
GAZİ ÜNİVERSİTESİ
Gazi
1926

DNA MİKRODİZİ VERİLERİ KULLANILARAK KANSER SINIFLANDIRMASINDA TOPLULUK ÖĞRENİMİ

Nazrin ABILOVA

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TEMMUZ 2025

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Nazrin ABİLOVA

27/06/2025

DNA MİKRODİZİ VERİLERİ KULLANILARAK KANSER SINIFLANDIRMASINDA TOPLULUK ÖĞRENİMİ

(Yüksek Lisans Tezi)

Nazrin ABİLOVA

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2025

ÖZET

Kanser, dünya genelinde en yüksek ölüm oranlarına sahip hastalık gruplarından biridir. Son yıllarda, biyoteknolojik gelişmelerin de etkisiyle, gen ekspresyon verileri gibi moleküler biyobelirteçlere dayalı veriler, kanserin tanı ve teşhisinde devrim niteliğinde katkılar sağlamıştır. Özellikle DNA mikrodizi teknolojisi sayesinde, aynı anda on binlerce genin ekspresyon düzeyi ölçülebilmekte ve bu sayede kanserin moleküler düzeydeki dinamikleri daha ayrıntılı olarak analiz edilebilmektedir. Bununla birlikte, bu tür veriler genellikle "yüksek boyutlu ve düşük örnek sayılı" bir yapıya sahiptir. Bu yapı, geleneksel makine öğrenmesi algoritmaları için aşırı öğrenme riski taşımakta ve analiz sürecini oldukça karmaşık hale getirmektedir. Bu karmaşık veri yapısıyla başa çıkmak ve sınıflandırma performansını artırmak amacıyla, öznitelik seçimi yöntemleri ön plana çıkmaktadır. Öznitelik seçimi, yüksek boyutlu gen ekspresyon verilerinde yalnızca sınıflandırma açısından anlamlı ve etkili genleri belirleyerek hem gürültüyü azaltmakta hem de modellerin genellenebilirliğini artırmaktadır. Bu kapsamda geliştirilen yöntemler, çeşitli kategorilere ayrılmakta ve her biri farklı avantajlar sunmaktadır. Bu tez çalışmasında, 11 farklı kanser türüne ait mikrodizi gen ekspresyon veri setleri kullanılarak kapsamlı bir analiz gerçekleştirilmiştir. Çalışma kapsamında, her veri kümesi üzerinde çeşitli öznitelik seçimi yöntemleri hem tek başına hem de farklı kombinasyonlarla uygulanmıştır. Özellikle özellik seçim yöntemlerinin birlikte kullanılmasıyla oluşturulan öznitelik kümeleri hem bireysel hem de birleşim/kesişim stratejileriyle analiz edilmiştir. Bu sayede 12 farklı öznitelik seçimi kombinasyonu oluşturulmuş ve her biri yedi farklı hem klasik hem topluluk sınıflandırma algoritmasıyla test edilmiştir. Ayrıca, en başarılı öznitelik seçimi-sınıflandırıcı kombinasyonları istatistiksel olarak Friedman ve Nemenyi testleri ile karşılaştırılmıştır. Bu yaklaşım, öznitelik seçiminin sınıflandırma başarısı üzerindeki etkisini çok yönlü bir şekilde ortaya koymakta ve yüksek boyutlu biyolojik verilerin işlenmesinde etkili yöntemlerin belirlenmesine katkı sağlamaktadır.

Bilim Kodu : 92444

Anahtar Kelimeler : Kanser, gen ekspresyonu, DNA mikrodizi, öznitelik seçimi, topluluk öğrenimi, sınıflandırma

Sayfa Adedi : 121

Danışman : Dr. Öğretim Üyesi Yılmaz ATAY

CANCER CLASSIFICATION USING DNA MICROARRAY DATA WITH ENSEMBLE LEARNING

(M. Sc. Thesis)

Nazrin ABILOVA

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2025

ABSTRACT

Cancer is one of the leading causes of mortality worldwide. In recent years, advances in biotechnology have significantly contributed to the diagnosis and prognosis of cancer through molecular biomarkers such as gene expression data. In particular, DNA microarray technology enables the simultaneous measurement of the expression levels of tens of thousands of genes, providing a more detailed understanding of the molecular dynamics of cancer. However, these datasets typically exhibit a "high-dimensional, low-sample-size" structure, which poses challenges for traditional machine learning algorithms due to the risk of overfitting and increased complexity in data analysis. To address these issues and improve classification performance, feature selection methods have become increasingly important. Feature selection aims to identify the most informative and relevant genes for classification, thereby reducing noise and enhancing model generalizability in high-dimensional gene expression data. These methods are categorized into various types, each offering distinct advantages. In this thesis, a comprehensive analysis was conducted using microarray gene expression datasets from 11 different cancer types. Various feature selection techniques were applied to each dataset both individually and in combination. In particular, feature subsets created through combinations of selection methods were evaluated using union and intersection strategies. As a result, 12 different feature selection combinations were generated and tested using seven different classification algorithms, including both traditional and ensemble methods. Furthermore, the most effective feature selection–classifier combinations were statistically compared using the Friedman and Nemenyi tests. This approach provides a multidimensional assessment of the impact of feature selection on classification performance and contributes to identifying effective strategies for processing high-dimensional biological data.

Science Code : 92444
Key Words : Cancer, gene expression, DNA microarray, feature selection, ensemble learning, classification
Page Number : 121
Supervisor : Asst. Prof. Dr. Yılmaz ATAY

TEŞEKKÜR

Bu tez çalışmasının her aşamasında akademik bilgisi, yol gösterici yaklaşımı ve sabrıyla bana rehberlik eden, desteğini ve güvenini daima hissettiren değerli danışmanım Dr. Öğr. Üyesi Yılmaz ATAY'a en derin teşekkürlerimi sunarım. Sadece bilimsel katkılarınızla değil, aynı zamanda bana olan inancınız ve teşvik edici tutumunuzla bu sürecin hem verimli hem de ilham verici geçmesini sağladınız. Tez konusunun şekillenmesinden elde edilen sonuçların değerlendirilmesine kadar geçen tüm süreçte gösterdiğiniz ilgi ve katkı, akademik yolculuğumda unutulmaz bir yer edindi. Size duyduğum saygı ve minnettarlığı ifade etmek kelimelerle sınırlı kalacaktır.

İÇİNDEKİLER

Sayfa

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

ÇİZELGELERİN LİSTESİ ix

ŞEKİLLERİN LİSTESİ xi

SİMGELER VE KISALTMALAR xii

1. GİRİŞ 1

2. LİTERATÜRDEKİ ÇALIŞMALAR 7

3. MATERYAL VE YÖNTEM 15

3.1. Veri Seti 17

3.2. Öznitelik Seçim Yöntemleri 18

3.2.1. Filtre yöntemler 19

3.2.2. Sarmal yöntemler 23

3.2.3. Topluluk yöntemleri 25

3.3. Makine Öğrenmesinde Modelleme ve Değerlendirme Süreçleri 29

3.3.1. Sınıflandırma algoritmaları 29

3.3.2. İstatistiksel testler 34

3.3.3. Değerlendirme ölçütleri 36

4. DENEYSEL ÇALIŞMALAR 37

4.1. Veri Setleri üzerinde Deneysel Sonuçlar 39

4.2. Literatürel Karşılaştırma 102

5. SONUÇ VE ÖNERİLER 109

Sayfa

KAYNAKLAR 113

ÖZGEÇMİŞ 121

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 3.1. Öznitelik Seçim Kombinasyonları ve Kısaltmaları 15

Çizelge 3.2. Gen Ekspresyonu Veri Setlerinin Özellikleri 18

Çizelge 4.1. Sınıflandırma Algoritmaları için Kullanılan Parametreler..... 38

Çizelge 4.2. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE70947)..... 41

Çizelge 4.3. Nemenyi Testi ile Performans Farklılıkları (GSE70947) 42

Çizelge 4.4. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE14520_U133A).... 46

Çizelge 4.5. Nemenyi Testi ile Performans Farklılıkları (GSE14520_U133A) 46

Çizelge 4.6. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE44076)..... 49

Çizelge 4.7. Nemenyi Testi ile Performans Farklılıkları (GSE44076) 50

Çizelge 4.8. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE53757)..... 53

Çizelge 4.9. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE6919_U95v2)..... 57

Çizelge 4.10. Nemenyi Testi ile Performans Farklılıkları (GSE6919_U95v2) 58

Çizelge 4.11. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE19804)..... 62

Çizelge 4.12. Nemenyi Testi ile Performans Farklılıkları (GSE19804) 62

Çizelge 4.13. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE63270)..... 66

Çizelge 4.14. Nemenyi Testi ile Performans Farklılıkları (GSE63270) 66

Çizelge 4.15. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE21510)..... 70

Çizelge 4.16. Nemenyi Testi ile Performans Farklılıkları (GSE21510) 71

Çizelge 4.17. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE50161)..... 75

Çizelge 4.18. Nemenyi Testi ile Performans Farklılıkları (GSE50161) 76

Çizelge 4.19. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE45827)..... 80

Çizelge 4.20. Nemenyi Testi ile Performans Farklılıkları (GSE45827) 81

Çizelge 4.21. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE28497)..... 85

Çizelge 4.22. Nemenyi Testi ile Performans Farklılıkları (GSE28497) 86

Çizelge 4.23. Özellik Seçim Yaklaşımları (FS0-2) ve Sınıflandırma Algoritmalarının Test Sonuçları 90

Çizelge **Sayfa**

Çizelge 4.24. Özellik Seçim Yaklaşımları (FS3-5) ve Sınıflandırma Algoritmalarının Test Sonuçları 93

Çizelge 4.25. Özellik Seçim Yaklaşımları (FS6-8) ve Sınıflandırma Algoritmalarının Test Sonuçları 96

Çizelge 4.26. Özellik Seçim Yaklaşımları (FS9-11) ve Sınıflandırma Algoritmalarının Test Sonuçları 99

Çizelge 4.27. GSE70947: Yöntem ve Literatür Karşılaştırması 102

Çizelge 4.28. GSE14520_U133A: Yöntem ve Literatür Karşılaştırması 103

Çizelge 4.29. GSE6919_U95v2: Yöntem ve Literatür Karşılaştırması 104

Çizelge 4.30. GSE19804: Yöntem ve Literatür Karşılaştırması 104

Çizelge 4.31. GSE44076: Yöntem ve Literatür Karşılaştırması 105

Çizelge 4.32. GSE53757: Yöntem ve Literatür Karşılaştırması 106

Çizelge 4.33. GSE63270: Yöntem ve Literatür Karşılaştırması 106

Çizelge 4.34. GSE45827: Yöntem ve Literatür Karşılaştırması 107

Çizelge 4.35. GSE28497: Yöntem ve Literatür Karşılaştırması 107

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 1.1. 2022'de erkeklerde en sık görülen kanser türlerinin coğrafi dağılımı..... 2

Şekil 1.2. 2022'de kadınlarda en sık görülen kanser türlerinin coğrafi dağılımı..... 3

Şekil 3.1. Önerilen yöntemin çalışma prensibi 16

Şekil 3.2. Filtre yöntemlerinin çalışma prensibi 19

Şekil 3.3. Sarmal yöntemlerinin çalışma prensibi 23

Şekil 3.4. Bagging yöntemlerin çalışma prensibi 26

Şekil 3.5. Boosting yöntemlerin çalışma prensibi 28

Şekil 3.6. Stacking yöntemlerin çalışma prensibi..... 29

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
ACO	Karınca Kolonisi Optimizasyonu
B	Birleşim
BSS	Between-class Sum of Squares
CD	Critical Difference
CFS-MLP	Correlation-based Feature Selection-Multi-layer Perceptron classifier
CHI	Ki-kare
CuMiDa	Curated Microarray Database
CSA	Guguk Kuşu Arama Algoritması
DEA	Differential Expression Analysis
DE	Diferansiyel Evrim
DISR-MLP	Double Input Symmetrical Relevance- Multi-layer Perceptron classifier
DNA	Deoksiribonükleik Asit
DÖ	Doğrusal Öğrenme
DSÖ	Dünya Sağlık Örgütü
DT	Karar ağacı
ERGS	Effective Range based Gene Selection
FCFB	Fast correlation-based filter
FGS	Fuzzy Gene Selection
FKs	Fisher Skor
FSVM	Fuzzy support vector machine
GA	Genetik Algoritma
GB	Gradient Boosting
GEO	Gene Expression Omnibus
GEP	Gen İfade Programlama
GSP	Gen Seçimi Programlaması

Kısaltmalar	Açıklamalar
HPV	Human Papillomavirus
IG	Bilgi Kazancı
K	Kesişim
kNN	En yakın komşu
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Lineer diskriminant analizi
LIMMA	Linear Models for Microarray Data
LR	Lojistik Regresyon
MI	Mutual Information
MLP	Multi-layer Perceptron classifier
mRMR	Minimum Fazlalılık Maksimum İlgililik
NB	Naive Bayes
NNE	Neural Network Ensemble
NSGA-II	Non-Dominated Sorting Genetic Algorithm 2
PCA	Principal Component Analysis
PSO	Parçacık Sürüsü Optimizasyonu
QMFOA	Quantum Moth Flame Optimization Algorithm
RFE	Recursive Feature Elimination
RF	Random Forest
ROC	Receiver Operating Characteristic
SAE	Yığınlanmış derin otokodlayıcı
SBS	Sıralı Geri Seçim
SCA	Sin-Cosine Algoritması
SEER	Surveillance, Epidemiology, and End Results Program
SFS	Sıralı İleri Seçim
SMOTE	Synthetic Minority Oversampling Technique
SNR	Sinyal-Gürültü Oranı
SVM	Support Vector Machine
TL	Transfer Learning
TS	Tabu Arama
WSS	Within-class Sum of Squares

Kısaltmalar	Açıklamalar
WVCNN	Weighted CNN
YSA	Yapay Sinir Ağları

1. GİRİŞ

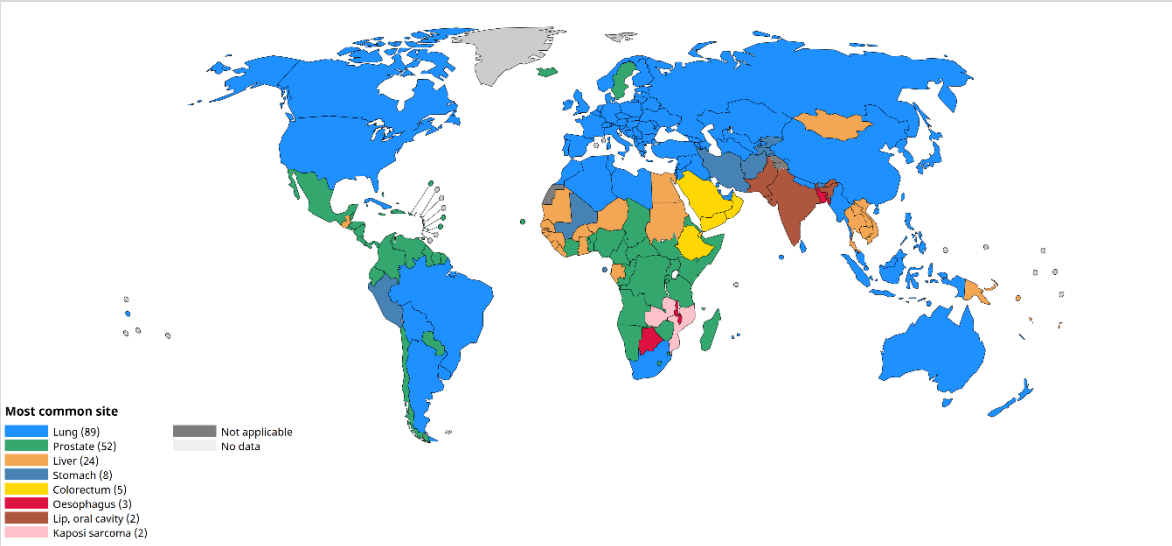
Vücuttaki sağlıklı hücreler kontrolden çıkıp diğer dokulara sızdığında veya diğer organlara yayıldığında, kanser adı verilen bir dizi hastalık gelişir. Normal hücreler, zamanla yaşlandıkça veya hasar gördükçe yenilenir. Bazen bu süreç tehlikeye girer ve anormal hücrelerin gereksiz yere çoğalmasına ve büyümesine izin verir. Bunun sonunda anormal türde hücreler tümöre dönüşebilir. Bu tümörler iyi veya kötü huylu olabilir. Kötü huylu tümörleri temsil eden kanserli hücreler, hızla çoğalarak diğer dokulara zarar verir. Bu hücreler organlar arasında hareket etme ve diğer yerlere yayılma yeteneğine sahiptir. Bu yayılma süreci "metastaz" olarak bilinir ve kanserin başlıca nedenidir. Vücudun farklı yerlerinde kanserin farklı türleri ortaya çıkabilir ve çevresel faktörler, bu süreçte genetik faktörler kadar etkilidir. Kansere sebep olan genetik değişiklikler, hücrelerin kontrolsüz bölünmesine neden olur. Tümörler, bu bozulmuş sürecin bir sonucu olarak gelişir. Dünya Sağlık Örgütü (DSÖ), kanseri "kontrolsüz şekilde çoğalan hücrelerin vücudun farklı bölgelerinde yayılmasıyla ortaya çıkan hastalık" olarak tanımlamaktadır [1]. Kanser türleri, farklı dokulara ve organlara yerleşebildiğinden, her bir türün tedavi yöntemleri de farklılık göstermektedir. Kansere sebep olan tümörlerin erken süreçte teşhis ve tespiti hastalığın yayılmasını durdurabilir.

Kanserin ilerlemesinde üç farklı genetik unsur rol oynar. Bunlar proto-onkogenler, tümör baskılayıcı genler ve Deoksiribonükleik Asit (DNA) onarım genleridir. Proto-onkogenler, hücresel büyüme ile bölünmede normal işlev gören olan geni ifade eder. Bu genlerde oluşan mutasyon, bu genleri onkogenlere dönüştürerek kontrolsüz büyümeye sebep olabilir. Tümör baskılayıcı genler hücre bölünmesini düzenleyen ve kanser gelişimini önlemeye yardımcı olan genleri temsil eder. Bu tür genleri etkileyen herhangi bir mutasyon kontrolsüz çoğalma riskini artırır. DNA onarıcı genler ise hücrelerdeki DNA hasarlarını düzeltmekle görevlidir fakat burada oluşan mutasyonlar diğer genetik değişimlere yol açarak kanserli transformasyona neden olabilir [2].

DSÖ'nün 2020 yılında sunduğu rapora göre kanserin yaklaşık 10 milyon insanın ölümüne yol açtığı belirtilmiştir [1]. En yaygın görülen kanser türleri arasında meme (göğüs), akciğer, prostat, kolon ve rektum gibi hastalıklar bulunmaktadır. Bu kanser türleri, gelişmiş ve gelişmekte olan ülkelerde yüksek ölüm oranlarına sebep olmaktadır. 2020 yılı itibarıyla, meme kanseri 2.26 milyon, akciğer kanseri 2.21 milyon, kolon ve rektum kanserleri ise

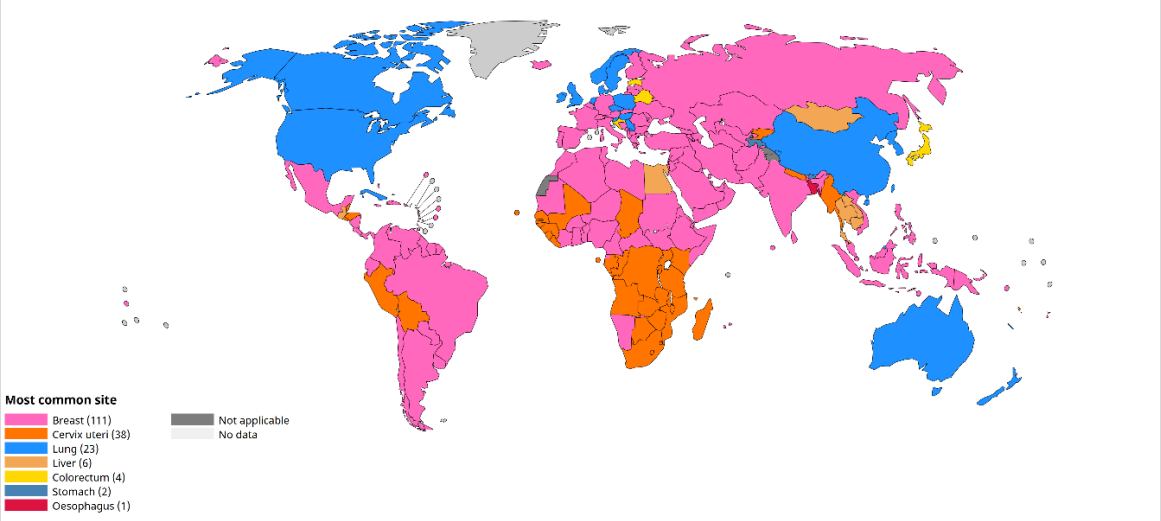
1.93 milyon vakayla en sık görülen kanser türleri olmuştur. Akciğer kanseri ise 1.80 milyon ölümle en fazla ölümle sonuçlanan kanser türüdür [1].

Dünya üzerindeki kanserin coğrafik dağılımı erkek ve kadınlar için farklılık göstermektedir. 2022 yılı itibariyle, erkeklerde en sık görülen kanser türlerine bakıldığında, akciğer kanseri dünya genelinde en yüksek ölüm oranına sahipken, özellikle gelişmiş bölgelerde daha yaygın görülmüştür. Bunun yanı sıra, prostat, kolon ve karaciğer gibi diğer kanser türleri de önemli yer tutmaktadır. Gelişmiş ülkelerde erken teşhis ve tedaviye erişim daha kolayken, gelişmekte olan ülkelerde bu oranlar daha düşüktür [3].



