813	80,95	499
Relief-F	80,36	76,47	89,66	70,37	0,8254	86,36	34
MRMR	66,07	64,71	75,86	55,56	0,6984	68,18	413
Laplace skor	62,50	65,38	58,62	66,67	0,6182	60,00	2
CFS	69,64	73,08	65,52	74,07	0,6909	66,67	7
Karşılıklı Bilgi	67,86	67,74	72,41	62,96	0,7000	68,00	9
Ki-kare	78,57	77,42	82,76	74,07	0,8000	80,00	18

Tablo Ek.1.23: SMK veri kümesi için KA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	75,00	71,43	86,21	62,96	0,7813	80,95	3
Fisher skor	75,00	71,43	86,21	62,96	0,7813	80,95	3
SNR	71,43	74,07	68,97	74,07	0,7143	68,97	709
T-Test	73,21	70,59	82,76	62,96	0,7619	77,27	22
F-Test	71,43	69,70	79,31	62,96	0,7419	73,91	395
Relief-F	75,00	72,73	82,76	66,67	0,7742	78,26	627
MRMR	71,43	70,97	75,86	66,67	0,7333	72,00	4
Laplace skor	75,00	71,43	86,21	62,96	0,7813	80,95	322
CFS	64,29	60,98	86,21	40,74	0,7143	73,33	36
Karşılıklı Bilgi	73,21	71,88	79,31	66,67	0,7541	75,00	569
Ki-kare	69,64	75,00	62,07	77,78	0,6792	65,63	1

Tablo Ek.1.24: SMK veri kümesi için İBYSA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	87,50	82,35	96,55	77,78	0,8889	95,45	38
Fisher skor	85,71	80,00	96,55	74,07	0,8750	95,24	604
SNR	80,36	78,13	86,21	74,07	0,8197	83,33	552
T-Test	83,93	77,78	96,55	70,37	0,8615	95,00	27
F-Test	78,57	72,97	93,10	62,96	0,8182	89,47	742
Relief-F	85,71	80,00	96,55	74,07	0,8750	95,24	56
MRMR	82,14	85,19	79,31	85,19	0,8214	79,31	497
Laplace skor	76,79	76,67	79,31	74,07	0,7797	76,92	52
CFS	78,57	72,97	93,10	62,96	0,8182	89,47	197
Karşılıklı Bilgi	80,36	76,47	89,66	70,37	0,8254	86,36	151
Ki-kare	78,57	74,29	89,66	66,67	0,8125	85,71	135

Tablo Ek.1.25: Meme kanseri veri kümesi için DDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	89,47	95,65	88,00	92,31	0,9167	80,00	106
Fisher skor	89,47	95,65	88,00	92,31	0,9167	80,00	106
SNR	92,11	95,83	92,00	92,31	0,9388	85,71	278
T-Test	92,11	95,83	92,00	92,31	0,9388	85,71	505
F-Test	86,84	85,71	96,00	69,23	0,9057	90,00	93
Relief-F	92,11	95,83	92,00	92,31	0,9388	85,71	33
Laplace skor	86,84	91,67	88,00	84,62	0,8980	78,57	288
Karşılıklı Bilgi	89,47	92,00	92,00	84,62	0,9200	84,62	21
Ki-kare	92,11	95,83	92,00	92,31	0,9388	85,71	1052

Tablo Ek.1.26: Meme kanseri veri kümesi için PDVM yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	92,11	95,83	92,00	92,31	0,9388	85,71	47
Fisher skor	89,47	92,00	92,00	84,62	0,9200	84,62	20
SNR	73,68	74,19	92,00	38,46	0,8214	71,43	40
T-Test	86,84	91,67	88,00	84,62	0,8980	78,57	5
F-Test	81,58	80,00	96,00	53,85	0,8727	87,50	5
Relief-F	89,47	88,89	96,00	76,92	0,9231	90,91	40
Laplace skor	84,21	80,65	100,00	53,85	0,8929	100,00	332
Karşılıklı Bilgi	86,84	88,46	92,00	76,92	0,9020	83,33	5
Ki-kare	86,84	85,71	96,00	69,23	0,9057	90,00	28

Tablo Ek.1.27: Meme kanseri veri kümesi için KNN yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	84,21	85,19	92,00	69,23	0,8846	81,82	698
Fisher skor	84,21	85,19	92,00	69,23	0,8846	81,82	701
SNR	86,84	91,67	88,00	84,62	0,8980	78,57	112
T-Test	84,21	88,00	88,00	76,92	0,8800	76,92	4
F-Test	78,95	75,76	100,00	38,46	0,8621	100,00	58
Relief-F	84,21	85,19	92,00	69,23	0,8846	81,82	218
Laplace skor	84,21	85,19	92,00	69,23	0,8846	81,82	82
Karşılıklı Bilgi	84,21	88,00	88,00	76,92	0,8800	76,92	2
Ki-kare	84,21	85,19	92,00	69,23	0,8846	81,82	113

Tablo Ek.1.28: Meme kanseri veri kümesi için NB yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	86,84	91,67	88,00	84,62	0,8980	78,57	386
Fisher skor	86,84	91,67	88,00	84,62	0,8980	78,57	388
SNR	76,32	76,67	92,00	46,15	0,8364	75,00	59
T-Test	86,84	91,67	88,00	84,62	0,8980	78,57	3
F-Test	84,21	88,00	88,00	76,92	0,8800	76,92	1039
Relief-F	86,84	91,67	88,00	84,62	0,8980	78,57	361
Laplace skor	84,21	88,00	88,00	76,92	0,8800	76,92	7
Karşılıklı Bilgi	86,84	91,67	88,00	84,62	0,8980	78,57	1
Ki-kare	84,21	88,00	88,00	76,92	0,8800	76,92	3

Tablo Ek.1.29: Meme kanseri veri kümesi için KA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	86,84	91,67	88,00	84,62	0,8980	78,57	1
Fisher skor	86,84	91,67	88,00	84,62	0,8980	78,57	1
SNR	81,58	90,91	80,00	84,62	0,8511	68,75	85
T-Test	86,84	91,67	88,00	84,62	0,8980	78,57	1
F-Test	84,21	85,19	92,00	69,23	0,8846	81,82	111
Relief-F	86,84	91,67	88,00	84,62	0,8980	78,57	1
Laplace skor	81,58	87,50	84,00	76,92	0,8571	71,43	4
Karşılıklı Bilgi	84,21	88,00	88,00	76,92	0,8800	76,92	2
Ki-kare	86,84	91,67	88,00	84,62	0,8980	78,57	392

Tablo Ek.1.30: Meme kanseri veri kümesi için İBYSA yöntemi sınıflandırma sonuçları.

Filtre Yöntemler	Doğruluk Oranı %	Kesinlik %	Duyarlılık %	Özgüllük %	F1	NKD %	Gen sayısı
Pearson Kor.	92,11	95,83	92,00	92,31	0,9388	85,71	100
Fisher skor	92,11	95,83	92,00	92,31	0,9388	85,71	32
SNR	92,11	95,83	92,00	92,31	0,9388	85,71	275
T-Test	92,11	95,83	92,00	92,31	0,9388	85,71	51
F-Test	89,47	86,21	100,00	69,23	0,9259	100,00	170
Relief-F	92,11	95,83	92,00	92,31	0,9388	85,71	81
Laplace skor	89,47	92,00	92,00	84,62	0,9200	84,62	282
Karşılıklı Bilgi	89,47	92,00	92,00	84,62	0,9200	84,62	21
Ki-kare	92,11	95,83	92,00	92,31	0,9388	85,71	274

ÖZGEÇMİŞ