Şekil 1.1. 2022'de erkeklerde en sık görülen kanser türlerinin coğrafi dağılımı [3]

Kadınlarda ise 2022 yılında en sık görülen kanser türleri meme kanseri, serviks kanseri ve kolon kanseri olmuştur. Coğrafi farklar, sağlık hizmetlerine erişim ve toplumların sağlık alışkanlıklarına göre değişiklik göstermektedir. Meme kanseri, dünya genelinde kadınlar arasında en yaygın kanser türü olarak öne çıkarken, gelişmekte olan bölgelerde HPV (Human Papillomavirus) enfeksiyonlarının etkisiyle serviks kanseri yüksek oranda görülebilmektedir. Bu coğrafi farklar, kanserin tedavi ve önlenmesi konusundaki eşitsizlikleri gözler önüne sermektedir [3].



Şekil 1.2. 2022'de kadınlarda en sık görülen kanser türlerinin coğrafi dağılımı [3]

2022'de dünya çapında kadınlar ve erkekler arasındaki kanser türlerinin coğrafi dağılımı, sağlık politikalarının ve erken tarama programlarının önemini bir kez daha vurgulamaktadır. Kanserin tedavisinde eşitlik sağlanması ve önlenmesi adına, tüm dünyada sağlık hizmetlerine erişimin artırılması büyük bir öneme sahiptir.

2025 ve sonrasında dünya genelinde kanser vakalarının giderek artmaya devam etmesi bekleniyor. SEER (Surveillance, Epidemiology, and End Results Program) verilerine göre, kanser nedeniyle ölüm oranları ve yeni vaka sayılarının fark edilir şekilde artması öngörülmektedir. 2025'te dünya çapında yaklaşık 9.6 milyon ölüm beklenirken, kanserin yol açtığı ölümler hala yüksek oranda kalmaya devam etmektedir. Ayrıca, kanser tanısı konulacak bireylerin sayısının 2030'a kadar iki katına çıkması öngörülmektedir. Gelişmiş ülkelerde kanser tedavisinin daha erişilebilir olmasına rağmen, gelişmekte olan ve düşük gelirli ülkelerde bu tedaviye ulaşım hala kısıtlıdır. Bu durum, küresel sağlık sorunu açısından ciddi etkilere sebep olmaktadır [4].

Kanser çeşitleri, yaygın olarak ortaya çıktığı organ veya dokuya göre adlandırılır. Bununla birlikte, kanserler gelişen hücre tipine bağlı olarak da sınıflandırılabilir. Çoğu kanser türü epitel veya kan hücrelerinden köken alır. Kanser türleri genel hatlarıyla beş ana kategori altında incelenebilir: karsinomlar, sarkomlar, lösemiler, lenfomalar ve melanomlar [5]. Karsinomlar en yaygın olan kanser türüdür ve epitel hücrelerinden üretilmiştir; bu hücreler vücudun iç yüzeylerini kaplayan hücrelerdir. Bu grup kendi içinde adenokarsinomlar, bazal hücreli karsinomlar gibi alt kategorilere ayrılabilir. Sarkomlar ise bağ dokularında,

kaslarda ya da kemiklerde meydana gelen tümörlerdir; osteosarkom en bilinen kemik kaynaklı kansere örnektir. Lösemiler ise hematopoietik sistemden türeyen kanser türüdür; bu durumda, beyaz kan hücrelerinin kontrolsüz şekilde çoğalmaları, söz konusu olup yüksek oranda, anormal beyaz kan hücresine sahip olmakla karakterizedir, ki, bunlar normal işlevleri bozarak oksijen taşımayı oluşturma gibi işlemleri engelleyebilirler. Lenfoma, lenfatik sistemdeki lenf düğümlerinden kaynaklanan bir başka kanser grubudur; özellikle T lenfositleri ile B lenfositlerine etkisi olduğu bilinmektedir, ki, Hodgkin lenfoma ile Non-Hodgkin lenfoma öne çıkan başlıca iki alt gruptur ve buralarda anormal hücrelerin birikimi görülebilir. Melanomların kaynağı melanin üreten melanositlere dayanmaktadır. Bunlar genellikle cildin benlerinde başlamakta ancak yüzde pigmentli alanlarda da meydana gelebilmektedir. Melanomlar, en tehlikeli tür olarak görülmekte ve hızlı metastaz yapabilmektedirler [2].

Bu tez çalışmasında, farklı kanser türlerine ait gen ekspresyon verileri üzerinde çeşitli öznitelik seçim yöntemlerinin karşılaştırılması ve bu yöntemlerin sınıflandırma performansına olan etkilerinin değerlendirilmesi amaçlanmaktadır. Özellikle, yüksek boyutlu biyomedikal verilerde öznitelik seçimi sürecinin, sınıflandırma başarısını nasıl etkilediğiyle ilgili farklı analizler yapılmakta ve en etkili yöntemlerin belirlenmesi hedeflenmektedir. Çalışma kapsamında, öznitelik seçimi amacıyla Ki-kare (CHI), Fisher Skor (FKs), Bilgi Kazancı (IG), ReliefF ve Genetik Algoritma (GA) gibi farklı yöntemler uygulanmakta ve bunların başarımı istatistiksel testler aracılığıyla analiz edilmektedir. Elde edilen öznitelik alt kümeleri doğru sınıflandırma açısından hem klasik hem de topluluk tabanlı algoritmalar ile değerlendirilmekte ve elde edilen sınıflandırma sonuçları istatistiksel yöntemlerle analiz edilmektedir.

Çalışmanın temel odak noktası, öznitelik seçimi yöntemlerinin başarım farklılıklarını göstermek ve hangi yöntemlerin kanser sınıflandırma problemlerinde daha etkili sonuçlar sunabildiğini ortaya koymaktır. Ayrıca, topluluk yaklaşımlarının öznitelik seçimi ve sınıflandırma süreçlerinde sağladığı katkılar da ele alınmaktadır. Böylece daha kararlı, güvenilir ve genellenebilir modellerin geliştirilmesi amaçlanmaktadır.

Tez motivasyonu

Bu tez çalışmasının motivasyonu, yüksek boyutlu biyomedikal verilerle çalışmanın getirdiği zorluklara rağmen, farklı öznitelik seçimi yöntemlerinin kanser sınıflandırma performansı üzerindeki etkilerini derinlemesine analiz etmektir. Gen ekspresyon verileri, çok sayıda öznitelik içermesi ve genellikle sınıf dağılımlarının dengesiz olması nedeniyle doğrudan sınıflandırma algoritmalarıyla işlenmesi zor olan veri kümeleridir. Bu durum, doğru öznitelik seçiminin model başarımı üzerindeki kritik rolünü daha da ön plana çıkarmaktadır. Tezin temel motivasyonu, farklı öznitelik seçim yaklaşımlarının hangi durumlarda daha etkili ve tutarlı sonuçlar verdiğini belirlemek ve bunların sınıflandırma performansları üzerindeki etkilerini sistematik bir şekilde ortaya koymaktır.

Problem tanımı ve karşılaşılan zorluklar

Kanserin genetik düzeyde farklı türlerde ve çeşitli yapılarla görülmesi ve tek tip bir analizle açıklanamaması günümüzün biyomedikal araştırmalarında en önemli meselelerden biri olmuştur. Modern zamanda genomik veriler, kanseri moleküler düzeyde anlama konusunda önemli bir kaynak sağlamaktadır. Ancak bu tür verilerin hem boyutu yüksek olup binlerce özellik içermekte, hem örnek sayısı son derece kısıtlı olmaktadır. Yüksek boyutlu verilerle çalışırken model gürültülü veya alakasız verilere maruz kalabildiğinden ve bu durum da hatalı sınıflandırma ya da aşırı öğrenme (overfitting) sorunu oluşturduğundan yalnızca gerçekten anlam taşıyan ve bilgi sağlayan özniteliklerin dahil edilmesi gerekmektedir [6]. Ayrıca öznitelik seçim süreci biyomedikal verilerin yüksek boyutlarıyla çalışma açısından en kritik unsurlardan biri olarak karşımıza çıkmaktadır. Ancak, her bir öznitelik seçim yöntemi farklı matematiksel temellere ve veri karakteristiklerine dayandığı için, her durumda aynı derecede başarılı sonuçlar elde edilmesi mümkün değildir. Geleneksel öznitelik seçme yöntemleri olan filtre, sarmalayıcı ve gömülü yaklaşımlar genellikle verilere özgü olup birbirleriyle önemli ölçüde farklı çıktılar sağlayabilmektedirler. Bu nedenle, hangi yöntemin belirli koşullar altında daha etkili olduğunu sistematik bir biçimde analiz etmek hem akademik bilgi üretimi hem de uygulamalarda güvenilir modeller geliştirilmesi açısından son derece önemlidir [7].

Bir diğer büyük zorluk ise sınıflandırma algoritmalarının karmaşık biyomedikal veriler üzerinde güvenilir şekilde çalışmasını sağlamaktır. Gen ifade verilerinin ortak bir özelliği sınıf dengesizliğidir. Bazı kanser türleri gen ekspresyonu verilerinde aşırı temsil edilirken,

diğerleri sınırlı sayıda temsil edilmektedir [8]. Sınıflardaki oluşmuş eşitsizlikler, özellikle sağlık hizmetleri gibi kritik konularda zorluklara neden olabilir. Sınıf dengesizliği ele alınırken duyarlılık (recall), precision (kesinlik) ve F1 skoru (F1 score) gibi metrikler, doğruluk kadar dikkate alınması gereken metriklerdir. Her bir metrik model performansının farklı bir yönünü vurgulayarak, daha kapsamlı bir anlayış sunar.

Sonuç olarak, bu tez çalışması yalnızca farklı kanser türlerine ait yüksek boyutlu gen ekspresyon verilerinin sınıflandırılmasını hedeflemekle kalmıyor; aynı zamanda öznitelik seçimi yöntemlerinin sınıflandırma performansına etkilerini sistematik olarak inceleyerek, bu süreçte karşılaşılan metodolojik ve teknik zorluklara çözüm arayan kapsamlı bir yaklaşım sunuyor.

Bu çalışmanın ikinci bölümünde, ilgili literatürde yapılmış benzer çalışmalar ele alınmakta ve mevcut yaklaşımların güçlü ve sınırlı yönleri tartışılmaktadır. Üçüncü bölümde, kullanılan veri seti, öznitelik seçim yöntemleri ve makine öğrenmesine dayalı modelleme süreci ayrıntılı olarak açıklanmakta; ayrıca sınıflandırma algoritmaları, istatistiksel testler ve değerlendirme ölçütleri sunulmaktadır. Dördüncü bölümde, gerçekleştirilen deneysel çalışmalar ve elde edilen sonuçlar detaylı biçimde verilmekte; farklı veri setleri üzerinde yapılan analizler ve literatürle karşılaştırmalar yer almaktadır. Son olarak, beşinci bölümde ise elde edilen bulgular ışığında genel bir değerlendirme yapılmakta ve gelecekteki çalışmalara yönelik öneriler sunulmaktadır.

2. LİTERATÜRDEKİ ÇALIŞMALAR

Kanserin özellikle, gen ifadesi ve diğer biyolojik veriler aracılığıyla yapılan analizlerle tanısı önemli bir araştırma alanı olmuştur. Bu alandaki kaynaklar, yüksek boyutlu veri kümelerinde öznitelik seçim tekniklerinin geliştirilmesine ve uygulanmasına odaklanılmıştır. Öznitelik seçimi, doğrudan sınıflandırma modellerinin doğruluğunu ve genellenebilirliğini yükseltmesi nedeniyle kritik bir öneme sahiptir. Kaynaklarda, klasik öznitelik seçimleri ve hibrit metaheuristik (metasezgisel) yöntemler ile beraber, son senelerde topluluk tabanlı öznitelik seçim yöntemleri de uygulanmakta ve daha istikrarlı, güvenilir öznitelik alt kümleri elde ederek daha fazla tercih edilmektedir.

Literatürde yapılan çalışmalar incelendiğinde Shen ve ark. (2008) gen seçimi için hibrit bir Parçacık Sürüsü Optimizasyonu (PSO) ve Tabu Arama (TS) yaklaşımı (HPSOTS) geliştirmiştir. PSO, arama uzayında potansiyel çözümleri temsil ederken, TS yerel iyileştirme sağlar. HPSOTS, PSO'nun çeşitliliğini artırarak yerel optimumlardan kaçınmak için TS'yi entegre etmiştir. Algoritma, ikili bit dizileri ve TS ile performansı iyileştirmiştir. Testlerde HPSOTS, TS ve PSO'ya kıyasla daha yüksek doğruluk sağlamış; kolon veri setinde %93,55, östrojen veri setinde %97,96 ve AML/ALL veri setinde %98,61 doğruluk elde edilmiştir [9].

Chandra ve Gupta (2011), kanser gen ekspresyonu verilerinin sınıflandırılmasında kullanılan çeşitli gen seçim yöntemlerini incelemişlerdir. Bu çalışmada, ERGS (Effective Range based Gene Selection) algoritması önerilmiş ve diğer popüler yöntemlerle karşılaştırılmıştır. ERGS, her özelliğin sınıfları net bir şekilde ayırma kapasitesine göre ağırlıklandırılmasını sağlayarak gen seçiminde verimli sonuçlar elde edilmesini sağlaryarak, aykırı değerlerin ve yüksek varyansa sahip sınıfların etkisini ortadan kaldırır. Çalışma, ERGS algoritmasının, Relief-F, Minimum Fazlalılık Maksimum İlgililik (mRMR), t-Statistic ve diğer yaygın yöntemlerle karşılaştırıldığında, gen ekspresyonu veri setlerinde daha yüksek doğruluk sağladığını göstermektedir. Özellikle, ALL_AML veri setinde top 10 genle %98.61 doğruluk elde edilmiştir [10].

Hui-Huang Hsu ve ark. (2011) tarafından geliştirilen çalışmada, hibrit öznitelik seçimi mekanizması tasarlanmış ve iki farklı biyoinformatik problem üzerinde test edilmiştir. Bu hibrit mekanizma, filtreleme yöntemleri ve sarmal yöntemlerin verimli bir şekilde

birleştirilerek, özelliklerin optimizasyonunu amaçlamaktadır. Çalışmanın amacı, gereksiz özniteliklerin ortadan kaldırılması ve sınıflandırma doğruluğunun artırılmasıdır. FKs ve IG gibi filtreleme tekniklerinin birleşimi, önemli özniteliklerin seçilmesinde etkili olmuştur. Denemeler, hibrit mekanizmanın, öznitelik kümesini küçültürken doğruluğu artırabileceğini göstermiştir. Disordered protein prediksiyon ve mikrodizi kanser veri setlerinde, orijinal veri setlerinden özellik sayısı azalırken doğruluk oranları önemli ölçüde artmıştır. Bu sonuçlar, yazarların hibrit öznitelik seçimi mekanizmasının biyoinformatik problemlerde etkili bir çözüm sunduğunu ortaya koymuştur [11].

Lorena ve ark. (2012) kanser gen ekspresyonu verilerinin sınıflandırılmasında kullanılan çeşitli teknikleri incelemişlerdir. Çalışmada ilk olarak BSS/WSS (Between-class Sum of Squares/Within-class Sum of Squares) gen seçimi kullanılmıştır. Bu gen seçimi yöntemi, Fisher lineer diskriminant analizi (LDA) ile birleştirilerek sınıflar arasındaki ayrımı artırmak için uygulanmıştır. Ardından, doğrusal SVM (Support Vector Machine) sınıflandırıcıları kullanılarak sınıflandırma işlemi yapılmıştır. Ayrıca, gürültülü veri setleriyle yapılan testlerde, sınıflar arasındaki farkları ayırt etmek için kullanılan karmaşıklık indeksleri ile veriler analiz edilmiştir. Sonuçta, BSS/WSS gen seçimi ile veri setlerinde sınıflandırma doğruluğu artırılmış, karmaşıklık azaltılmış ve genetik verilerle yapılan sınıflandırmalarda daha iyi performans elde edilmiştir [12].

Mahmut Kaya (2014) sunulan tez çalışmasında, özellik seçimi ve problem temelli ağırlıklandırma tekniklerinin sınıflandırma başarısını nasıl etkileyebileceği incelemiştir. Çalışmanın temel amacı, genetik veri kümelerinde özellik seçimi yöntemlerinin sınıflandırma doğruluğunu artırmak olmuştur. Yapılan deneyler, FKs ve t-Test gibi istatistiksel yöntemlerle özniteliklerin seçilmesinin başarıyı artırdığını göstermiştir. Sonuçlar, Naive Bayes (NB) sınıflandırıcısının genellikle en iyi sonuçları verdiğini, ancak SVM sınıflandırıcısının geri arama yöntemleriyle daha iyi performans sergilediğini ortaya koymuştur. Ayrıca, öznitelik ağırlıklandırma ile k-en yakın komşu (kNN) sınıflandırıcısının başarı oranında belirgin bir artış gözlemlenmiştir [13].

Chen ve ark. (2014) mikrodizi teknolojisi ile gen seçimini optimize etmek için PSO ve C4.5 Karar Ağacı (DT) tabanlı bir yöntem (PSODT) önermiştir. PSO, gen alt kümelerini temsil ederken C4.5, sınıflandırma doğruluğunu değerlendirir. 11 veri seti üzerinde test edilen PSODT, %97,26 doğruluk ile SVM ve diğer yöntemlerden üstün sonuçlar elde

etmiştir. Çalışma, PSO ve C4.5 birleşiminin gen seçimi sorunlarında etkili ve maliyet açısından avantajlı olduğunu göstermiştir [14].

Shanab ve ark. (2014) mikrodizi veri setlerindeki yüksek boyutluluk ve gürültü sorunlarını ele almak için sarmalayıcı tabanlı öznitelik seçimini değerlendirmiştir. NB algoritmasıyla yapılan öznitelik seçimi sonrası, farklı algoritmalarla sınıflandırma gerçekleştirilmiştir. Sonuçlar, sarmal yönteminin özellikle MLP (Multi-layer Perceptron classifier) için faydalı olduğunu ancak diğer algoritmalarda performansı düşürdüğünü göstermiştir. Çalışma, özniteliklerin önemini belirlemede ve tekrarlayan öznitelikleri azaltmada sarmalayıcı yönteminin etkili olduğunu vurgulamaktadır [15].

Tabakhi ve ark. (2015) gen seçimi için yeni bir Karınca Kolonisi Optimizasyonu (MGSACO) yöntemi geliştirmiş ve mikrodizi veri setleri üzerinde test etmiştir. Beş kanser türüyle yapılan çalışmada, MGSACO'nun SVM sınıflandırıcısı ile tüm veri setlerinde en düşük hata oranını sağladığı ve lösemi veri setinde %17,94 hata oranı ile diğer yöntemleri geride bıraktığı görülmüştür. Ayrıca NB ve DT sınıflandırıcılarıyla da üstün performans sergilemiş ve mRMR yöntemine kıyasla %11 hata oranıyla önemli bir avantaj sağlamıştır. Yöntem, maksimum ilgililik ve minimum fazlalılık içeren genleri etkili bir şekilde seçerek yüksek boyutlu veri setlerinde başarılı sonuçlar sunmaktadır [16].

Aslı Taşçı, 2016 tarihli tez çalışmasında, mikrodizi veri kümelerine uygulanan gen seçimi ve sınıflandırma tekniklerini inceleyerek çeşitli özellik seçim yöntemlerinin sınıflandırma başarısı üzerindeki etkisini değerlendirmeyi amaçlamıştır. Araştırmanın birincil amacı, farklı kanser türleri için yüksek sınıflandırma doğruluğu elde etmek amacıyla mikrodizi verilerinden en alakalı genleri titizlikle belirlemek olmuştur. Yapılan deneylerde farklı öznitelik seçme algoritmalarının kullanılmasıyla gen alt kümeleri seçilmiş ve bu alt kümeler üzerinde sınıflandırma işlemleri gerçekleştirilmiştir. Sonuçlar, özellikle CFS-MLP (Correlation-based Feature Selection-Multi-layer Perceptron classifier) ve DISR-MLP (Double Input Symmetrical Relevance-Multi-layer Perceptron classifier) kombinasyonlarının veri setlerinde tam doğrulukla başarı sağladığını göstermiştir [17].

Abubakhari Sserwadda (2027), çalışmasında meme kanseri metastazının tahmin edilmesi için gen seçimi ve sınıflandırma stratejilerini araştırmıştır. Çalışmanın amacı, genetik verilerdeki özniteliklerin doğru seçilmesiyle sınıflandırma doğruluğunun arttırılmasıydı.

Beş farklı meme kanseri mikrodizi veri seti üzerinde yapılan deneylerde, LASSO (Least Absolute Shrinkage and Selection Operator), DEA (Differential Expression Analysis) ve Random Forest (RF) gibi öznitelik seçim yaklaşımlarının performansı değerlendirilmiştir. Sonuçlar, özellikle kesişim alt kümelerinin kullanılmasıyla sınıflandırma doğruluğunun %73-82 arasında değişen oranlara ulaştığını ve bu yöntemlerin genetik verilerde daha güvenilir biyomarkerler seçilmesine yardımcı olduğunu göstermiştir [18].

Alanni ve ark. (2019) kanser sınıflandırması için Gen Seçimi Programlaması (GSP) adlı bir yöntem önermiştir. GSP, Gen İfade Programlama (GEP) algoritmasını temel alarak, rastgele işlemleri sistematik gen sıralama tabanlı seçimle değiştirmiştir. 10 mikrodizi veri seti üzerinde doğrusal çekirdekli SVM ile yapılan testlerde, GSP %99.92 doğruluk ve en az sayıda genle yüksek performans göstermiştir. Yüksek doğruluk ve istikrar sağlasa da işlem süresinin uzunluğu geliştirme gereksinimi olarak belirtilmiştir [19].

Gizem Yağmur (2019), meme kanseri hastalarının sağ kalım tahminini artırmak amacıyla SEER veri setini kullanarak öznitelik seçimi ve sınıflandırma algoritmalarını incelemiştir. Çalışmanın amacı, 73 farklı öznitelikten en faydalı olanları seçmek ve sınıflandırma doğruluğunu artırmaktı. Üç farklı filtreleme yöntemi kullanarak, Sıralı İleri Seçim (SFS) yöntemiyle öznitelikler daha da azaltılmıştır. Sonuçlar, J48 ve SVM algoritmalarının en yüksek doğruluk oranlarını ve F-skorlarını verdiğini, öznitelik sayısının azalmasının başarıyı artırmaya veya korumaya yardımcı olduğunu göstermektedir. Bu çalışma, meme kanseri sağ kalım tahmini için öznitelik seçimi ve sınıflandırma doğruluğunu artırmak adına başarılı bir yaklaşım sergilemiştir [20].

Candan ve ark. (2019) akciğer ve beyin kanseri verilerinde GA ile öznitelik seçimini uygulayarak NB, kNN, RF ve SVM gibi algoritmaların performansını karşılaştırmıştır. Çalışma, GA'nın SVM'in doğruluğunu artırdığını (%93,59'dan %94,09'a) ve beyin kanseri verilerinde %4,44 iyileşme sağladığını göstermiştir. Ancak NB'de bazı durumlarda performans düşüşü gözlemlenmiştir. GA ile öznitelik seçiminin, kanser teşhisinde önemli avantajlar sunduğu vurgulanmıştır [21].

Özlem Arık (2020), altı farklı kanser türü üzerinden genetik verilere uygulanan çeşitli özellik seçim yöntemlerinin ve sınıflandırma algoritmalarının etkinliğini değerlendirmiştir. Araştırmanın amacı, gen ifadesi profillerindeki çok sayıda özniteliklerden yalnızca en

anlamlı özellikleri belirleyerek sınıflandırma doğruluğunu artırmaktı. Çeşitli öznitelik seçim yöntemleri, özellikle LASSO ve LIMMA (Linear Models for Microarray Data) ile birlikte kullanıldığında doğruluk oranlarını artırmıştır. Doğrusal Öğrenme (DÖ) algoritması, en iyi sonuçları verirken, Yapay Sinir Ağları (YSA) genellikle düşük başarılar göstermiştir. Çalışmada, öznitelik seçimiyle sınıflama doğruluğunun arttığı ve doğru özniteliklerin seçilmesinin genetik veri analizinde başarıyı artırdığı vurgulanmıştır [22].

Emanet ve ark. (2021), CSE-CIC-IDS2018 veri kümesinde Ki-kare, Spearman Korelasyonu ve Tekrarlayan Özellik Eleme (RFE) özellik seçimi yöntemlerinin performansını değerlendirmek için bir çalışma yürütmüş, ardından AdaBoost, DT, Lojistik Regresyon (LR) ve Gradient Boosting (GB) gibi sınıflandırma algoritmaları ile birlikte test edilmişdir. Yöntemler arasında RFE, %98,76'lık bir doğruluk elde ederek en yüksek performansı göstermiştir. Ki-kare ve Spearman Korelasyonu veri boyutunu azaltmış ve işlem sürelerini kısaltmış olsa da doğrulukta bir azalmaya neden olmuştur. Çalışma, gelecekteki araştırmalarda hibrit modeller kullanılarak daha etkili saldırı tespit sistemlerinin geliştirilebileceğini öne sürmektedir [23].

Wang ve ark. (2022) mikrodizi gen ifade verilerinden biyomarker seçimi ve kanser sınıflama doğruluğunu artırmak için bir topluluk öznitelik seçimi yöntemi önermiştir. SRBCT, Kolon, DLBCL ve Lösemi gibi dört veri seti kullanılarak yapılan çalışma, yüksek boyutlu veri problemlerini değerlendirmiştir. Sonuçlar, önerilen yöntemin hem özelliklerin kararlılığını hem de sınıflandırma doğruluğunu artırdığını ve mevcut yöntemlerden daha iyi performans gösterdiğini göstermiştir [24].

Umut Akcan (2022), Karınca Kolonisi Optimizasyonu'na (ACO) dayalı bir özellik seçimi yöntemi geliştirerek sınıflandırma başarısını iyileştirmeyi hedeflemiştir. Çalışma boyunca, farklı öznitelik seçimi teknikleri kullanılarak her bir yöntem değerlendirilmiş ve ACO, kosinüs benzerlik yöntemiyle öznitelikler arasındaki sezgisel uzaklıkları hesaplamıştır. Elde edilen sonuçlar, önerilen ACO tabanlı yaklaşımın, özellikle SVM sınıflandırma yönteminde daha yüksek doğruluk oranları sağladığını ve literatürdeki diğer yöntemlerle karşılaştırıldığında WDBC, Dermatology ve Spambase veri setlerinde daha başarılı olduğunu göstermektedir. Sonuç olarak, bu çalışma, ACO tabanlı öznitelik seçiminin yüksek doğruluk oranları sağladığını ve gelecekte büyük veri setlerinde daha kapsamlı çalışmalarla daha iyi sonuçlar elde edilebileceğini ortaya koymuştur [25].

KhosroRezaee ve ark. (2022) mikrodizi verilerinden gen seçimi ve kanser sınıflandırması için derin öğrenme tabanlı bir yaklaşım önermiştir. Yığınlanmış derin otokodlayıcı (SAE) algoritmasıyla yüksek seviye özellikler elde edilmiş ve RNA-Seq verileri 3 farklı kanser türü üzerinde test edilmiştir. Sinyal-Gürültü Oranı (SNR), Wilcoxon testi ve ROC (Receiver Operating Characteristic) yöntemleriyle gen seçimi yapılmış; aşırı uyum, rastgele karıştırma ve normalleştirme ile önlenmiştir. Önerilen yöntemin, veri setlerinde yüksek doğruluk sağlayarak mevcut yöntemlerden üstün olduğu vurgulanmıştır [26].

Simangunsong ve Hidayat (2023), DNA mikrodizi verileri üzerinde Mutual Information (MI) ve RFE yöntemlerini birleştirerek özellik indirgeme kullanarak sınıflandırma süreçlerini optimize etmiştir. GSE18842, GSE19804 ve GSE27262 veri setlerinde MI yöntemiyle önemli öznitelikler seçilmiş, RFE ile indirgenmiş ve 10 katlı çapraz doğrulama uygulanmıştır. PyCaret destekli modeller, indirgenmiş özelliklerle %95-%98 doğruluk sağlamıştır. Çalışma, gen sayısını optimize ederek yüksek doğruluk ile düşük maliyetle akciğer kanseri teşhisi için etkili bir yöntem sunmaktadır [27].

Omar Ali Almajmaee (2023), öznitelik seçim tekniklerinin makine öğrenmesi modellerinin doğruluğunu nasıl artırabileceğini araştırmış ve çeşitli öznitelik seçimi stratejilerinin etkilerini değerlendirmiştir. Farklı denetimli, denetimsiz ve hibrit öznitelik seçimi yöntemleri ile öznitelikler belirlenmiş ve her bir yöntem, çeşitli veri setleri ve makine öğrenmesi görevleri üzerinde test edilmiştir. Özellikle, filtre, sarmal ve gömülü yöntemlerin doğru şekilde seçilmesinin modelin doğruluğunu artırdığı, ayrıca veri kalitesi ve gürültü ile baş eden tekniklerin önemli bir araştırma alanı olduğu vurgulanmıştır [28].

Melike Kılıç (2023), çalışmasında, farklı öznitelik seçimi yöntemlerini kullanarak mide kanseri verileri üzerinde yapılan makine öğrenmesi modellerinin doğruluğunu artırmayı hedeflemiş ve Boruta ve Anova gibi yöntemlerle miRNA biyomarkerlerini belirlemiştir. GSE106817 ve GSE113486 veri setlerini kullanarak, her iki yöntemle yapılan öznitelik seçiminde yüksek doğruluklar elde edilmiştir. Özellikle, Boruta ve Anova yöntemleri ile seçilen 293 miRNA, RF ve SVM gibi modellerle test edilmiştir. Çalışma sonuçları, miRNA'ların kanser türlerini ayırt etmede etkili biyomarkerler olduğunu göstermiş ve Boruta yöntemi genellikle daha yüksek doğruluk sağlasa da Anova yöntemi daha verimli sonuçlar üretmiştir. Bu çalışma, pan-kanser analizlerinde miRNA'ların potansiyelini ortaya

koyarak, gelecekteki çalışmalarda derin öğrenme ve yapay sinir ağları gibi gelişmiş tekniklerle daha ileri düzey modellerin geliştirilmesini önermektedir [29].

Kinjal ve diğerleri (2023), makine öğreniminde topluluk öğrenme yöntemlerinin uygulanmasını araştırmış ve bu yöntemlerin birden fazla modeli birleştirerek daha yüksek doğruluk oranlarına ulaşabileceğini göstermiştir. Araştırma, topluluk yöntemlerinin, özellikle de bagging, boosting ve stacking tekniklerinin, karmaşık ve büyük veri kümeleriyle çalışırken geleneksel yöntemlere kıyasla üstün performans sağladığını vurgulamıştır. Bu yöntemler, modellerin varyansını azaltarak daha stabil sonuçlar elde edilmesini sağlarken, zayıf öğrenicileri güçlü öğrenicilere dönüştürerek doğruluğu artırmaktadır. Ayrıca, topluluk yöntemleri, aşırı uyum sorununu azaltarak modelin genelleme kabiliyetini geliştirmektedir [30].

Buse Nur Baltacıoğlu (2024), çalışmasında, makine öğrenmesi öznitelik seçimi yöntemlerinin sağlık verileri üzerindeki etkinliğini araştırmış ve farklı tekniklerle elde edilen sonuçları karşılaştırmıştır. Çalışma, dört veri seti üzerinde test edilmiştir. Boruta ve Anova gibi yöntemlerle seçilen miRNA'lar, RF ve SVM modelleriyle test edilerek yüksek doğruluk oranları elde edilmiştir. Özellikle, gömülü yöntemlerin sağlık veri kümelerinde daha yüksek doğruluk sağladığı gözlemlenmiştir. Çalışma, sağlık verileri gibi karmaşık veri kümelerinde doğru öznitelik seçim tekniklerinin kullanımının model başarısını artırmak için kritik olduğunu ve gelecekte bu yöntemlerin daha büyük veri kümeleriyle otomatikleştirilerek geliştirilmesi gerektiğini vurgulamaktadır [31].

Yaqoob ve ark. (2024) gen seçimi ve kanser sınıflandırmasını optimize etmek için Sin-Cosine Algoritması (SCA) ile Guguk Kuşu Arama Algoritması (CSA) birleşimini içeren hibrit bir yöntem (SCACSA) önermiştir. Çalışmada, mRMR yöntemi ile gereksiz özellikler filtrelenmiş ve ardından SCACSA kullanılarak gen seçimi gerçekleştirilmiştir. SVM ile yapılan sınıflandırma testlerinde, önerilen yöntemin diğer öznitelik seçme algoritmalarına kıyasla daha yüksek doğruluk sağladığı belirlenmiştir. Deneysel sonuçlara göre, SCACSA, 30 gen seçerek %99.53 doğruluk oranına ulaşmış ve mevcut yöntemlerden üstün performans göstermiştir. Çalışma, doğadan esinlenen optimizasyon algoritmalarının biyomedikal veri analizinde güçlü bir araç sunduğunu ve kanser teşhisinde gen seçiminin doğruluk üzerindeki kritik rolünü vurgulamaktadır [32].

Literatürde yapılan çalışmalar, gen seçimi ve sınıflandırma algoritmalarının kanser teşhisi için önemli katkılar sunduğunu göstermektedir. Çalışmaların çoğu ya tekil algoritmalara ya da belirli veri setlerine odaklandığından topluluk tabanlı yaklaşımların hem öznitelik seçimi hem de sınıflandırma süreçlerinde sistematik ve bütünsel bir şekilde uygulanması hala gelişime açıktır. Bu tez çalışması, literatürdeki mevcut çalışmalardan farklı olarak, farklı kanser türlerinde etkili özelliklerin belirlenmesinde kapsamlı analizlerle birlikte, birden fazla ensemble öznitelik seçim yöntemini karşılaştırmalı olarak değerlendirerek, yüksek boyutlu biyomedikal verilerde yüksek başarımlı, açıklanabilir, daha hızlı ve kararlı sonuçlar elde etmeyi amaçlamaktadır.

3. MATERYAL VE YÖNTEM

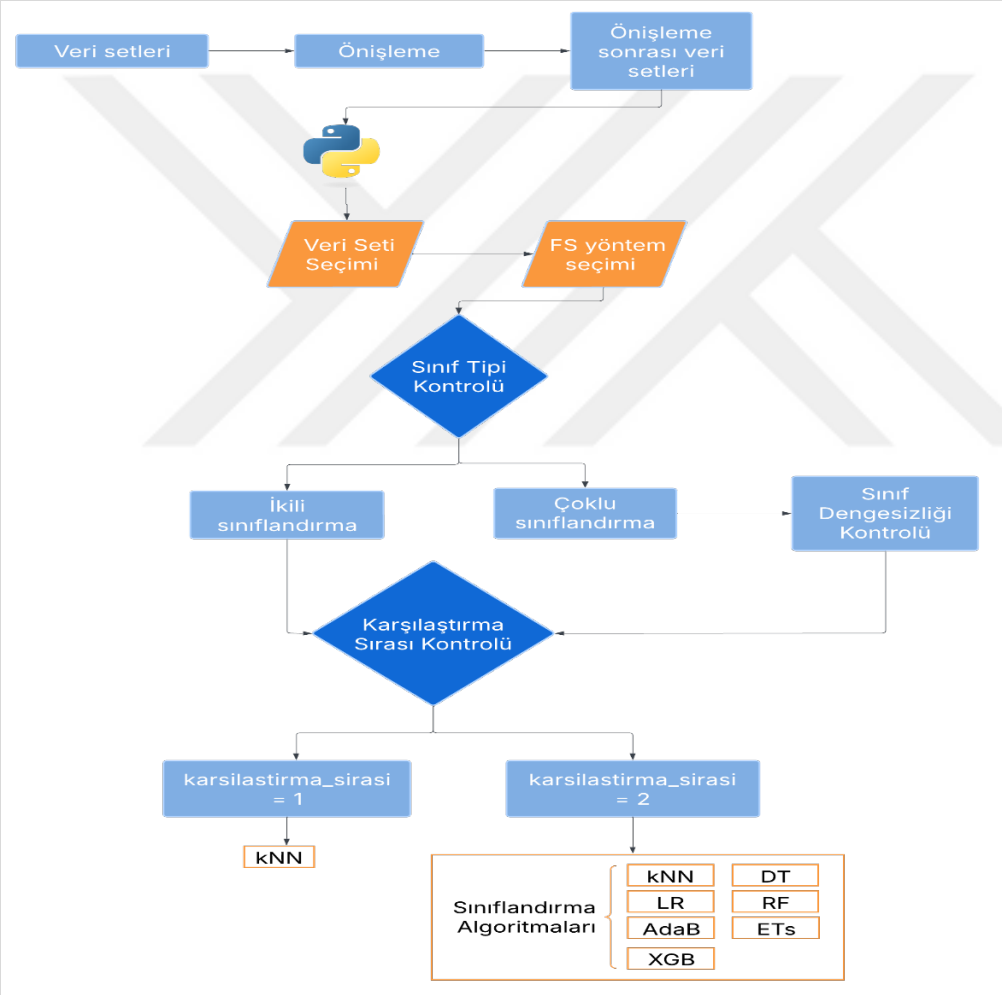
Çalışmada önerilen yöntem, biyomedikal veri setleri üzerinde sınıflandırma problemlerine yönelik sistematik bir modelleme ve değerlendirme süreci içermektedir. Bu süreç, kullanıcı tarafından incelenmek istenen veri setinin girdi olarak sunulmasıyla başlamaktadır. İlk aşamada öncelikle veri ön işleme adımları uygulanmakta; eksik değerler ortalama ile tamamlanmakta, sınıf etiketleri sayısal formata dönüştürülmekte ve bilgi taşımayan öznitelikler veri setinden çıkarılmaktadır. Bu sayede analiz için daha uygun bir veri yapısı elde edilmektedir. Veri temizliğinin ardından, sınıf yapısı kontrol edilmekte ve veri setinin sınıf sayısına göre farklı işlem akışları uygulanmaktadır. Bu çalışmada ikili ve çoklu sınıflandırma problemine odaklanılmaktadır. Sınıflar arasında dengesizlik tespit edilmesi durumunda, azınlık sınıf örneklerinin artırılması amacıyla SMOTE (Synthetic Minority Oversampling Technique) yöntemi devreye alınmaktadır.

Öznitelik seçim sürecinde hem filtre tabanlı hem de sarmal yöntemler birlikte değerlendirilmektedir. Ki-kare, FKs, IG ve ReliefF gibi filtre yöntemlerinin yanı sıra, GA kullanılarak sarmal seçim yapılmaktadır. Elde edilen öznitelik listeleri tek-tek değerlendirilmenin yanı sıra, birleşim (B) ve kesişim (K) stratejileriyle de birleştirilerek 12 farklı öznitelik seçimi kombinasyonu oluşturulmuş ve Çizelge 3.1'de sunulmuştur. Birleşim yaklaşımıyla daha kapsamlı öznitelik kümeleri elde edilirken, kesişim yaklaşımı tüm yöntemlerin ortaklaştığı özniteliklere odaklanmaktadır.

Çizelge 3.1. Öznitelik Seçim Kombinasyonları ve Kısaltmaları

Kombinasyon	Kısaltma
Özellik seçimi olmadan	*FS0*
Ki-kare	*FS1*
FKs	*FS2*
IG	*FS3*
ReliefF	*FS4*
Ki-kare + FKs + IG + ReliefF + Genetik Algoritma + B	*FS5*
Ki-kare + FKs + IG + ReliefF + Genetik Algoritma + K	*FS6*
Ki-kare + FKs + IG + ReliefF + B	*FS7*
Ki-kare + Genetik Algoritma + B	*FS8*
FKs + Genetik Algoritma + B	*FS9*
IG + Genetik Algoritma + B	*FS10*
ReliefF + Genetik Algoritma + B	*FS11*

Sınıflandırma işlemi 2 senaryoda gerçekleştirilmektedir. İlk senaryoda (karsilastirma_sirasi = 1), yalnızca k-NN algoritması kullanılarak farklı öznitelik kombinasyonlarının performansı karşılaştırılmaktadır. Böylece öznitelik seçim yöntemlerinin k-NN üzerindeki etkisi detaylı biçimde incelenmektedir. İkinci senaryoda (karsilastirma_sirasi = 2), daha geniş bir sınıflandırıcı seti kullanılarak değerlendirme yapılmaktadır. Bu aşamada RF, AdaBoost, Extra Tress (ETs) ve XGBoost gibi topluluk öğrenme algoritmalarının yanı sıra, DT, LR ve kNN gibi temel yöntemlerle her öznitelik seçimi kombinasyonu test edilmektedir. Önerilen yöntemin akış şeması Şekil 3.1'de sunulmuştur.



Şekil 3.1. Önerilen yöntemin çalışma prensibi

Bu bölümün devamında, tez çalışmasında kullanılan veri kümesi kaynakları, öznitelik seçimi teknikleri, sınıflandırma modelleri için performans değerlendirme ölçütleri ve istatistiksel analiz yöntemleri ayrıntılı bir şekilde sunulmuştur.

3.1. Veri Seti

Çalışmada kullanılan veri setleri Curated Microarray Database (CuMiDa) veri tabanından temin edilmiştir [33]. CuMiDa, gen ekspresyonu üzerine yapılan araştırmaların merkezinde yer alan ve dünya genelinde geniş çapta kullanılan Gene Expression Omnibus (GEO) veritabanında bulunan 30.000'den fazla çalışmanın sistematik olarak incelenmesi sonucunda oluşturulmuştur. Bu veri tabanı, hem ikili (binary), hem çoklu (multi) sınıflandırma problemlerine uygun olacak şekilde, farklı kanser türlerine ait genetik ifadeleri içeren veri setlerini kapsamaktadır.

Çalışmada, CuMiDa veri tabanından seçilen farklı organ ve dokulardaki çeşitli kanser türlerini temsil eden hem ikili sınıflandırma hem de çoklu sınıflandırma problemlerine yönelik toplam 11 gen ifadesi veri kümesi kullanılmıştır. İkili sınıflandırma veri setleri arasında meme (GSE70947), karaciğer (GSE14520_U133A), kolorektal (GSE44076), böbrek (GSE53757), prostat (GSE6919_U95Av2), akciğer (GSE19804) ve lösemi (GSE63270) kanserlerine ait gen ekspresyon profilleri yer almaktadır. Bu veri setleri genel olarak 101 ila 357 arasında değişen örnek (sample) sayısına ve 12.626 ile 54.676 arasında değişen gen (öznitelik) sayısına sahiptir. Her bir örnek, ilgili kanser türüne ait hastalık ya da sağlıklı birey bilgisi ile etiketlenmiştir. İkili sınıflandırma veri setlerine ek olarak, daha karmaşık analizler için dört adet çoklu sınıflandırma veri seti de kullanılmıştır. Bu setler, birden fazla hastalık alt tipini veya farklı prognostik grupları ayırt etme imkanı sunmaktadır. Çoklu sınıflandırma için kullanılan veri setleri ve GEO kimlikleri ise sırasıyla kolorektal kanser (GSE21510) (3 sınıf), beyin kanseri (GSE50161) (5 sınıf), meme kanseri (GSE45827) (6 sınıf) ve lösemi (GSE28497) (7 sınıf) şeklindedir. Bu zengin ve çeşitli veri seti seçimi, kanser biyolojisi alanındaki farklı sınıflandırma zorluklarını ele alarak, kapsamlı ve güvenilir analizler yapılmasına olanak tanımıştır [33].

Çizelge 3.2, çalışmada kullanılan veri setlerinin temel özelliklerini sunmaktadır. Bu çizelgede veri setinin ait olduğu kanser türü, GEO kimliği, örnek sayısı, gen (özellik) sayısı ve sınıf sayısı bilgileri yer almaktadır.

Çizelge 3.2. Gen Ekspresyonu Veri Setlerinin Özellikleri

Kanser Türü	GEO kimliği	Örnek sayısı	Gen sayısı	Sınıf sayısı
Meme (göğüs)	*GSE70947*	289	35,982	2
Karaciğer	*GSE14520_U133A*	357	22,278	2
Kolorektal	*GSE44076*	194	49,387	2
Böbrek	*GSE53757*	143	54,676	2
Prostat	*GSE6919_U95Av2*	124	12,626	2
Akciğer	*GSE19804*	114	54,676	2
Lösemi	*GSE63270*	101	54,676	2
Kolorektal	*GSE21510*	147	54,676	3
Beyin	*GSE50161*	130	54,676	5
Meme	*GSE45827*	151	54,676	6
Lösemi	*GSE28497*	281	22,284	7

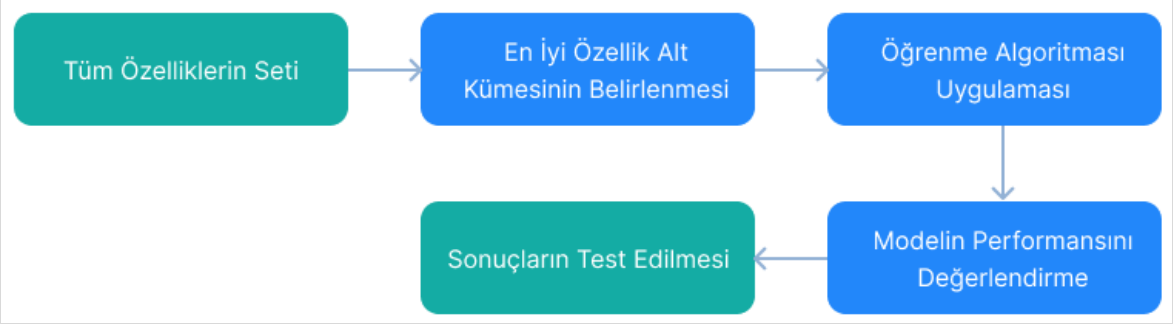
3.2. Öznitelik Seçim Yöntemleri

Son yıllarda yüksek boyutlu veri kümlerine duyulan ilgi ve artan kullanım, sınıflandırma, regresyon, kümeleme gibi birçok alanda öznitelik seçimine olan ilgiyi artırmış ve mevcut öğrenme algoritmalarının performansını ciddi bir şekilde zorlaştırmıştır.

Bu alandaki araştırmalar 1960'lı yıllara kadar uzanmakta olup, son yıllarda daha fazla ilgi görmektedir [34]. Bugün, özellikle mikrodizi teknolojisi gibi büyük veri kümelerinin kullanılabildiği alanlarda, çok sayıda gen açıklama verisinin aynı anda analiz edilebilmesi büyük bir avantaj sağlamaktadır. Bu teknoloji, birçok tıbbi alanda araştırmacılara binlerce gen verisini inceleme imkanı sunsa da, büyük boyutlu gen ekspresyon verileriyle yapılan analizlerde, çok sayıda gen ile temsil edilen az sayıda örneğin varlığı, sınıflandırma işlemlerinde bazı dezavantajlara yol açabilmektedir [35]. Bu tür dezavantajlar öznitelik seçim teknikleri ile ortadan kaldırılarak, daha doğru sınıflandırma modelleri geliştirilmesine olanak tanımıştır. Özellik seçiminde amaç, sınıflandırma performansını artırmaya yardımcı olmayan gereksiz ve alakasız özelliklerin veriden çıkarılmasıdır. Bununla uygun öznitelikler seçilerek model karmaşıklığı azaltılarak, aşırı öğrenme riski ortadan kaldırılabilir ve böylece genelleme performansı iyileştirilebilir [36]. Öznitelik seçim işlemleri üç farklı yaklaşımla sunulabilir. Genel olarak bu yaklaşımlar; filtre, sarmal ve gömülü yöntemler olarak temsil edilir [37]. Ancak, çalışmada yalnızca filtre ve sarmal yöntemlere odaklanılmış, gömülü yöntemler değerlendirme kapsamı dışında bırakılmıştır.

3.2.1. Filtre yöntemler

Filtre yöntemler, en uygun öznitelik alt uzayını belirlemede yaygın olarak kullanılan yöntemlerdir. Özellikle yüksek boyutlu veri setlerinde hızlı hesaplama açısından verimli olan bu yöntem, makine öğrenmesi modellerinden bağımsız çalışarak özniteliklerin hedef değişkenle ilişkisini istatistiksel ölçütlere dayanarak değerlendirir. Değerlendirmeyi öznitelikler arasındaki ilişkiyi göz önünde bulundurmadan yaparak, yöntemin zaman karmaşıklığını önemli ölçüde azaltır [38]. Bu işlem ile en küçükten en büyüğe bir değer elde edilir ve bu değerler, en iyi ayırt edici öznitelikleri belirlemek için kullanılır. En yüksek değere sahip olan öznitelikler, sınıflandırma başarısını artırır ve bu özniteliklerin daha yüksek ayrım yeteneğine sahip olduğunu gösterir. Seçilmek istenen öznitelik sayısına ulaşana kadar en iyi başarı oranı elde edilir ve bu noktada en anlamlı öznitelik alt kümesi belirlenmiş olur [39]. Şekil 3.2'de, filtre yöntemlerinin çalışma şekli akış diyagramı şeklinde daha detaylı verilmiştir.



Şekil 3.2. Filtre yöntemlerinin çalışma prensibi

Şekil 3.2'deki akış şemasına göre ilk adımda, tüm özellikler veri kümesinden alınır ve hiçbir özellik ayrım yapılmadan tüm özellikler değerlendirmeye alınır. Bu aşamada, verideki her bir özellik, modelin başarı açısından potansiyel taşıyan bir öznitelik olabilir. Burada en önemli nokta, başlangıçta verilerin tüm boyutlarının kullanılabilir olmasıdır, çünkü hangi özelliklerin sınıflandırma için etkili olduğunu bu aşamada bilmek mümkün değildir. İkinci adımda ise, özelliklerin her biri, hedef değişkenle ilişkisi açısından istatistiksel ölçütlere göre değerlendirilir. Bu aşamada, her bir özelliğin hedef değişkenle ne kadar güçlü bir ilişkisi olduğunu ölçmek için bazı filtre yöntemler kullanılır. Ve bununla, hedef değişkenle en güçlü ilişkisi olan özellikler seçilir. Sonraki adımda, seçilen özellik alt kümesi kullanılarak öğrenme algoritması ile veri sınıflandırma veya regresyon görevleri için modellenir. Öğrenme algoritması uygulandıktan sonra, model performansı

performans metrikleri ile değerlendirilir. Bu adımda, seçilen en iyi özellikler ve öğrenme algoritmalarıyla, tüm eğitim verisi kullanılarak model eğitilir ve bu özelliklerin model başarısına katkısı ölçülerek son model eğitilir. Son modelin eğitilmesi, filtre yöntemlerinin başarısını belirleyen kritik bir adımdır. Eğer bu süreçte en iyi sonuçları veren model elde edilirse, bu modelin doğruluğu daha önce belirlenen test verisi ile değerlendirilebilir. Son adımda ise eğitimden sonra oluşturulan model test edilir. Bu aşamada modelin doğruluğu, başarısı ve genellenebilirliği kontrol edilir. Eğer model, test verisi üzerinde yüksek başarı gösteriyorsa, özellik seçimi ve sınıflandırma işlemi başarılı kabul edilir.

Bilgi kazancı

IG, bir veri setindeki özniteliklerin hangi kombinasyonlarının daha yüksek verim sağlayacağını belirlemek için kullanılan önemli bir öznitelik seçme yöntemidir. Sınıf etiketleriyle öznitelikler arasındaki ilişkiyi ölçmek için entropi kavramı kullanılır. Entropi, bir veri kümesindeki düzensizlik düzeyini ölçer [40]. Eğer bir öznitelik tüm örneklerde aynı sınıfa ait değerler içeriyorsa, bu öznitelik saf (homojen) kabul edilir ve entropi değeri sıfır olur. Eğer farklı sınıflara ait değerlere sahipse, entropi değeri yükselir [41]. Entropi hesaplamasında, her bir sınıfa ait olasılık değerlerinin logaritmik olarak ağırlıklı toplamı kullanılır (Eş. 3.1). Burada, *m*, analiz edilen sınıfların sayısını, *X*, sınıf değerini, *p(x)* ise *X* değişkeninin olasılık yoğunluğunu belirtir [40].

$$E(X) = \sum_{i=1}^{m} -p(x_i)\log_2 p(x_i) \tag{3.1}$$

Bilgi kazancının amacı, daha iyi sınıflandırma modelleri üretmek için veri düzensizlik derecesini azaltmaktır. Yüksek entropili öznitelikler üzerinde yapılacak bir bölme işlemiyle, bu düzensizlik azalır ve bilgi kazanımı elde edilir. Bilgi kazanımı, tüm veri kümesinin entropisi ile, öznitelik A'ya göre bölündükten sonraki ağırlıklı ortalama entropisi arasındaki fark olarak tanımlanır (Eş. 3.2) [40]:

$$IG = E(X) - E_A(X) \tag{3.2}$$

Sonuç olarak, bilgi kazanımı, gereksiz bilgileri kaldırmak ve büyük veri kümelerindeki özelliklerin sınıflandırma gücünü değerlendirmek için yaygın olarak yararlı bir teknik olarak kabul edilir. Özellikle biyolojik verileri ve kanser gibi karmaşık hastalıkları sınıflandırmak için çok önemlidir [42].

Ki-kare istatistiği

Ki-kare testi, iki kategorik değişkenin birbirine bağımlı olup olmadığını belirlemek için kullanılır. Özniteliklerin bağımsız olup olmadığını değerlendirerek, öznitelikler arasında istatistiksel olarak anlamlı bir ilişki olup olmadığını tespit eder [43]. Bu test, her bir öznitelik için ayrı ayrı olarak hesaplanır ve test istatistiği belirli bir eşik değeri üzerinde sıralanarak en anlamlı öznitelikler seçilir. Test ile yüksek skorlara sahip olan öznitelikler, modelin doğruluğunu artırmaya en çok katkı sağlayacak öznitelikler olarak seçilir [10]. Eğer öznitelikler sayısal değerler içeriyorsa, bu değerler önce kategorilere ayrılır ve ardından ki-kare testi uygulanır. Bu durumda, özniteliklerin her birinin, ilgili sınıfla olan ilişkisinin gücü, hesaplanan ki-kare değeriyle belirlenir. Öznitelikler, daha yüksek ki-kare değeri elde eden özellikler olarak sıralanır ve seçim yapılır [44].
Ki-kare testinin uygulanabilmesi için verinin ayrık hale getirilmesi gerekmektedir (Eş. 3.3).

$$E_{ij} = \frac{R_i C_j}{N} \tag{3.3}$$

Burada, R_i sırasındaki örnek sayısını, C_j ise sütundaki sınıf sayısını temsil eder. N ise toplam örnek sayısını ifade eder. Her öznitelik için hesaplanan istatistiksel değer, o özniteliğin sınıflar arasında anlamlı bir ayrım yapma yeteneğini yansıtır. Aşağıda verilen ifade (Eş. 3.4), bu hesaplama sürecini tanımlar:

$$X^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\frac{O_{ij} - E_{ij}}{E_{ij}}\right) \tag{3.4}$$

Özellik seçiminin bu şekilde yapılması hem istatistiksel anlamlılık hem de hesaplama açısından verimli sonuçlar elde edilmesine olanak tanır. Bu test özniteliklerin sınıflara göre ayrılmasını sağlayarak, iyi sınıflandırma performansına ulaşmada etkili bir yöntem sunar.

ReliefF algoritması

ReliefF yöntemi, özniteliklerin ağırlıklarını hesaplamak için kullanılan filtre tabanlı bir öznitelik seçme algoritmasıdır ve sınıflar arasındaki ayrım gücünü değerlendirmek amacıyla her bir özniteliği ayrı ayrı inceleyerek hangi özelliklerin daha belirleyici olduğunu ölçer. Bu algoritma, sınıflar arasında ayrım yapma yeteneğini değerlendirmek için her bir özelliği ayrı ayrı analiz ederek ilgili özelliklerin önemini değerlendirir. ReliefF, aynı sınıftan en yakın komşular (near-hit) ile farklı sınıflardan komşuları (near-miss) karşılaştırarak her bir özelliğin önemini belirler. Bu yöntem, özellikle iki sınıf arasında ayrım yapmayı hedefleyen veri kümeleri için uygundur [45].

Bu yöntemin temel mantığı, her bir özniteliği, aynı sınıf içindeki örnekleri ne kadar iyi kümelerken, diğer sınıflardan olan örnekleri ne kadar iyi ayırt edebildiğine göre değerlendirmektir. Bu mantıkla, özelliklere bir ilişki ağırlığı atanır ve bu ağırlıklar her bir örneğin en yakın sınıf içi ve sınıflar arası farklarına göre güncellenir. Bu işlemde, farklı sınıfların örneklerinden ayırt edebilen özelliklere daha fazla ağırlık verilir. Ayrıca, ağırlıklar, her sınıfın olasılıkları da göz önünde bulundurularak güncellenir [10].

Fisher korelasyon skorlama

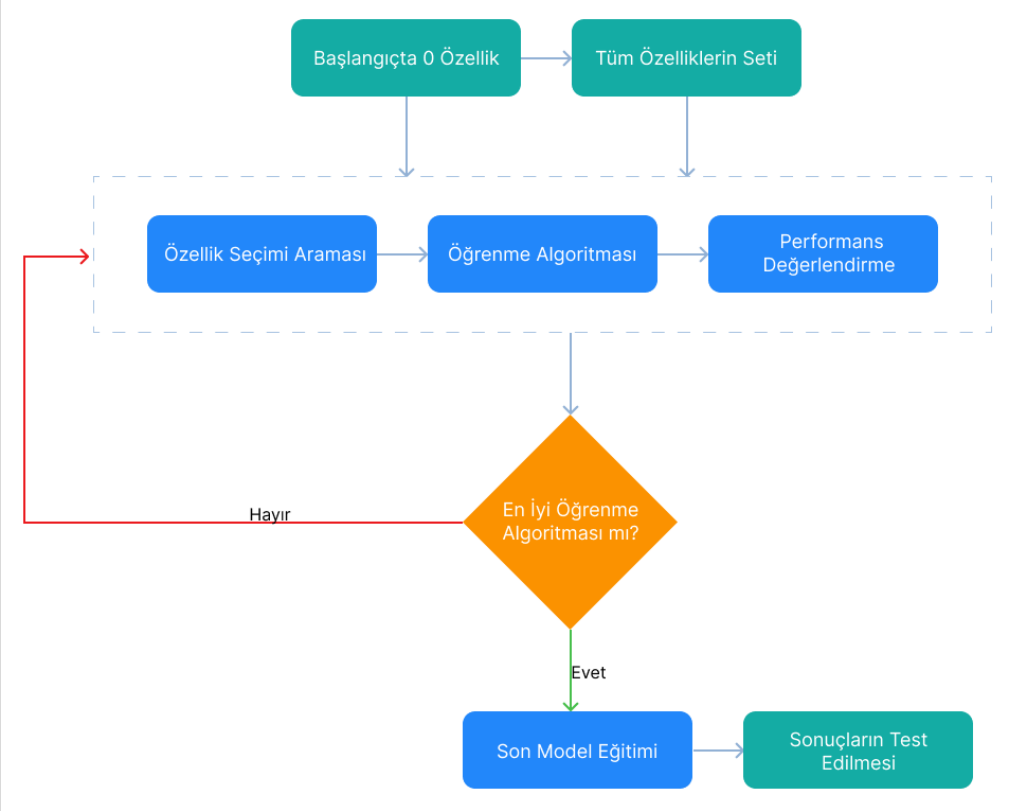
Fisher Skoru (FKs), veri kümesindeki öznitelikler arasındaki ayrım gücünü maksimize etmeye çalışan bir yöntemdir. Özniteliklerin sayısını azaltarak, sınıflandırma doğruluğunu artırır. Yüksek bilgi kazancı sağlayan öznitelikler seçildiğinde, modelin doğruluğu artar ve daha hızlı eğitim süreci sağlanır. Gürültülü ve gereksiz veriler filtrelendiğinden bu yöntemde daha doğru sonuçlar elde edilmektedir. Aynı zamanda sınıflar arasındaki farkları arttırarak, aynı sınıf içindeki benzerlikleri de azaltmayı hedefler. FKs yüksek öznitelikleri seçtiğinden, birlikte anlam taşıyan ve düşük skorlarla değerlendirilen öznitelikleri göz ardı edebilir. Bu durum ise onun dezavantajı olarak öne çıkmaktadır [38]. FKs'u elde etmek için her bir özniteliğin ortalama değeri ile standart sapma değeri kullanılarak bir skor hesaplanır (Eş. 3.5).

$$FKS(x_i) = \frac{|u_i^+ - u_i^-|}{\sigma_i^+ + \sigma_i^-} \tag{3.5}$$

Burada u_i^+ ve u_i^- sınıfların ortalamalarını, σ_i^+ ve σ_i^- ise her sınıfın standart sapmalarını temsil eder.

3.2.2. Sarmal yöntemler

Sarmal yöntemler, öznitelik seçiminde sürekli bir arama işlemi gerçekleştiren ve her adımda bir sınıflandırıcı algoritmasına dayalı olan yöntemlerdir. Bu yöntemlerin temel avantajı, çözüm uzayını filtre yöntemlere kıyasla daha etkili bir şekilde keşfetmeleridir. Ancak bu yöntemlerin her adımda bir sınıflandırıcıya ihtiyaç duyması ve bu nedenle işlem süresinin artması bu yöntemlerin dezavantajıdır. Sarmal yöntemlerde her adımda sınıflandırıcı algoritmalarının çalıştırılması, yöntemin daha yavaş işlem süresine sahip olmasına neden olur [15]. Bu yöntemler, çözüm uzayını adım adım oluşturur ve her adımda mevcut çözüm uzayı değişebilir. Bu, en iyi öznitelik alt kümesinin belirlenmesi işleminin zaman alıcı olmasına yol açar. Bununla birlikte, çözüm uzayı elde edene kadar, sınıflandırıcıya dayalı olarak en iyi ayrım gücüne sahip öznitelikler kesin bir şekilde seçilemeyebilir. Bu yöntemlerin, genel çalışma mekanizması Şekil 3.3'te verilmiştir.



Şekil 3.3. Sarmal yöntemlerinin çalışma prensibi

Bu yöntem, özellik seçiminde ileri seçim (forward selection) ve geri eleme (backward elimination) işlemlerini birleştirir ve her iki tekniği de kullanarak en iyi özellik alt kümesini bulmaya çalışır. İlk adımda hiçbir özellik seçilmez. Eğer ileri seçim yöntemi

kullanılacaksa, sıfır özellik ile başlanır. Bu adımda, modelde kullanılacak özellikler seçilmeden önce hiçbir özellik verilmez. Eğer geri eleme yöntemi kullanılacaksa, başlangıçta tüm özellikler kullanılır. Sonraki aşamada özniteliklerin alt kümesi belirlenir. İleri seçimde, her defasında yeni bir özellik eklenirken, geri elemede ise mevcut bir özellik çıkarılır. Bu adım, sürekli olarak özellikler eklenip çıkarılarak en iyi performansı sağlayacak özellik kümesinin bulunmasını amaçlar. Seçilmiş en iyi özellik alt kümesiyle bir öğrenme algoritması uygulanır ve performans, performans metrikleri ile değerlendirilir. Bu noktada, elde edilen performansın "en iyi" olup olmadığı kontrol edilir. Eğer "evet" ise, en iyi özellik alt kümesi belirlenir ve işlem sonlandırılır. Eğer "hayır" ise, ileri seçim veya geri eleme yöntemi tekrar uygulanarak daha iyi sonuçlar alınmaya çalışılır. En iyi özellikler ve öğrenme algoritması belirlendikten sonra, son model eğitilir. Bu adımda, eğitim verisi ile model son bir kez eğitilerek en iyi performans elde edilir. Son adımda, eğitilen model test edilir ve doğruluğu değerlendirilir. Modelin test verisi üzerinde gösterdiği performans, seçilen özelliklerin ve algoritmaların başarısını belirler.

Genetik algoritma

Genetik Algoritma (GA) optimizasyon ve arama problemlerini çözmek için kullanılan bir yöntem olup, 1970'li yıllarda John Holland tarafından geliştirilmiştir [46]. Algoritma, ilk başta bir popülasyon oluşturarak, popülasyonun her bireyini, çözüm uzayındaki olası çözümleri temsil eden kromozomlarla ifade eder. Her bir kromozom, bir çözüm önerisini temsil eder ve bu çözüm önerisi, genetik operasyonlar aracılığıyla geliştirilir. Genetik algoritma, genellikle dört ana adımda çalışır; seçim, çaprazlama, mutasyon ve sonlandırma. Bu adımlar sayesinde, daha iyi çözümler elde edilir [38].

GA'nın bir diğer önemli özelliği olan sonlandırma kriteri ise algoritmanın ne zaman durması gerektiğini belirler. Çoğu durumda, GA belirli bir iterasyon sayısı tamamlandığında ya da belirli bir başarıya ulaştığında durdurulur. Bu yaklaşım, hesaplama kaynaklarının daha verimli kullanılmasını sağlayarak, algoritmanın uzun süre çalışmasını önler. GA'nın önemli adımlarından biri olan genetik kodlama, problemi çözmek için doğru şekilde yapılandırılmalıdır [47]. Algoritma genellikle ikili kodlama kullanır, ancak prombel türüne bağlı olarak onluk sistemde de kodlama yapılabilir. Bu, algoritmanın esnekliğini artırır ve farklı problemler için uygun çözüm stratejileri oluşturulmasına olanak tanır. Algoritma, geniş arama alanlarında çözüm arayışı sürdürdüğü için, çözüm bulma süreci genellikle daha esnek gerçekleşir. Bu ise algoritmanın farklı veri türleriyle

çalışabilmesini ve daha geniş bir problem alanında uygulanabilir olmasını sağlar. Ancak, algoritmanın her adımında yapılan hesaplamalar, zaman alıcı olabildiğinden genetik algoritmaların etkili bir şekilde çalışabilmesi için parametrelerin dikkatli bir şekilde ayarlanması ve her probleme özgü genetik operatörlerin seçilmesi gereklidir.

Sonuç olarak, genetik algoritmalar, karmaşık ve yüksek boyutlu problemlerin çözülmesinde oldukça etkili araçlar sunmaktadır. Ancak, algoritmanın tam potansiyeline ulaşabilmesi için doğru ayarlamalar ve uygun parametre seçimleri gereklidir.

3.2.3. Topluluk yöntemleri

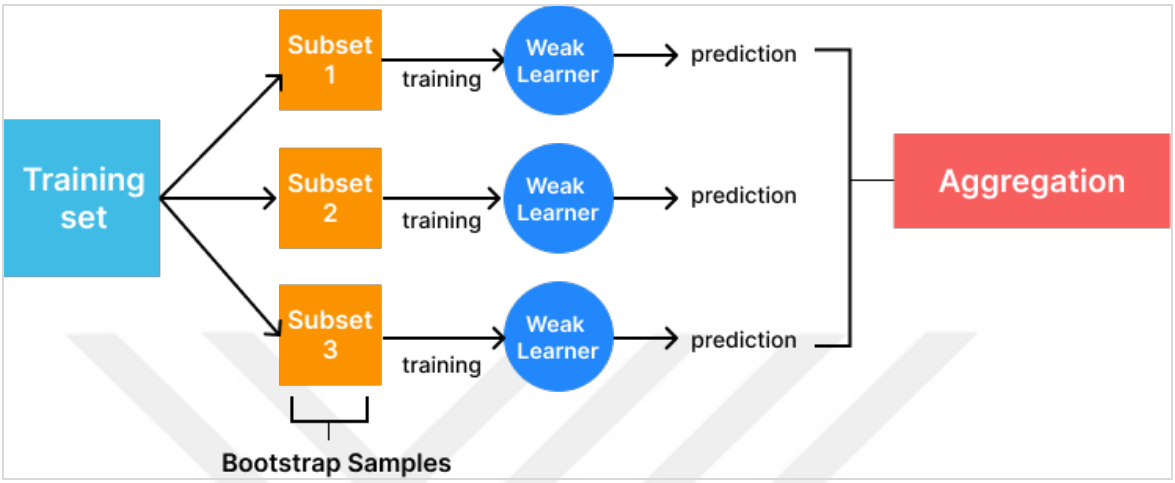
Topluluk yöntemleri, birden fazla temel modelin birleşimiyle oluşan çok güçlü bir öğrenme yöntemidir. Bu yöntemler, temel modellerin nasıl eğitildiği ve birleştirildiği konusunda çeşitlilik göstermektedir [48]. Topluluk yöntemleri genellikle, temel sınıflandırıcıları birleştirerek tek bir tahmin oluşturmak için birden fazla öğrenci kullanır. Bu temel sınıflandırıcılar, belirli veri kümeleri üzerinde eğitildikten sonra, bunların tahminleri birleştirilerek nihai sonuç elde edilir [49]. Bu yöntemler aynı zamanda regresyon ve diğer makine öğrenimi disiplinlerinde kullanılmaktadır. En yaygın kullanılan ensemble teknikleri arasında bagging, stacking ve boosting yer almaktadır.

Bagging yöntemi

Bagging (Bootstrap Aggregating), birden fazla modelin birleşiminden oluşan modelin çeşitlendirilmesini sağlayarak daha doğru tahminler yapan bir topluluk öğrenimi yöntemidir. Yöntemin amacı, eğitim setindeki küçük değişikliklerin model tahminlerinde önemli farklar yaratmasına olanak tanıyarak, modelin genelleme yeteneğini artırmak ve aşırı öğrenme riskini azaltmaktır [50]. Ancak bagging'in bir dezavantajı, yüksek hesaplama maliyeti ve yüksek bias değerine sahip olmasıdır [51]. Bagging, genellikle regresyon problemlerinde tahminlerin ortalamasını alırken, sınıflandırma problemlerinde ise çoğunluk oyu kullanır. Bu, her modelin sınıflandırma yaptığı durumu ve en fazla oy alan sınıfı belirler [52]. Bagging yönteminde tahmin, tüm alt modellerin çıktılarının ortalaması alınarak elde edilir (Eş. 3.6).

$$\hat{f}_{bag}(X) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(X) \tag{3.6}$$

Burada, B bagging ile oluşturulan model sayısını, b b-inci modeli, $\hat{f}_b(X)$ b-inci modelin tahmini ifade eder. Bagging yöntemlerinin genel çalışma prensibi, Şekil 3.4'teki grafikte verilmiştir.



Şekil 3.4. Bagging yöntemlerin çalışma prensibi

Şekil 3.4'deki akış şemasına göre, ilk adımda eğitim verisi alınarak, rastgele örnekleme yöntemiyle farklı alt kümler oluşturulur. Bu alt kümeler üzerinde eğitim yapılarak zayıf öğrenciler oluşturulur. Son aşamada ise her bir zayıf öğrencinin yaptığı tahminler çoğunluk oylaması gibi birleştirme yöntemleri ile birleştirilir. Bagging yöntemi, zayıf öğrenicilerin bir araya getirilmesiyle güçlü bir tahmin modelinin oluşturulmasını sağlar ve bu süreç, modelin doğruluğunu artırarak aşırı uyum riskini azaltır.

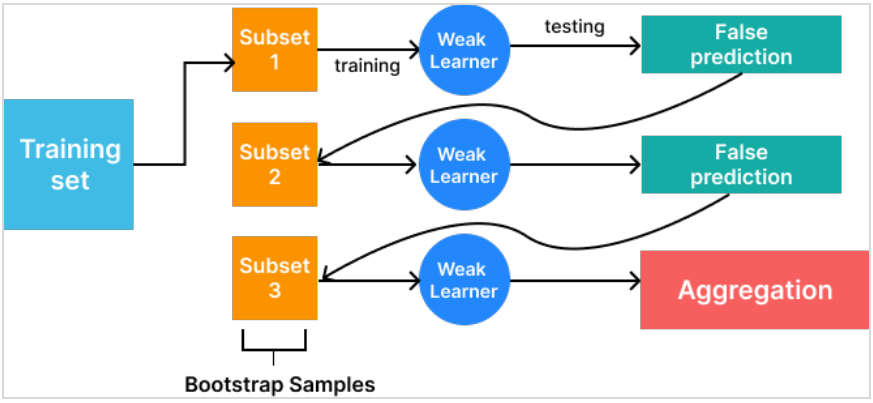
Boosting yöntemi

Boosting yöntemi, bir dizi zayıf öğrenicinin ardışık olarak eğitildiği bir topluluk öğrenme yöntemidir. Bu yöntem, her bir modelin, bir önceki modelin hatalarını düzeltmeye yönelik olarak eğitildiği bir süreçten geçer. Her bir ardışık model, önceki modelin zayıf olduğu örneklere daha fazla odaklanarak, eğitim sürecinde hata düzeltmesi yapar [53]. Bu yaklaşım, daha güçlü ve doğru bir model oluşturmak amacıyla her bir öğreti modelinin hata oranını minimize etmeye çalışır. Aynı zamanda her bir modelin eğitim verisinde göz ardı edilen, hatalı sınıflandırılmış örneklere odaklanarak, modelin performansını önemli ölçüde iyileştirir. Ancak boosting'in bir dezavantajı, her bir sınıflandırıcının bir öncekinin hatalarını düzeltmeye çalışması nedeniyle modelin eğitim süresinin uzaması ve aşırı uyum riskinin artmasıdır. Bu nedenle, boosting algoritmalarının doğru bir şekilde çalışabilmesi için hiperparametre ayarlarının yapılması gerekmektedir [51].

Boosting yönteminde, zayıf öğrenicilerin tahminleri belirli ağırlıklarla çarpılarak toplanır ve böylece nihai tahmin elde edilir (Eş. 3.7).

$$\hat{f}(x) = \sum_{t=1}^{B} \lambda_t \hat{f}_t(x) \tag{3.7}$$

Burada, $\hat{f}(x)$ final modelin tahmini, B toplam ağaç sayısı, $\hat{f}_t(x)$ t-inci zayıf öğrenicinin tahmini, λ_t her bir modelin ağırlığını gösteren öğrenme oranıdır. Bu formülde, B sayısı kadar model ardışık olarak eğitildikten sonra, her bir modelin tahminlerinin ağırlıklı toplamı alınarak nihai tahmin yapılır [54]. Boosting yöntemlerinin genel çalışma mekanizması Şekil 3.5'te sunulmuştur.



Şekil 3.5. Boosting yöntemlerin çalışma prensibi

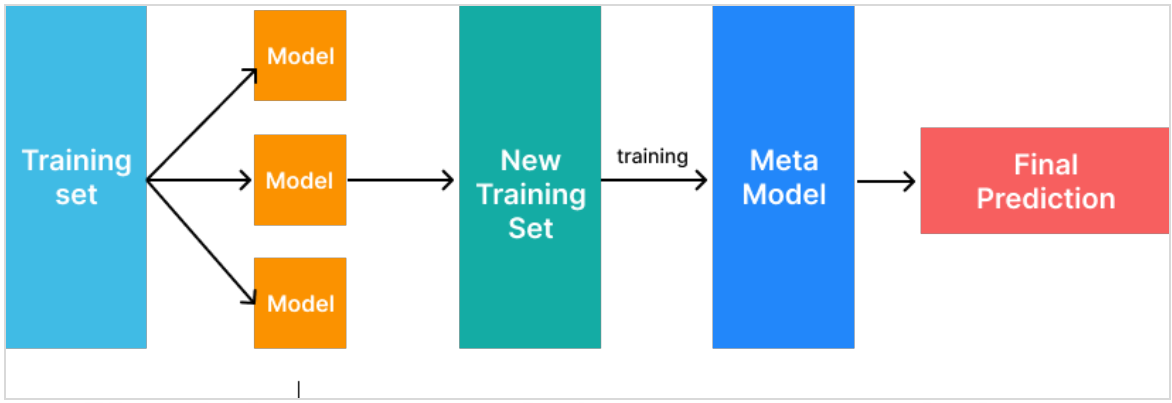
Şekil 3.5'deki akış şemasına göre, ilk adımda her alt küme üzerinde eğitim yapılır. Bu algoritmada önemli bir fark, her yeni modelin önceki modelin yaptığı hataları düzeltmeye odaklanmasıdır. Yani, bir zayıf öğrenici eğitildikten sonra, yanlış tahmin edilen örnekler üzerine daha fazla ağırlık verilerek bir sonraki öğrenici eğitilir. Bu işlem tekrarlanarak daha önce yanlış sınıflandırılan örneklere daha fazla dikkat edilir. Sonda tüm bu öğrenciler birleştirilerek güçlü ve doğru bir model ortaya çıkar. Boosting yöntemi, özellikle yanlış tahminlerin üzerine gitmek için güçlü bir yöntemdir.

Stacking yöntemi

Stacking yöntemi, Wolpert (1992) tarafından önerilen ve farklı modellerin bir araya getirilerek daha başarılı tahminler elde edilmesini amaçlayan bir topluluk öğrenme yöntemidir [55]. Bu yöntem, çok katmanlı bir yapıya dayalıdır ve genellikle iki seviyeden oluşur; birinci seviye (level-0) temel modelleri ve ikinci seviye (level-1) meta-model. İlk seviyedeki modeller, eğitim verisi üzerinde eğitilmiş ve tahminler üreten bireysel öğrenicilerdir. İkinci seviyede ise meta-model, bu tahminleri alarak en uygun şekilde birleştirmeyi öğrenir ve nihai sonucu üretir. Burada, temel modellerin çıkışları genellikle sınıflandırma problemlerinde sınıf etiketleri veya olasılık değerleri olarak kullanılır [56, 57].

Stacking yönteminin bir dezavantajı, temel modellerin sayısını ve türünü doğru belirlemenin zor olmasıdır. Hangi modellerin daha iyi tahminler üreteceğini öngörmek her zaman kolay değildir ve bu da model performansını doğrudan etkileyebilir. Ayrıca, meta modelin anlaşılabilirliğini sağlamak ve işlem süresini makul seviyede tutmak, özellikle

veri seti büyüdükçe daha karmaşık hale gelir [58]. Şekil 3.6'da Stacking yöntemlerinin çalışma prensibini sunan genel işleyiş verilmiştir.



Şekil 3.6. Stacking yöntemlerin çalışma prensibi

Şekil 3.6'daki akış şemasına göre yöntem iki ana seviyeden oluşur. İlk aşamada, eğitim verisi kullanılarak birden fazla temel model eğitilir ve bu modellerin tahminleri meta-modelin (Level-1) girdisi olarak kullanılır. Meta-model, temel modellerden nihai tahmini üretir. Her temel model, farklı veri örnekleri üzerinde eğitilerek meta-modelin daha doğru ve güvenilir sonuçlar üretmesini sağlar. Bu yaklaşım, birden fazla model türünün tahminlerini birleştiren topluluk öğrenme mantığına dayanır.

3.3. Makine Öğrenmesinde Modelleme ve Değerlendirme Süreçleri

Bu bölümde, çalışmada kullanılan sınıflandırma algoritmaları, performans değerlendirme ölçütleri ve istatistiksel testlere genel hatlarıyla yer verilmektedir. Sınıflandırma, seçilen özniteliklerin ayırt ediciliğini test etmek amacıyla uygulanmaktadır. Değerlendirme ölçütleri, modellerin başarımını çok yönlü olarak incelemek için kullanılmakta; istatistiksel testler ise elde edilen sonuçların anlamlılığını ortaya koymayı hedeflemektedir.

3.3.1. Sınıflandırma algoritmaları

Sınıflandırma algoritmaları, biyomedikal veri analizi gibi yüksek boyutlu problemlerde örneklerin ait oldukları sınıfların doğru bir şekilde belirlenmesini sağlamak amacıyla yaygın olarak kullanılmaktadır. Bu çalışmada, seçilen özniteliklerin sınıflandırma performansına etkisini değerlendirmek için çeşitli denetimli öğrenme algoritmalarından yararlanılmıştır. Farklı yapıya sahip bu algoritmaların kullanılması, elde edilen sonuçların güvenilirliğini ve genellenebilirliğini artırmak açısından önem taşımaktadır.

K-Nearest Neighbors

kNN algoritması hem sınıflandırma hem regresyon problemlerinde yaygın kullanılan basit ama etkili bir makine öğrenme algoritmasıdır. Temel çalışma prensibi, sınıflandırılacak veri noktasının etrafındaki en yakın komşuların oy çokluğuna dayanır. kNN algoritmasının temel çalışma prensibi, sınıflandırılacak veri noktasını çevreleyen en yakın komşuların çoğunluk oyuna dayanır. kNN algoritması, verilerin dağıtım yapısı hakkında herhangi bir ön varsayımda bulunmaz, bu da algoritmanın esnekliğini artırır ve çok çeşitli veri türlerine uygulanmasına olanak tanır.

Aynı zamanda, gürültülü verilere karşı dayanıklı olan kNN algoritması veri kalitesi yüksek olduğu sürece oldukça başarılı sonuçlar verir. Bu algoritmanın bir dezavantajı, tahmin yaparken eğitim setindeki tüm veri noktalarıyla hesaplama yapması nedeniyle hesaplama maliyeti ve hafıza gereksinimi duymasıdır. Hem hesaplama süresi hem bellek kullanımını önemli ölçüde arttırdığından yüksek boyutlu ve büyük veri setlerinde sınıflandırma sürecini yavaşlatabilir [59]. Sonuç olarak, KNN algoritması özellikle veri yapısının iyi anlaşıldığı, yüksek kaliteli ve nispeten küçük ya da orta büyüklükteki veri setlerinde oldukça etkili bir yöntemdir. Büyük ve karmaşık veri setlerinde ise algoritmanın performansını artırmak için indeksleme teknikleri, özellik seçimi ve boyut indirgeme yöntemleriyle desteklenmesi gerekmektedir.

Decision Tree

DT, özellikle karmaşık ve yüksek boyutlu veri kümelerinde, sınıflandırma ve regresyon görevlerine sıklıkla uygulanan popüler bir makine öğrenme tekniğidir. Algoritma, hedef değişkeni tahmin etme nihai hedefi ile belirli özelliklere göre veri kümesini karar noktalarında (veya düğümlerde) yinelemeli olarak bölerek çalışır. Her düğümde, algoritma belirli bir özelliğe göre bir karar verir ve bu daha sonra verilerin nasıl dallanacağını belirler. Ağacın yaprak düğümleri, nihai sınıflandırma sonuçlarını veya tahmin edilen değerleri temsil eder [59].

Basitlikleri ve yorumlama kolaylıkları bu algoritmanın güçlü yanlarından biridir. Aynı zamanda, karar ağaçları hızlı işleme ve düşük bellek gereksinimleri sayesinde önemli performans sorunları olmadan büyük veri kümelerini işlemelerine olanak tanır [60]. Bu algoritmanın dezavantajı ise aşırı öğrenmeye oldukça yatkın olmasıdır. Yani, eğitim verisine fazla uyum sağlayarak yeni verilere genelleme yapmada başarısız olabilirler. Böylece büyük ve karmaşık karar ağaç yapıları oluştuğunda, bu modellerin yorumlanması aşırı zorlaşarak, sonuçları ciddi bir şekilde etkileyebilir [60].

Logistic Regression

Lojistik regresyon, sınıflandırma problemlerine yaygın kullanılan güçlü bir yöntemdir. Özellikle ikili sınıflandırma problemlerinde tercih edilen bu yöntem, 0 ile 1 arasında olasılık değerleri üreterek bir olayın gerçekleşme ya da gerçekleşmeme durumunu tahmin eder. LR, bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkiyi en iyi şekilde modellemeyi amaçlar. Sigmoid (lojistik) fonksiyon kullanılarak, her veri noktası için belirli bir sınıfa ait olma olasılığı hesaplanır ve bu olasılığa bağlı olarak sınıflandırma yapılır. Bu sayede, model sonuçları olasılık biçiminde yorumlanabilir ve belirlenen eşik değeri ile sınıflar ayrılır [60].

LR'nin başlıca avantajlarından biri uygulama kolaylığı, hesaplama verimliliği ve modelin düzenlenmesinin kolay olmasıdır. Girdi özelliklerinin ölçeklenmesine gerek olmaması ve küçük veri gürültüsüne karşı dayanıklı olması da önemli avantajlardandır. Ayrıca, modelin çıktısı bir olasılık değeri olduğundan, model performansı özelleştirilmiş metriklerle değerlendirilebilir. Ancak, lojistik regresyonun bazı sınırlamaları vardır. Modelin doğrusal bir karar sınırı varsayımı, karmaşık ve doğrusal olmayan ilişkileri etkili bir şekilde yakalamasını engeller. Aynı zamanda tüm bağımsız değişkenlerin modele dahil edilmesi gerekir ve değişkenler arasında çoklu bağlantı olması modelin performansını olumsuz etkileyebilir [60]. Lojistik regresyon, sağlık sektöründe hastalık risk tahmininden, doğal dil işleme ve kelime sınıflandırmaya; finansal modellemeden hava durumu tahminlerine kadar geniş bir uygulama alanına sahiptir. Ayrıca, kanser tanısı, süreç veya sistem arızası olasılığı gibi mühendislik uygulamalarında da yaygın olarak kullanılmaktadır.

Random Forest

Random Forest, bagging yönteminin önemli bir modifikasyonu olarak ortaya çıkmış, doğruluk oranını artırmayı ve ağaçlar arasındaki korelasyonu azaltmayı hedefleyen bir algoritmadır. Bu algoritmada ilk olarak, her bir karar ağacı orjinal veri setinden rastegele örneklem alınarak eğitilir ve bu örneklem, eğitim setinin büyüklüğüyle aynı boyutta olur. Bu şekilde ağaçlar arasında daha fazla çeşitlilik sağlanır ve her ağacın tahminleri daha güvenilir olur. RF'de ikinci rastegleleştirme kaynağı ise öznitelik örneklemidir. Bu yöntemle, her düğümdeki bölme (split) işlemi için, giriş değişkenlerinin bir alt kümesi rastgele seçilir ve en iyi bölme işlemi aranmaya çalışılır ve bu sayede her ağaç farklı özelliklere odaklanarak genel tahmin doğruluğunu artırır [52, 49]. RF hem sınıflandırma hem regresyon problemlerinde etkili bir şekilde kullanılmaktadır. Bu algoritmanın en büyük avantajlarından biri her bir ağacın doğruluğunu değil, aynı zamanda ağaçlar arasındaki çeşitliliği de artırır, böylece modelin aşırı öğrenme riskini azaltır [61]. Bu yöntem sadece performans açısından değil, aynı zamanda açıklanabilirlik açısından da avantaj sunmaktadır. Ağaçlardan çıkarılan kurallar sayesinde kullanıcılar hangi değişkenlerin karar süreçlerine daha fazla katkı sağladığını görebilirler [62]. Böylece, veri setinde önemsiz değişkenlerin elenmesi sağlanabilir.

Özellikle büyük veri setlerinde kullanılabilirliği artırmak adına, RF algoritmaları paralel hesaplama teknikleriyle entegre edilebilir. Bu sayede işlem süreleri azaltılır ve büyük veri uygulamalarında etkin bir çözüm sunulur. Random Forest, sadece tablo bazlı verilerde değil, görüntü işleme, sınıflandırma, segmentasyon ve uzaktan algılama gibi yüksek boyutlu veri içeren alanlarda da yaygın şekilde kullanılmaktadır.

Extra Tree

ETs, karar ağaçlarının geliştirilmiş bir versiyonundan oluçan bir makine öğrenme algoritmasıdır. Güçlü, doğru ve esnek modeller üretmek için birden fazla karar ağacını bir araya getirerek çalışır. Hem doğruluğu hesaplamada hem de büyük ve yüksek boyutlu veri setlerinde hesaplama açısından verimli bir algoritmadır. RF yöntemine benzer şekilde çalışmakla birlikte, ETs rastgele olarak tüm eğitim veri setini kullanır ve düğüm bölme noktalarını tamamen rastgele seçer. Bu rastgelelik, modelin varyansını azaltmaya yardımcı olur ve ayrıca eğitim sürecini hızlandırarak daha stabil hale getirir. Algoritma, her düğümde rastgele seçilen sınırlı sayıda özelliklere göre bölme yapmayı içerir ve tüm

ağaçlar oluşturulduktan sonra, nihai tahmin, sınıflandırma durumunda çoğunluk oyuyla, regresyon durumunda ise ortalama alınarak elde edilir. Bu nedenle, ET'ler doğruluk ve hesaplama açısından verimli bir yaklaşımdır [63].

XGBoost

XGBoost, karar ağaçlarına dayalı, Gradient Boosting prensibini temel alan bir makine öğrenme algoritmasıdır [64]. Bu algoritmanın öne çıkan özelliklerinden biri, sadece karar ağaçlarını temel öğrenici olarak kullanması ve bu ağaçların karmaşıklığını doğrudan kontrol altına almak için özel bir düzenleme terimi kullanmasıdır (Eş. 3.8).

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{3.8}$$

Burada T, ağacın yaprak sayısını, ω yapraklara karşılık gelen tahmin değerlerini ifade eder. Bu formül, ağaçların ne kadar karmaşık olacağını denetlemek amacıyla kullanılır. Yüksek γ değerleri, daha basit ağaçlara yol açar çünkü iç düğüm bölünmeleri için daha yüksek kazanç gerektirir [64]. Bu algoritmanın diğer bir güçlü yanı ise rastgelelik tekniklerini entegre ederek aşırı öğrenmenin önüne geçmesi aynı zamanda eğitim süresini de kısaltmasıdır. XGBoost'un diğer bir dikkat çeken başka yönü ise alan bilgisine dayalı kısıtlamaların uygulanabilmesidir. Ayrıca, özellik etkileşim kısıtlamaları sayesinde, belirli özelliklerin aynı ağaç yolunda birlikte kullanılmaması sağlanarak model açıklanabilirliği artırılır. Model, sadece doğruluk açısından değil, aynı zamanda eğitim süresi ve kaynak verimliliği açısından da avantaj sağlar. Eksik verileri ele alış biçimi, düzenleyici terimleri ve paralel hesaplama desteği sayesinde XGBoost, gradient boosting algoritmalarının daha güçlü ve optimize edilmiş bir versiyonu olarak kabul edilmektedir [65].

AdaBoost

AdaBoost (Adaptive Boosting), 1996 yılında Freund ve Schapire tarafından geliştirilen, temelinde zayıf sınıflandırıcıları bir araya getirerek güçlü bir model oluşturan makine öğrenme algoritmasıdır [66]. Bu algoritma, özellikle ikili sınıflandırma problemleri için geliştirilmiş olup, her iterasyonda yanlış sınıflandırılan örneklere daha fazla odaklanarak yeni bir model üretir. Böylece, her yeni zayıf öğrenici bir öncekinin hatalarını azaltmaya çalışır. Genellikle karar ağaçlarının yalnızca bir seviye derinlikte olan basitleştirilmiş halleri, yani decision stumps, AdaBoost'ta zayıf öğrenici olarak kullanılır. Eğitim süreci,

örneklerin ağırlıklandırılmasıyla başlar. Başlangıçta tüm örneklere eşit ağırlık verilir, ancak her iterasyonun sonunda yanlış sınıflandırılan örneklerin ağırlıkları artırılır. Bu şekilde model, hatalı sınıflandırılan verileri bir sonraki iterasyonda daha fazla dikkate alır [67]. Her bir zayıf öğrenicinin modele katkısı, doğruluk oranına bağlı olarak belirlenir ve bu katkı α_t ile ifade edilir (Eş. 3.9):

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{3.9}$$

Burada ϵ_t, t'inci zayıf öğrenicinin hatalı sınıflandırma oranını temsil eder [67]. Bu katsayı, öğrenicinin son sınıflandırmadaki etkisini belirler. Doğru sınıflandırılan örneklerin ağırlıkları azaltılırken, yanlış sınıflandırılanlarınki artırılır. Böylece, AdaBoost'un sonraki adımları önceki hatalara duyarlı hale gelir ve her yeni model, öncekilerin eksik bıraktığı alanları tamamlamayı hedefler.

AdaBoost'un en temel avantajı, zayıf öğrenicilerle bile yüksek doğruluğa ulaşabilmesidir. Ancak algoritma, verideki gürültüye ve aykırı değerlere karşı hassastır. Aşırı uyum riski bu nedenle yüksektir. Her örneğe eşit yaklaşması, hatalı etiketlenmiş verilerin model üzerinde aşırı etki yaratmasına neden olabilir [66].

3.3.2. İstatistiksel testler

Literatürde, makine öğrenimi algoritmalarının farklı veri kümelerinde ne kadar iyi performans gösterdiğini karşılaştırmak, dikkatli bir şekilde ele alınması gereken sık karşılaşılan bir sorundur. İlk bakışta, çeşitli algoritmaların, özellik seçimi tekniklerinin veya sınıflandırma stratejilerinin sonuçları metrik değerleri açısından önemli görünebilir. Ancak, bu farklılıkların istatistiksel olarak anlamlı olup olmadığını düşünmek önemlidir. Bu durumda, modelin üstünlüğü hakkında sağlam bir sonuca varmak, metriklerin yüzeysel bir karşılaştırmasından daha fazlasını gerektirebilir.

Bu çalışmada, sınıflandırma algoritmalarının ve öznitelik seçim yöntemlerinin farklı veri kümeleri üzerindeki başarım farklılıklarını değerlendirmenin ötesine geçilerek, bu farkların rastlantısal mı yoksa sistematik mi olduğunun anlaşılması hedeflenmiştir. Bu amaçla, birden fazla algoritmanın aynı veri kümeleri üzerinde elde ettiği performansların sıralı değerlendirmesini yaparak, genel anlamda yöntemler arasında istatistiksel olarak anlamlı

bir fark olup olmadığını test etmek için Friedman testi, yaranmış farkın hangi yöntem çiftleri arasında ortaya çıktığını belirleyebilmek için Nemenyi post-hoc testi uygulanmıştır.

Friedman testi, birden fazla algoritmanın birden fazla veri kümesi üzerindeki sıralı performanslarını değerlendirerek, genel olarak modeller arasında anlamlı fark olup olmadığını test eden çoklu karşılaştırma testlerinden biridir. Bu test, özellikle veri normal dağılım göstermediğinde veya varyans homojenliği gibi varsayımlar sağlanmadığında tercih edilen güvenilir bir yöntemdir [68]. Friedman testi sonucunda algoritmalar arasında anlamlı farklar tespit edilmesi durumunda, bu farkların hangi algoritmalar arasında ortaya çıktığını belirlemek amacıyla Nemenyi post-hoc testi uygulanmaktadır. Nemenyi testi, çoklu karşılaştırmalar için kullanılan parametrik olmayan bir yöntem olup, her bir algoritma çiftinin ortalama sıralama farkını esas alır. Yöntemin temel amacı, istatistiksel olarak anlamlı performans farklarını belirlemek ve sıralamalara dayalı güvenilir ikili karşılaştırmalar sunmaktır. Nemenyi testinde kullanılan temel ölçüt, kritik fark (Critical Difference – CD) değeridir (Eş. 3.10) [69]:

$$CD = q_{\alpha}\sqrt{\frac{k(k+1)}{6N}} \tag{3.10}$$

Burada, k karşılaştırılan algoritma sayısını, N kullanılan veri seti sayısını, q_{α} seçilen anlamlılık düzeyine karşılık gelen studentized range katsayısını ifade eder. İki algoritmanın ortalama sıralama farkı bu kritik farkın üzerinde ise, bu fark istatistiksel olarak anlamlı kabul edilir. Aksi takdirde, fark istatistiksel olarak önemsiz sayılır ve algoritmaların benzer performans gösterdiği değerlendirilir.

3.3.3. Değerlendirme ölçütleri

Makine öğrenimi tabanlı sınıflandırma problemlerinde farklı algoritmaları karşılaştırmak ve geliştirilen modellerin ne kadar iyi performans gösterdiğini ölçmek için farklı değerlendirme ölçütleri kullanılır. Bu değerlendirmede temel yapı taşı karışıklık matrisidir (confusion matrix). Bu matris, her bir örneğin gerçek sınıfı ile modelin tahmini arasındaki ilişkiyi detaylı olarak sunar. İkili sınıflandırma (binary classification) problemlerinde bu matris dört temel değeri içerir: Doğru Pozitif (TP), Yanlış Pozitif (FP), Doğru Negatif (TN) ve Yanlış Negatif (FN). Bu değerlerden türetilen metrikler sayesinde, modelin hem pozitif hem de negatif sınıfları ne derece doğru tanımladığı ölçülür [70]. Doğruluk (ACC), en yaygın kullanılan ölçütlerden biridir; doğru pozitif ve doğru negatif tahminlerin toplam örnek sayısına oranı ile hesaplanır (Eş. 3.11).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{P+N} \tag{3.11}$$

Ancak doğruluk özellikle dengesiz sınıf dağılımına sahip veri kümelerinde yanıltıcı olabilir. Bu tür durumlarda, kesinlik ve duyarlılık gibi özel metrikler kullanılır. Kesinlik (PRE), pozitif tahminlerin ne kadarının doğru olduğunu gösterirken duyarlılık (SN), gerçek pozitiflerin ne kadarının doğru tespit edildiğini ifade eder. Bu iki metrik arasındaki denge ise F1 skoru (F Score) ile sağlanır. F1 skoru, kesinlik ve duyarlılığın harmonik ortalamasıdır ve özellikle pozitif sınıfın kritik olduğu durumlarda ön plana çıkar (Eş. 3.12, 3.13, 3.14) [71]:

$$PRE = \frac{TP}{TP+FP} \tag{3.12}$$

$$F\ Score = \frac{2 \times precision \times recall}{precision \times recall} \tag{3.13}$$

$$SN = \frac{TP}{TP+FN} = \frac{TP}{P} \tag{3.14}$$

4. DENEYSEL ÇALIŞMALAR

Bu bölümde, önerilen yöntemlerin etkinliğini değerlendirmek amacıyla yapılan deneysel çalışmalar ve çıktıları sunulmaktadır. Tüm deneyler Python programlama dili kullanılarak gerçekleştirilmiştir. Bu süreçte scikit-learn [72], imblearn [73], matplotlib [74], seaborn [75], scikit-posthocs [76], yellowbrick [77] ve skrebate [78] gibi çeşitli kütüphanelerden faydalanılmıştır.

Her bir sınıflandırma algoritması için hiperparametre ayarlamaları, modelin performansını artırmak amacıyla parametre kombinasyonlarının daha verimli ve rastlantısal olarak denenmesini sağlayan RandomizedSearchCV yöntemi tercih edilmiştir. Böylece, her sınıflandırıcı için geniş bir parametre aralığı içinden en uygun değerler belirlenmiş ve modellerin sınıflandırma başarımı optimize edilmiştir.

kNN algoritması için n_neighbors, weights, algorithm, metric ve p parametreleri değerlendirilmiştir. n_neighbors parametresi, komşu sayısını belirtirken; weights, komşuların eşit mi yoksa uzaklığa göre ağırlıklı mı değerlendirileceğini belirler. algorithm parametresi, komşu arama algoritmasını (ball_tree veya kd_tree) seçmekte kullanılmıştır. metric ile uzaklık ölçütü olarak euclidean ve minkowski tercih edilmiştir. p parametresi ise Minkowski metriğinin derecesini belirtmektedir.

DT modeli için ise yalnızca criterion parametresi optimize edilmiştir. Bu parametre, düğüm bölme kriteri olarak gini, entropy ve log_loss seçeneklerini içermektedir. LR algoritmasında, düzenlileştirme gücünü belirleyen C parametresi kullanılmış ve bu parametre için 10^-4 ile 10^4 arasında logaritmik aralıkta 20 farklı değer test edilmiştir.

RF modeli için üç farklı hiperparametre (n_estimators, criterion, max_features) optimize edilmiştir. n_estimators parametresi, oluşturulacak karar ağaçlarının sayısını belirlemekte olup, modelin genelleme yeteneğini artırmak amacıyla çeşitli değerlerle denenmiştir. criterion parametresi, her ağacın hangi kritere göre bölüneceğini belirlerken (gini, entropy, log_loss), max_features parametresi ise her bölünmede kaç özelliğin göz önünde bulundurulacağını belirlemektedir. Farklı kombinasyonlar denenerek en iyi sınıflandırma başarımı sağlayan parametre değerleri belirlenmiştir.

ETs algoritmasında da criterion ve max_features parametreleri optimize edilmiştir. Bu parametreler RF modelinde olduğu gibi çalışmaktadır. AdaBoost sınıflandırıcısı için yalnızca random_state parametresi sabit bir değer (1) olarak belirlenmiştir. Bu, deneylerin tekrarlanabilirliğini sağlamak için yapılmıştır.

XGBoost algoritması için hiperparametre optimizasyonu kapsamında n_estimators, learning_rate ve max_depth parametreleri dikkate alınmıştır. n_estimators parametresi, modelde oluşturulacak zayıf öğrenicilerin (ağaçların) sayısını belirtirken; learning_rate, her bir öğrenicinin modele olan katkı oranını belirlemektedir. max_depth parametresi ise ağaçların maksimum derinliğini kontrol ederek modelin karmaşıklığını etkilemektedir. Yapılan deneyler sonucunda, bu parametreler için en iyi sınıflandırma performansını sağlayan değer kombinasyonları belirlenmiştir.

Söz konusu parametreler ve değer aralıkları Çizelge 4.1'de açıklamalarıyla sunulmuştur.

Çizelge 4.1. Sınıflandırma Algoritmaları için Kullanılan Parametreler

Sınıflandırıcı	Parametreler	Hiperparametre Değerleri
kNN	*n_neighbors*	2, 5, 10
	weights	uniform, distance
	algorithm	ball_tree, kd_tree
	metric	euclidean, minkowski
	p	1
DT	*criterion*	gini, entropy, log_loss
LR	*C*	logspace(-4, 4, 20)
RF	*n_estimators*	10, 50, 100
	criterion	gini, entropy, log_loss
	max_features	sqrt, log2
ETs	*criterion*	gini, entropy, log_loss
	max_features	sqrt, log2
AdaBoost	*random_state*	1
XGBoost	*n_estimators*	randint(50, 200)
	learning_rate	uniform(0.01, 0.2)
	max_depth	randint(3, 10)

Model değerlendirmesi için veri setleri %80 eğitim, %20 test olarak bölünmüştür. Her kombinasyon için 10 farklı rastgele eğitim/test ayrımı yapılmış ve modelin kararlılığı ile genellenebilirliği incelenmiştir. Böylece, farklı öznitelik seçim yöntemlerinin ve sınıflandırıcıların performansları objektif olarak karşılaştırılmıştır.

Performans ölçütleri doğruluk, kesinlik, duyarlılık (hassasiyet) ve F1 skoru olarak belirlenmiş ve sonuçlar karşılaştırılmıştır. Sonuçların istatistiksel anlamlılığını değerlendirmek için doğruluk ortalamaları üzerinde parametrik olmayan Friedman testi yapılmış ve anlamlı farklılık durumunda Nemenyi post-hoc testi ile ikili karşılaştırmalar gerçekleştirilmiştir.

4.1. Veri Setleri üzerinde Deneysel Sonuçlar

Bu bölümde, farklı biyomedikal gen ekspresyon veri setleri üzerinde gerçekleştirilen deneysel çalışmaların bulguları sunulmaktadır. Çalışma kapsamında, yüksek boyutlu yapıya sahip veri setleri kullanılmış, ikili ve çoklu sınıflandırma problemi çerçevesinde değerlendirilmiş ve çeşitli makine öğrenimi algoritmalarıyla analiz edilmiştir. Özellikle, 11 farklı öznitelik seçim kombinasyonunun (FS0–FS11) sınıflandırma performansına etkisi kapsamlı biçimde incelenmiştir. Her kombinasyon, klasik istatistiksel yöntemlerden bilgi temelli yaklaşımlara, GA gibi sezgisel arama stratejilerinden çoklu yöntem birleşimlerine kadar geniş bir yelpazeyi kapsamaktadır. Elde edilen sonuçlar doğrultusunda, her veri setinde farklı öznitelik seçim stratejilerinin sınıflandırma başarısını ve işlem süresini anlamlı düzeyde etkileyebildiği gözlemlenmiştir. Bu bölümde, veri setlerine özel bulgular detaylandırılarak karşılaştırmalı analizler ve istatistiksel değerlendirmeler sunulmaktadır.

GSE70947 veri seti ile elde edilen bulgular

Meme kanserine ait GSE70947 veri seti, toplam 289 örnekten oluşmakta olup, ikili sınıflandırma problemi için uygun bir yapıya sahiptir. Veri setindeki örnekler, "normal" ve "breast_adenocarcinoma" olmak üzere iki sınıfa ayrılmıştır. Her bir örnek, gen ekspresyon düzeylerini temsil eden yüksek boyutlu öznitelikleri içermektedir.

Veri setinde model performansını iyileştirmek ve gereksiz verileri kaldırmak için farklı özellik seçimi teknikleri kullanılmıştır. Her gen yerine yalnızca en önemli ve benzersiz özelliklerin seçilmesini sağlayarak, bu teknikler modelin öğrenme sürecini hızlandırır ve doğruluk oranlarını artırır. Çalışma kapsamında, veri seti üzerinde yapılan özellik seçimi işleminin ardından veri setinin boyutunda önemli bir değişiklik gözlemlenmiştir. Başlangıçta, veri seti 289 örnek ve 35,983 özellikten oluşmaktayken, özellik seçimi sonrası veri setinin boyutu 289 örnek ve 35,981 özellik olarak güncellenmiştir. Bu değişiklik, özellik seçimi sürecinin gereksiz veya düşük öneme sahip özellikleri çıkarmasından kaynaklanmaktadır. Sonuç olarak, veri setinin boyutu küçülmüş ve modelin eğitimi için daha verimli hale gelmiştir.

Çeşitli özellik seçimi tekniklerinin etkilerini araştırmak için, kNN algoritması temel sınıflandırıcı olarak kullanılmıştır. Seçilen gen sayısı, eğitim süresi ve performans göstergeleri her FS yaklaşımı için karşılaştırılmıştır. Çeşitli özellik seçimi kombinasyonları, kNN sınıflandırıcısı kullanılarak karşılaştırılmış, elde edilen doğruluk oranları, kesinlik, duyarlılık, F1 skorları ve işlem süreleri değerlendirilmiştir.

Yapılan analizlerde, özellik seçimi uygulanmamış durumda (FS0), modelin %77 ile %91 arasında değişen doğruluk oranlarıyla ortalama %84 doğruluk sağladığı, ancak yüksek boyutlu veri nedeniyle işlem süresinin oldukça uzun (15-17 saniye) olduğu gözlemlenmiştir. Bu durum, özellik seçiminin gerekliliğini ortaya koymuştur. Klasik istatistiksel yöntemler olan FS1 ve FS2, doğruluk açısından FS0'a benzer ya da hafif değişken sonuçlar üretmiş; ancak işlem süresini ciddi oranda kısaltarak (yaklaşık 0.15-0.2 saniye) yüksek boyutlu verilerde pratiklik kazandırmıştır. FS3 ve FS4 yöntemleri, hem doğruluk (%81-%93) hem de diğer metriklerde iyileşme sağlamış, işlem süresini de kısa tutmuştur. Özellikle IG ve ReliefF yöntemleri yüksek doğruluk oranlarıyla öne çıkarken, ReliefF bazen beklenmedik süre artışları gösterse de genel verimlilik açısından etkili olmuştur.

Birleşik yöntemlerin kullanıldığı FS5 ve FS6 kombinasyonlarında, modeller dengeli ve başarılı performans sergilemiştir. FS5'te yüksek ve istikrarlı doğruluklar (%74-%93) elde edilirken, işlem süreleri de makul seviyelerde (0.16-0.23 saniye) kalmıştır. FS6'da özelliklerin kesişimiyle elde edilen seçim, model genelinde anlamlılığı artırarak %75-%90 arasında tutarlı doğruluk sağlamıştır. FS7'de ise GA dışındaki dört temel yöntemin

birleşimiyle kNN algoritması %81-%95 aralığında yüksek doğruluklara ulaşmış; genellikle kısa işlem süreleri kaydedilmiş olsa da bazı durumlarda işlem süresinin beklenenden uzun olduğu gözlemlenmiştir. GA'nın dahil edildiği FS8-FS11 kombinasyonlarında ise doğruluk oranlarında dalgalanmalar görülmüş, genel olarak orta düzey performanslar elde edilmiş; FS9 ve FS10'da belirgin değişkenlik gözlemlenirken, FS11'de işlem süresi bir miktar daha artmıştır. Bu sonuçlar, özellik seçimi yöntemlerinin doğruluk, denge ve işlem süresi açısından modele önemli katkılar sunduğunu ortaya koymaktadır. Bu performans metriklerinin ortalama değerleri Çizelge 4.2'de sunulmuştur.

Çizelge 4.2. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE70947)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,84	0,86	0,84	0,84	16,10
FS1	0,82	0,83	0,82	0,82	16,39
FS2	0,85	0,87	0,85	0,85	0,16
FS3	0,87	0,89	0,87	0,89	0,16
FS4	0,85	0,86	0,85	0,85	0,33
FS5	0,84	0,83	0,85	0,84	0,16
FS6	0,84	0,84	0,85	0,84	0,16
FS7	0,89	0,88	0,90	0,89	0,37
FS8	0,80	0,71	0,81	0,80	0,16
FS9	0,80	0,72	0,73	0,80	0,16
FS10	0,81	0,70	0,80	0,81	0,17
FS11	0,76	0,76	0,77	0,76	0,16

kNN sınıflandırıcısındaki çeşitli özellik seçimi yöntemlerinin performansını karşılaştırmak için bazı istatistiksel testler uygulanmıştır. Bu testler, herhangi bir önemli performans artışı olup olmadığını ve hangi yöntemlerin performansının diğerlerine göre üstünlük sağladığını görmek için motivasyonu belirlemek üzere seçilmiş ve tüm yöntemlerin genel performansları başlangıçta Friedman testi ile karşılaştırılmıştır. Daha sonra, çiftler arası karşılaştırma için Nemenyi testi uygulanmıştır.

kNN sınıflandırıcısı üzerine uygulanmış Friedman testine göre, özellik seçimi yöntemleri arasında ortalama olarak önemli bir fark bulunmamıştır ($Chi^2 = 6,81$, $p = 0,6573$). Bu, yöntemlerin genel performanslarının benzer ve ayırt edilemez olduğunu göstermiştir. Ancak, Friedman testi yalnızca genel bir fark olup olmadığını ortaya koyar ve çiftler arasında bir fark olup olmadığını göstermez. Bu nedenle, Nemenyi post hoc testini kullanarak çiftler arası bir karşılaştırma gerçekleştirdi. Nemenyi post-hoc testleri, bazı

yöntem çiftleri arasındaki performanslarda önemli farklılıklar ($p < 0,05$) olduğunu gösterdi. Özellikle, bilgi tabanlı özellik seçimi yöntemleri ile bazı diğer özellik seçimi yöntemlerinde önemli farklılıklar ortaya çıkmıştır. Bu ayrıntılı karşılaştırma Çizelge 4.3'te gösterilmiştir.

Çizelge 4.3. Nemenyi Testi ile Performans Farklılıkları (GSE70947)

Yöntem Çifti	P-Değeri
FS2 vs FS3	0,000655
FS2 vs FS5	0,007590
FS9 vs FS2	0,000429
FS6 vs FS3	0,028637
FS9 vs FS4	0,042794
FS9 vs FS6	0,020899

Çalışmanın devamında, yalnızca kNN ile sınırlı kalmayarak, farklı sınıflandırıcılar kullanılarak öznitelik seçim yöntemlerinin etkileri kapsamlı bir şekilde değerlendirilmiştir. Çeşitli sınıflandırıcılar, seçilen genlerin performansa katkısını ortaya koymak amacıyla test edilmiştir. Her bir sınıflandırıcı için farklı öznitelik seçim teknikleri uygulanmış, böylece model başarısının algoritma ve öznitelik kombinasyonuna göre nasıl değiştiği incelenmiştir. Bu kapsamlı analiz hem model doğruluğu hem de işlem süresi açısından optimizasyon olanaklarını belirlemeyi hedeflemiştir.

Analiz sonuçları, özellik seçim yöntemlerinin sınıflandırma modellerinin performansını önemli ölçüde etkilediğini ve her yöntemin farklı algoritmalarla farklı düzeylerde etkileşim gösterdiğini ortaya koymuştur. Genel olarak, tekil özellik seçimi yöntemleri (FS1, FS2, FS3, FS4), işlem sürelerinde belirgin kısalmalar sağlarken, doğruluk oranlarında da önemli iyileşmeler göstermiştir. Özellikle FKs (FS2) ve IG (FS3) gibi yöntemler, LR, XGBoost, ETs ve RF gibi modellerle yüksek doğruluk ve dengeli metrikler elde edilmesini sağlamıştır. Bu yöntemler, modelin karmaşıklığını azaltırken genelleme yeteneğini artırma potansiyellerini kanıtlamıştır. ReliefF (FS4) ise kNN ve LR ile yüksek ve istikrarlı doğruluklar sergileyerek öne çıkmıştır.

Çoklu özellik seçimi yöntemlerinin birlikte kullanımı (FS5, FS6, FS7) ise model performansında daha da dikkat çekici iyileşmeler sağlamıştır. Özellikle FS5 kombinasyonunda, kNN, LR, RF ve ETs algoritmaları %92 ila %96.5'e varan yüksek

doğruluk oranlarına ulaşmıştır. Bu durum, farklı özellik seçim yöntemlerinin birleşiminden elde edilen zengin ve anlamlı özellik setlerinin, özellikle topluluk modellerin karmaşık öğrenme kapasitesini optimize ettiğini göstermektedir. Ancak, FS6 kombinasyonunda doğruluk oranları bir miktar daha düşmüş, bu da kesişim yönteminin bazen çok seçici davranarak önemli bilgileri göz ardı edebileceğine işaret etmektedir. FS7 de genel olarak yüksek ve stabil doğruluklar sunmuştur.

GA dahil kombinasyonlar (FS8, FS9, FS10, FS11) ise performans açısından daha değişken sonuçlar sergilemiştir. GA'nın tek başına veya sınırlı kombinasyonlarda performansı, diğer yöntemlerin sağladığı tutarlılık ve yüksek doğruluk seviyelerine tam olarak ulaşamamıştır. Bazı algoritmalarda, özellikle FS9 ve FS10'da doğrulukda dalgalanmalar olmuştur. Ancak FS8'de kNN ve AdaBoost'un, FS11'de ise kNN ve LR'nin yüksek verimliliği, GA'nın uygun algoritmalarla birleştirildiğinde yararlı olabileceğini göstermiştir. Ancak GA'nın bazı kombinasyonlarda performansı stabilize etme ve iyileştirme konusunda diğer yöntemler kadar başarılı olamadığı görülmüştür.

Özetle, bu analizler farklı özellik seçimi yaklaşımlarının sınıflandırıcıların başarısı için kritik bir rol oynadığını net bir şekilde göstermektedir. Daha spesifik olarak, bilgi-teorik tekniklerin ve bu tekniklerin kombinasyonlarının bir araya getirilmesinin doğruluk ve hesaplama süresi açısından en umut verici sonuçları verdiği görülmüştür.

Sınıflandırma işleminin ardından her bir sınıflandırıcı için, uygulanan farklı özellik seçimi yöntemlerinin performanslarının istatistiksel olarak karşılaştırılması Friedman ve Nemenyi post-hoc testleri uygulanmıştır.

Yedi farklı sınıflandırıcı (kNN, DT, LR, RF, ETs, AdaBoost, XGBoost) üzerinde yapılan Friedman testleri sonucunda, yalnızca ETs sınıflandırıcısı anlamlı farklılık göstermiştir (Chi^2=18.345, p=0.031), diğer modellerde ise p-değerleri 0.09–0.263 arasında değişmiş ve anlamlılık eşiğini aşmamıştır. Ancak ikili karşılaştırmalar için uygulanan Nemenyi testi, her sınıflandırıcıda bazı yöntem çiftleri arasında istatistiksel olarak anlamlı farklar olduğunu göstermiştir. Özellikle FS0 yöntem ile FS2, FS4, FS6, FS10 ve FS11 yöntemleri arasında sıkça anlamlı farklar gözlenmiştir. Örneğin, RF sınıflandırıcıda FS10 ile FS0 (p=0.000072) ve FS6 ile FS10 (p=0.000024) arasında, ETs sınıflandırıcıda FS6 ile FS10 (p=0.000053) ve FS11 ile FS10 (p=0.000240) gibi anlamlı karşılaştırmalar yapılmıştır.

Özellikle FS10 ve FS11, birçok sınıflandırıcıda diğer yöntemlerle karşılaştırıldığında anlamlı farklılıklar yaratmıştır. kNN'de FS11 ile FS10 arasında p=0.000028, LR'da FS11 ile FS0 p=0.000033, AdaBoost'ta FS11 ile FS10 p=0.000864, XGBoost'ta FS10 ile FS11 p=0.000014 olarak belirlenmiş ve bu durum GA'nın etkisinin önemli olduğunu göstermiştir. Ayrıca FS6'nın da hem ETs hem de RF'de öne çıktığı görülmektedir. Sonuç olarak, genel testlerde tüm yöntemler arasında güçlü farklar çıkmasa da özellikle GA içeren ve özellik kesişimi yöntemlerinin, bazı modellerde doğruluğu ve dengeli metrikleri anlamlı şekilde etkilediği ortaya çıkmıştır. Bu da model performansını artırmada uygun özellik seçimi stratejilerinin kritik olduğunu göstermektedir.

Bu veri setinin genelinde elde edilen bulgular, özellik seçimi yöntemlerinin gen ekspresyon verileri üzerinde sınıflandırma başarısını önemli ölçüde etkilediğini açıkça ortaya koymuştur. Hem kNN algoritması hem de diğer sınıflandırıcılar üzerinden yapılan kapsamlı analizlerde, tekil ve birleşik öznitelik seçimi stratejilerinin model doğruluğunu artırdığı, işlem süresini önemli ölçüde azalttığı ve modelin genelleme yeteneğini geliştirdiği gözlemlenmiştir. Özellikle bilgi temelli yöntemler (FS2, FS3) ve bu yöntemlerin kombinasyonları (FS5, FS6, FS7), birçok sınıflandırıcıda yüksek doğruluk (%89-96) ve dengeli metrikler sunarak öne çıkmıştır. GA tabanlı yöntemler ise bazı durumlarda yüksek performans sağlasa da tutarlılık konusunda dalgalı sonuçlar vermiştir.

Özetlemek gerekirse, özellik seçimi sınıfları daha iyi ayırma ve işlem verimliliği açısından önemli bir adımdır. Özellikle yüksek boyutlu biyolojik veri kümelerinde, sınıflandırma doğruluğunun yanı sıra eğitim süresi ve kararlılık, optimum özellik seçimi yöntemi ile önemli ölçüde artırılabilir. Yapılan istatistiksel testler, belirli yöntem çiftleri arasında anlamlı performans farklarını ortaya koymuş, bu da özellik seçiminde stratejik yaklaşımın gerekliliğini bir kez daha vurgulamıştır. Çalışmanın genel çıkarımı, özellikle bilgi temelli ve kombinasyon yöntemlerinin, farklı sınıflandırıcılarla birlikte kullanıldığında en yüksek başarıyı sağladığı yönündedir. Bu nedenle, model kurulumunda yalnızca algoritma seçimi değil, uygun özellik seçimi yönteminin tercih edilmesi de başarıyı belirleyen anahtar faktörlerden biridir.

GSE14520_U133A veri seti ile elde edilen bulgular

Karaciğer kanserine ait GSE14520_U133A veri seti, toplam 357 örnekten oluşmakta olup, ikili sınıflandırma problemi için uygun bir yapıya sahiptir. Bu veri setindeki örnekler, "HCC" ve normal" olmak üzere iki sınıfa ayrılmıştır. Her bir örnek, gen ekspresyon düzeylerini temsil eden yüksek boyutlu öznitelikleri içermektedir. Veri seti üzerinde yapılan özellik seçimi işlemi sonrası, başlangıçtaki veri seti 357 örnek ve 22,279 özellikten oluşurken, bu sayı 357 örnek ve 22,277 özellik olarak güncellenmiştir. Bu küçücük boyut farkı, özellik seçimi sürecinde veri setindeki gereksiz veya düşük öneme sahip özelliklerin kaldırılmasından kaynaklanmaktadır. İlk başta bu veri seti üzerinde diğer veri setinde olduğu gibi kNN algoritması kullanılarak, çeşitli öznitelik seçim yöntemlerinin model performansı üzerindeki etkileri incelenmiştir.

Yapılan analizlerde, özellik seçimi uygulanmadığı durumda (FS0), kNN modeli %93 ile %98.6 arasında değişen doğruluk oranlarıyla yüksek bir başarı sergilemiş, ancak yüksek boyutlu veri nedeniyle işlem süreleri ortalama 15–17 saniye gibi oldukça uzun olmuştur. Klasik istatistiksel yöntemler olan FS1 ve FS2, doğruluğu korurken (%93–98.6), eğitim süresini ciddi şekilde düşürmüş (0.17–0.23 saniye) ve bu yönüyle yüksek boyutlu verilerde pratiklik sağlamıştır. Bilgi teorik temelli FS3 ve FS4 yöntemleri ise benzer şekilde yüksek doğruluklar (%93–100) ve dengeli metrikler sunmuş; özellikle ReliefF yöntemi bazı senaryolarda %100 doğruluk ve F1 skoru ile öne çıkmıştır. Eğitim süresi genellikle düşük kalmış olsa da ReliefF'te nadiren 1.9 saniyeye ulaşan süreler gözlemlenmiştir.

Kombine yöntemlerin kullanıldığı FS5-FS7, genellikle en yüksek performansı sağlayan kombinasyonlar olmuştur. FS5, %95-100 doğruluk aralığıyla dikkat çekerken, işlem süresi 0,17-0,26 saniye aralığında kalmış, GA ve kesişim yöntemleri hariç tutulduğunda FS7 ise %91-98 doğrulukla yüksek başarı sağlamış, ancak bazı durumlarda 1,95 saniyeye kadar işlem süreleri kaydedilmişdir. Öte yandan, FS6 %79–93 arasında değişen doğruluk oranlarıyla daha değişken bir performans göstermiştir. GA içeren FS8–FS11 kombinasyonları ise genel olarak %87–98 doğruluk aralığında istikrarlı sonuçlar üretmiş ve işlem süreleri 0.14–0.23 saniye aralığında makul seviyelerde kalmıştır. Tüm bu sonuçlar, uygun özellik seçimi yöntemlerinin modelin doğruluğunu koruyarak eğitim süresini önemli ölçüde azaltabileceğini ve performansı artırabileceğini göstermektedir. Bu performans metriklerinin ortalama değerleri Çizelge 4.4'de sunulmuştur.

Çizelge 4.4. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE14520_U133A)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,94	0,94	0,94	0,94	13,75
FS1	0,95	0,95	0,95	0,95	0,17
FS2	0,96	0,96	0,96	0,96	0,17
FS3	0,96	0,96	0,96	0,96	0,17
FS4	0,96	0,96	0,96	0,96	0,59
FS5	0,98	0,98	0,98	0,98	0,19
FS6	0,88	0,88	0,88	0,88	0,17
FS7	0,98	0,96	0,96	0,96	0,54
FS8	0,91	0,91	0,91	0,91	0,17
FS9	0,94	0,94	0,94	0,94	0,17
FS10	0,95	0,95	0,95	0,95	0,18
FS11	0,95	0,95	0,95	0,95	0,18

Diğer veri setine benzer şekilde, kNN sınıflandırıcısı üzerinde yapılan analizlerde Friedman testi sonuçları, farklı özellik seçimi yöntemleri arasında genel anlamda istatistiksel olarak anlamlı bir fark olmadığını göstermiştir ($Chi^2 = 10.91$, $p = 0.2816$). Bununla birlikte, Nemenyi testine dayanarak, Çizelge 4.5'de bildirildiği gibi, bazı yöntem çiftleri arasında anlamlı farklılıklar bulunmuştur.

Çizelge 4.5. Nemenyi Testi ile Performans Farklılıkları (GSE14520_U133A)

Yöntem Çifti	P-Değeri
FS3 vs FS11	0,0017
FS4 vs FS11	0,0032
FS5 vs FS11	0,0053
FS8 vs FS10	0,0037
FS10 vs FS11	5,0e-6
FS1 vs FS11	0,0317
FS2 vs FS11	0,0388
FS9 vs FS11	0,0351

Çalışmanın devamında yalnızca kNN yaklaşımıyla kısıtlı kalmayarak, farklı sınıflandırıcılar ile seçim yöntemlerinin etkileri kapsamlı bir şekilde değerlendirilmiştir.

Elde edilen sonuçlar, öznitelik seçim yöntemlerinin sınıflandırıcıların performansını önemli ölçüde etkilediğini göstermiştir. Tekil yöntemler arasında FKs (FS2) ve IG (FS3) öne çıkmış; LR, XGBoost, ETs ve RF modellerinde %85 ile %94 aralığında doğruluk oranları sağlarken, işlem süresinde %30'a varan azalma gözlenmiştir. ReliefF (FS4)

yöntemi ise kNN ve LR algoritmaları ile %82–%90 arasında yüksek ve tutarlı doğruluk değerleri sunmuştur.

Çoklu öznitelik seçimi yöntemlerinin birleşimi (FS5, FS6, FS7) performansta daha belirgin iyileşmeler sağlamıştır. Özellikle FS5 kombinasyonunda kNN, LR, RF ve ETs modelleri %92 ila %96,5 arasında oldukça yüksek doğruluk oranları elde etmiştir. Ancak FS6'da doğruluklar %88–%91 arasında biraz gerileyerek, kesişim tabanlı seçimin bazen önemli genleri göz ardı ettiğini göstermiştir. FS7 ise genel olarak %90–%95 aralığında yüksek ve stabil sonuçlar vermiştir.

GA içeren kombinasyonlar ise performans açısından daha değişken olmuştur. FS8 kombinasyonunda kNN ve AdaBoost modelleri %90'ın üzerinde doğruluk gösterirken, FS9 ve FS10'da bazı algoritmalarda %80–%85'e kadar düşen doğruluklar ve performans dalgalanmaları gözlemlenmiştir. FS11'de kNN ve LR %91–%94 doğrulukla GA'nın uygun kombinasyonlarla faydalı olabileceğini göstermiştir.

İşlem süreleri açısından, tekil öznitelik seçimleri ortalama 20-35 saniye aralığında hızlı sonuç verirken, çoklu yöntem kombinasyonları 40-70 saniye arası işlem süreleri göstermiştir. GA içeren kombinasyonlar ise zaman zaman 80 saniyeyi aşarak daha yüksek hesaplama maliyeti getirmiştir.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. Yapılan istatistiksel analizler, farklı sınıflandırıcılar üzerinde uygulanan özellik seçim yöntemlerinin performans üzerinde anlamlı etkiler yarattığını göstermektedir. Friedman testleri genel olarak sınıflar arasında istatistiksel fark olmadığını ortaya koyarken (örneğin kNN: $p=0.726$, DT: $p=0.183$, LR: $p=0.672$, RF: $p=0.266$), yalnızca AdaBoost sınıflandırıcısında anlamlı fark bulunmuştur ($p=0.023$). Ancak detaylı ikili karşılaştırmalar için uygulanan Nemenyi testleri, özellikle FS8 yöntemi ile diğer yöntemler arasında tutarlı şekilde anlamlı farklar olduğunu ortaya koymuştur. FS3, FS5, FS0 ve FS1 gibi yöntemlerle yapılan karşılaştırmalarda, FS8'in neredeyse tüm sınıflandırıcılarda istatistiksel olarak daha düşük performans gösterdiği belirlenmiştir. Örneğin, kNN'de FS3 vs FS8 ($p=0.000033$), LR'de FS9 vs FS8 ($p=0.000017$), RF'de FS9 vs FS8 ($p=0.000005$), ETs'de FS9 vs FS8 ($p=0.000007$) ve XGBoost'ta yine FS9 vs FS8 ($p=0.000005$) gibi son derece anlamlı farklar elde edilmiştir.

Bu sonuçlar, özellikle GA içeren FS8 yönteminin birçok sınıflandırıcıyla zayıf uyum sağladığını ve model performansını düşürebileceğini göstermektedir. Buna karşılık, FS3, FS5 ve FS0 yöntemlerinin FS8'e kıyasla anlamlı ölçüde daha başarılı olduğu görülmüştür. AdaBoost sınıflandırıcısında genel anlamda farklılıklar gözlemlenmesi ve yine FS8'in en zayıf yöntem olarak öne çıkması, bu yöntemin dikkatli ve kontrollü bir şekilde kullanılmasını gerektirdiğini ortaya koymaktadır. Genel olarak, Nemenyi testleri FS8'in sınıflandırma modelleri arasında en zayıf yöntem olduğuna dair güçlü kanıtlar sunmaktadır.

Özetle, GSE14520_U133A veri setine dayalı kapsamlı bir analize göre, özellik seçimi yöntemlerinin sınıflandırma modellerinin doğruluğu, hesaplama süresi ve verimliliği üzerinde doğrudan bir etkiye sahip olabilir. Özellikle bilgi temelli yöntemler (FS2, FS3) ve bu yöntemlerin birleşimleri (FS5, FS7), birçok modelde yüksek doğruluk oranları (%92–%98) ve kısa işlem süreleri ile öne çıkmıştır. Buna karşın, GA içeren FS8 yöntemi, neredeyse tüm sınıflandırıcılarda en zayıf performansı göstermiş ve istatistiksel testlerde sıkça anlamlı şekilde başarısız sonuçlar üretmiştir. Nemenyi testi sonuçları da FS8'in birçok güçlü yöntemle karşılaştırıldığında anlamlı derecede düşük performans gösterdiğini desteklemiştir. Dolayısıyla, özellik seçimi stratejisi belirlenirken yalnızca doğruluk değil, işlem süresi, istikrar ve modelle uyumluluk gibi faktörler de göz önünde bulundurulmalı; özellikle GA gibi yöntemlerin kombinasyonlarında dikkatli olunmalıdır.

GSE44076 veri seti ile elde edilen bulgular

Kolorektal kanserine ait GSE44076 veri seti, toplam 194 örnekten oluşmakta olup, ikili sınıflandırma problemi için uygun bir yapıya sahiptir. Bu veri setindeki örnekler, "Adenokarsinom" ve "normal" olmak üzere iki sınıfa ayrılmıştır. Veri seti üzerinde yapılan özellik seçimi işlemi sonrası, başlangıçtaki veri seti 194 örnek ve 49,388 özellikten oluşurken, bu sayı 194 örnek ve 49,386 özellik olarak güncellenmiştir. Bu veri seti üzerinde bir önceki analizde olduğu gibi kNN algoritması kullanılarak, çeşitli öznitelik seçim yöntemlerinin model performansı üzerindeki etkileri incelenmiştir.

Yapılan analizler sonucunda, FS0 durumunda modelin %100 doğruluk, F1 skoru, kesinlik ve hassasiyet gibi mükemmel performans değerleri sunduğu, ancak işlem süresinin 13.37 saniye gibi oldukça yüksek olduğu görülmüştür. Bu durum, özellikle büyük veri setlerinde işlem süresinin sınırlayıcı bir faktör olabileceğini göstermektedir. FS1 ve FS2 gibi klasik

istatistiksel yöntemler, doğruluktan ödün vermeden (%97–98) işlem süresini önemli ölçüde azaltarak (yaklaşık 0.15–0.17 saniye) pratik bir avantaj sağlamıştır. FS3–FS7 arasında yer alan bilgi teorik temelli ve kombinasyon yöntemleri, %93–100 aralığında değişen doğruluklarla yüksek başarı sunmuş, ancak özellikle ReliefF ve GA içeren yöntemlerde işlem süreleri 1.9 saniyeye kadar çıkabilmiştir. GA tabanlı FS8–FS11 kombinasyonlarında da doğruluk oranları yüksek kalmış, ancak işlem süreleri ortalama 1.2 saniyeye kadar artmıştır. Bu performans metriklerinin ortalama değerleri Çizelge 4.6'da sunulmuştur.

Çizelge 4.6. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE44076)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,99	0,99	0,99	0,99	11,52
FS1	0,99	0,99	0,99	0,99	0,16
FS2	0,99	0,99	0,99	0,99	0,15
FS3	1.00	1.00	1.00	1.00	0,16
FS4	0,98	0,98	0,99	0,98	0,32
FS5	0,99	0,99	0,99	0,99	0,16
FS6	0,95	0,95	0,95	0,95	0,15
FS7	0,99	0,99	0,99	0,99	0,34
FS8	0,97	0,97	0,98	0,97	0,16
FS9	0,97	0,97	0,98	0,97	0,15
FS10	0,97	0,97	0,98	0,97	0,16
FS11	0,96	0,96	0,96	0,96	0,17

Diğer veri setlerinde olduğu gibi kNN sınıflandırıcısında kullanılan farklı özellik seçimi yöntemlerinin performansını değerlendirmek amacıyla yapılan analizlerde, istatistiksel testler önemli bulgular ortaya koymuştur. Öncelikle, Friedman testi sonuçları ($Chi^2 = 12.36$, $p=0.1935$) genel olarak farklı özellik seçimi yöntemleri arasında istatistiksel olarak anlamlı bir fark olmadığını göstermiştir. Yöntem çiftleri arasındaki daha ince farklılıkları ortaya koymak için Nemenyi post-hoc testi uygulanmıştır. Çizelge 4.7, Nemenyi testi sonuçlarına göre anlamlı performans farkı gösteren yöntem çiftlerini özetliyor.

Çizelge 4.7. Nemenyi Testi ile Performans Farklılıkları (GSE44076)

Yöntem Çifti	P-Değeri
FS2 vs FS3	0,0168
FS2 vs FS9	0,0388
FS2 vs FS10	0,0472
FS3 vs FS11	0,0025
FS9 vs FS11	0,0067
FS10 vs FS11	0,0085

Bu çalışmada da sadece kNN algoritmasıyla kısıtlı kalınmayıp, çeşitli sınıflandırıcılar kullanılarak öznitelik seçim yöntemlerinin etkileri detaylı bir şekilde incelenmiştir.

FS0 kombinasyonunda, tüm sınıflandırıcılar %100 doğruluk, F1 skor, kesinlik ve duyarlılık değerlerine ulaşarak iyi bir performans sergilemiştir. Bu, veri setinin modelin tam potansiyelini ortaya çıkarabilecek kadar iyi dengelenmiş veya ayrılabilir olduğunu göstermektedir. Ancak, modellerin işlem süreleri arasında belirgin farklılıklar gözlemlenmiştir. ETs (1.36 saniye) en hızlı model olarak öne çıkarken, kNN (11.17 saniye) ve XGBoost (9.95 saniye) gibi modeller daha uzun süreli işlemler gerçekleştirmiştir.

Tekil öznitelik seçimi yöntemlerinin incelendiği kombinasyonlarda, modellerin genel performansının yüksek doğruluk oranlarıyla (%94.7 - %100) korunduğu gözlemlenmiştir; özellikle kNN, tüm bu durumlarda %100 doğrulukla tutarlı bir şekilde en yüksek performansı sürdürmüştür. Bu yöntemlerin en dikkat çekici etkisi ise işlem sürelerindeki düşüş olmuştur; örneğin, FS1'de kNN'nin işlem süresi 11.17 saniyeden yalnızca 0.15 saniyeye inmiştir. Ayrıca, DT (0.05 saniye), AdaBoost (0.05 saniye) ve XGBoost (FS1'de 0.07 saniye, FS3 ve FS4'te 0.05 saniye) gibi modeller de oldukça hızlı çalışmıştır. Yöntem özelinde bakıldığında, FS2 ve FS3, LR, XGBoost, ETs ve RF modellerinde %85 ila %94 arasında doğruluk sağlarken işlem süresinde %30'a varan azalmalar görülmüştür; FS4 ise kNN ve LR ile %82–%90 arasında yüksek ve tutarlı doğruluk değerleri sunmuştur.

Farklı öznitelik seçimi yöntemlerinin birleşimiyle oluşturulan kombinasyonlarda, modellerin performansı daha da belirgin iyileşmeler göstermiştir. Özellikle FS5'te kNN, LR, RF ve ETS modelleri %92 ila %96.5 arasında oldukça yüksek doğruluk oranları elde ederken, XGBoost modeli 0.08 saniye ile en hızlı işlem süresini sunmuştur. FS6'da ise Ki-

kare, FKs, IG, ReliefF ve GA gibi yöntemlerin birleşimiyle çoğu model %90 ve üzerinde doğruluk sergilemiş, RF ve XGBoost gibi güçlü modellerin verimli sonuçlar verdiği gözlemlenmiştir. FS7 de benzer şekilde, genetik algoritma ve kesişim gibi ek yöntemler kullanılmadan dahi %90 üzeri doğruluk oranlarına ulaşarak, RF ve XGBoost modelleriyle en iyi performansı yansıtmıştır.

GA kombinasyona dahil edildiğinde performans açısından farklılıklar olmuş, FS8 kombinasyonunda, kNN ve AdaBoost modellerinin doğrulukları %90'ın üzerine çıkarak GA'da görülebilir başarı elde edilmiştir. Ancak bazı algoritmalar %80-%85'in altında doğruluklar sunmuş, FS9 ve FS10 kombinasyonları, özellikle XGBoost ve RF modellerinde yüksek doğrulukla birlikte bazı dalgalanmalar gözlemlenmiştir. Bununla birlikte, ReliefF ve GA'nın birleşiminden oluşan FS11 kombinasyonunda birçok modelin performansı iyileşmiş; kNN ve LR %91-%94 doğrulukta, XGBoost ve RF ise en iyi performans gösteren modeller arasında olmuştur. Bu kombinasyonlar, GA'nın doğru uygulandığında model doğruluğunu daha da artırmak için güçlü bir araç olabileceğini ancak bazen hesaplama açısından yüksek hesaplama maliyetleri getirebileceğini göstermiştir.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. Friedman testleri genel olarak farklı yöntemler arasında istatistiksel olarak anlamlı bir fark olmadığını göstermiştir (örneğin kNN: p=0.461, DT: p=0.178, LR: p=0.169, RF: p=0.136, ETs: p=0.252). Sadece AdaBoost (p=0.071) ve XGBoost (p=0.229) sınıflandırıcılarında anlamlılığa yaklaşan sonuçlar elde edilmiştir. Ancak detaylı ikili karşılaştırmalar için uygulanan Nemenyi testleri, birçok sınıflandırıcıda belirli yöntem çiftleri arasında anlamlı performans farklılıkları olduğunu ortaya koymuştur. Özellikle FS1'in diğer yöntemlerle (örneğin FS2, FS3, FS11) karşılaştırıldığında farklı sınıflandırıcılarda anlamlı üstünlük sağladığı görülmüştür. Ayrıca FS3, FS5, FS7 ve FS9 gibi yöntemler genellikle FS8 ve FS11'e kıyasla daha başarılı sonuçlar vermiştir. ETs ve RF sınıflandırıcılarında da FS11'in daha zayıf kaldığı ve FS1 ile FS3'e karşı anlamlı farklar yarattığı belirlenmiştir. AdaBoost ve XGBoost'ta FS5 ile FS8/FS11 ve FS2 ile FS9 arasında önemli farklar bulunmuştur. Bu bulgular, belirli özellik seçimi tekniklerinin belirli modellerde daha etkili olduğunu ve performansı en üst düzeye çıkarmak için her model için özel bir seçim stratejisinin gerekli olduğunu göstermektedir.

Sonuç olarak, yapılan kapsamlı analizler, öznitelik seçim yöntemlerinin sınıflandırma performansını hem doğruluk hem de işlem süresi açısından doğrudan etkilediğini açıkça ortaya koymuştur. Özellikle klasik istatistiksel yöntemler (FS1, FS2) ve bilgi temelli yöntemler (FS3, FS5, FS7), çoğu sınıflandırıcıda yüksek doğruluk (%94–100) ve düşük işlem süresi ile öne çıkarken, GA içeren yöntemlerin (özellikle FS8 ve FS11) bazı modellerde başarı sağlasa da daha yüksek işlem süreleri ve performans dalgalanmaları gösterdiği gözlemlenmiştir. Friedman testleri genellikle genel anlamlı farklara işaret etmese de Nemenyi testleri belirli yöntem çiftleri arasında anlamlı farklar bulunduğunu ortaya koymuştur. Özellikle FS3, FS5 ve FS7'nin, FS8 ve FS11 gibi GA tabanlı yöntemlere kıyasla daha tutarlı ve başarılı performans sunduğu saptanmıştır. Ayrıca, her sınıflandırıcının belirli özellik seçim yöntemleriyle daha uyumlu çalıştığı görülmüş; bu da model tasarımında "özelleştirilmiş öznitelik seçim stratejilerinin" kullanılmasının, sınıflandırma doğruluğunu ve verimliliğini optimize etmede kritik bir rol oynadığını güçlü şekilde ortaya koymuştur.

GSE53757 veri seti ile elde edilen bulgular

Böbrek kanserine ait GSE53757 veri seti, toplam 143 örnekten oluşmaktadır ve her bir örnek 54,676 öznitelik içermektedir. Bu veri seti, "ccRCC" ve "normal" olmak üzere iki ana sınıfa ayrılmıştır. Veri seti üzerinde gerçekleştirilen özellik seçimi işlemi sonucunda, başlangıçta 143 örnek ve 54,677 özellik bulunan veri setinin boyutu 143 örnek ve 54,675 özelliğe düşmüştür. Bu veri seti üzerinde de kNN algoritması kullanılarak, çeşitli öznitelik seçim yöntemlerinin model performansı üzerindeki etkileri incelenmiştir.

Yapılan analizler, FS0 kombinasyonunda modelin ortalama %86.2 doğruluk ve benzer seviyelerde F1 skoru, kesinlik ve duyarlılık değerleri ile kabul edilebilir bir performans sunduğunu göstermiştir; ancak işlem süresinin 33.69–36.17 saniye gibi oldukça yüksek değerlerde olması, özellikle büyük veri setleri için önemli bir dezavantaj oluşturmuştur. FS1 ve FS2 gibi klasik istatistiksel yöntemler, hem doğruluğu yüksek seviyede tutmuş (%72.41–%100 aralığında) hem de işlem süresini ortalama 0.4 saniyeye düşürerek önemli bir kazanım sağlamıştır. FS3–FS6 yöntemleri doğruluk açısından genellikle %79–93 aralığında stabil sonuçlar üretmiş, bazı durumlarda GA veya K gibi bileşenlerin etkisiyle işlem süresinde küçük artışlar ya da doğrulukta dalgalanmalar gözlemlenmiştir. Özellikle FS4 ve FS5'te nadiren istisnai şekilde uzun işlem süreleri kaydedilmiştir. FS7, GA içermeyen klasik yöntem kombinasyonu olarak %96.55'e varan doğruluklarla oldukça

başarılı sonuçlar sunmuş, işlem süreleri de genel olarak düşük kalmıştır. GA'nın dahil edildiği FS8–FS11 kombinasyonlarında ise doğruluk %75.86–%93.10 aralığında değişmiş ve bazı denemelerde üst düzey başarı yakalansa da genel olarak metriklerde dalgalanma gözlemlenmiştir. Bu kombinasyonların işlem süresi ortalama 0.4–0.8 saniye aralığında seyretmiştir. Tüm bulgular, öznitelik seçimiyle hem doğruluk oranlarının artırılabileceğini hem de işlem süresinin optimize edilebileceğini göstermekte; ancak GA gibi yöntemlerin daha dikkatli ve dengeli bir şekilde uygulanması gerektiğini ortaya koymaktadır. Bu performans metriklerinin ortalama değerleri Çizelge 4.8'de sunulmuştur.

Çizelge 4.8. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE53757)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,87	0,86	0,87	0,87	34,94
FS1	0,86	0,86	0,87	0,86	0.44
FS2	0,83	0,83	0.84	0,83	0,43
FS3	0,88	0,88	0,89	0,88	0,64
FS4	0,87	0,87	0,87	0,87	0,91
FS5	0,86	0,86	0,87	0,86	0,57
FS6	0,82	0,82	0,82	0,82	0,46
FS7	0,88	0,88	0,89	0,88	0,75
FS8	0,82	0,82	0,82	0,82	0,59
FS9	0,87	0,87	0,88	0,87	0,59
FS10	0,85	0,85	0,86	0,85	0,58
FS11	0,87	0,87	0,87	0,87	0,67

Diğer veri setlerinde olduğu gibi kNN sınıflandırıcısında kullanılan farklı özellik seçimi yöntemlerinin performansını değerlendirmek amacıyla yapılan bazı testler yapılmış ve Friedman testi sonuçları (Chi^2=17,9155, p=0,03616) genel olarak farklı özellik seçimi yöntemleri arasında istatistiksel olarak anlamlı bir fark olduğunu göstermiştir. Elde edilen p-değeri (0,03616), belirlenen 0,05 anlamlılık düzeyinden küçük olduğu için, bu yöntemlerden en az birinin diğerlerinden farklı bir performansa sahip olduğu sonucuna varılmıştır. Ancak, bu genel farklılığın hangi yöntem çiftleri arasında olduğunu daha detaylı incelemek için uygulanan Nemenyi post-hoc testi, matristeki p-değerlerinin tamamının 0.05 anlamlılık düzeyinin üzerinde olması nedeniyle hiçbir spesifik varyant çifti arasında istatistiksel olarak anlamlı bir fark tespit edememiştir. Bu durum, Friedman testinin genel bir etkiyi işaret etmesine karşın, Nemenyi testinin uyguladığı daha sıkı çoklu karşılaştırma düzeltmeleri nedeniyle bireysel farkların istatistiksel olarak belirgin kabul edilecek eşiği aşmadığını ortaya koymaktadır. Dolayısıyla, varyantlar arasında performans

farklılıkları mevcut olmakla birlikte, bu farklılıkların Nemenyi'nin katı kriterlerini karşılayacak büyüklükte olmadığı sonucuna varılmıştır.

Çalışmanın devamında çeşitli sınıflandırıcılar kullanılarak öznitelik seçim yöntemlerinin etkileri detaylı bir şekilde incelenmiştir. Yapılan deneyler, öznitelik seçimi yöntemlerinin çeşitli sınıflandırıcı algoritmalarının performansını ve işlem sürelerini önemli ölçüde etkilediğini açıkça ortaya koymuştur.

FS0 kombinasyonunda kNN, LR, RF ve ETs modelleri %93,1 ile %96,5 arasında yüksek doğruluk seviyelerine ulaşmış, bu da veri setinin doğrusal olarak iyi ayrılabilir olduğunu göstermiştir. Özellikle, kNN ve XGBoost'un bazı denemelerde %96,55 doğrulukla en iyi yöntem olduğu görülmüş ve en düşük doğruluğun %58,62 ile DT ve AdaBoost olduğu gözlemlenmiştir. Yine de, işlem sürelerinde büyük bir farklılık oluşmuş: ETs en hızlı - ortalama 7-9 ve LR- ortalama 50-65; AdaBoost en yavaşı - 200 saniyenin üzerinde olmuştur.

Tek özellik seçimi yaklaşımlarını uygulayan FS1, FS2, FS3 ve FS4 kombinasyonları, algoritmaların metrikleri için önemli iyileştirmelere ve işlem süresinin azaltılmasına yol açmıştır. Genel olarak, bu kombinasyonlarda ortaya çıkan doğruluklar %58,62 ile %100 arasında geniş bir aralıkta değişmiştir. Özellikle KNN, LR, RF ve ET modelleri, çoklu denemelerde %82,75 ile %96,55 arasında oldukça kararlı ve yüksek doğruluklar göstermiştir. Performans açısından RF, ET, AdaBoost ve XGBoost'un FS3 kombinasyonunda %100 doğruluk elde ettiği, DT performansının ise tüm tek özellik seçimi senaryolarında tutarsız ve zayıf olduğu gözlemlenmiştir. Bu kombinasyonlar için işlem süreleri önemli ölçüde iyileşmiş; en hızlı modeller XGBoost (0,05-1,18 sn) ve DT (0,07-0,19 sn) olmuştur. Diğer modeller de FS0'a kıyasla önemli hızlanmalar göstermiştir.

Farklı öznitelik seçimi yöntemlerinin birleşimiyle oluşturulan FS5, FS6 ve FS7 kombinasyonları, sınıflandırma algoritmalarının performansı ve işlem süreleri üzerinde belirgin etkiler yaratmıştır. Doğruluk oranları %58.62 ile %100 arasında geniş bir yelpazede dağılım göstermiştir. Özellikle FS5'te kNN ve ETs'in, FS6'da ise kNN, LR ve AdaBoost'un %100 doğruluk oranlarına ulaşması dikkat çekicidir. Genel olarak, kNN, LR ve ETs gibi modeller bu senaryolarda yüksek ve tutarlı doğruluklar sergilerken, DT'nin performansı genellikle daha değişken ve düşük kalmıştır. İşlem süreleri açısından, bu

birleşim senaryoları algoritma hızlarında önemli iyileşmeler sağlamıştır. XGBoost ve DT yine en hızlı modeller olurken (ortalama 0.05-0.20 saniye), diğer algoritmalar da FS0'a kıyasla önemli ölçüde hızlanmıştır. FS7'de GA ve K'in devre dışı bırakılması, işlem sürelerinde bir miktar artışa neden olarak bu iki yöntemin optimizasyondaki rolünü göstermiştir. Birden fazla öznitelik seçim yönteminin birleşimi, algoritma performansını büyük ölçüde optimize ederek işlem sürelerini düşürmüş ve çoğu model için yüksek doğruluk oranlarını korumasına veya artırmasına yardımcı olmuştur.

GA'nın dahil edildiği öznitelik seçimi kombinasyonları olan FS8, FS9, FS10 ve FS11 kombinasyonlarda, sınıflandırma algoritmalarının performansı ve işlem sürelerinde dikkate değer değişiklikler gözlemlenmiştir. Doğruluk oranları %65.51 ile %96.55 arasında bir aralıkta değişmekle birlikte, özellikle LR, kNN ve AdaBoost gibi modeller tutarlı ve yüksek doğruluk değerleri sergilemiştir. FS9'da LR %96.55 doğruluk oranına ulaşırken, FS8 ve FS10'da da benzer yüksek değerler görülmüştür. RF ve ETs'in doğruluğu bazı denemelerde değişkenlik gösterirken, DT ve XGBoost genellikle ortalama veya düşük doğruluk oranlarıyla performans sergilemiştir. İşlem süreleri açısından, bu senaryolar algoritmaların genel olarak hızlı çalıştığını göstermektedir. XGBoost ve DT yine en hızlı modeller arasında yer alırken (ortalama 0.05-0.71 saniye), diğer modeller de kabul edilebilir sürelerde tamamlanmıştır. GA'nın dahil edilmesi, öznitelik seçim süreçlerinin etkinliğini artırarak genel performansa olumlu katkı sağlamış ve işlem sürelerini makul seviyelerde tutmuştur.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. Yapılan istatistiksel analizler, farklı öznitelik seçim yöntemlerinin sınıflandırıcı performansları üzerindeki etkilerini değerlendirmiştir. kNN sınıflandırıcısı için Friedman testi istatistiksel olarak anlamlı bir fark olduğunu göstermiştir ($Chi^2 = 19.26$, $p = 0.023$), bu da uygulanan özellik seçim yöntemlerinin kNN performansını belirgin şekilde etkilediğini ortaya koymuştur. Ancak Nemenyi testi sonuçlarında, p-değerlerinin yüksek olması nedeniyle doğrudan anlamlı farklılıklar sınırlı kalmış, genel farkın bazı spesifik eşleşmelerden kaynaklandığı anlaşılmıştır. LR sınıflandırıcısında da anlamlı fark bulunmuştur ($Chi^2 = 18.35$, $p = 0.031$), fakat yine Nemenyi testi ikili karşılaştırmalarda belirgin farklar ortaya koymamıştır. DT, AdaBoost, ETs, RF ve XGBoost modellerinde Friedman testleri genel anlamda anlamlı farklılık göstermemiştir. Buna rağmen, RF, ETs ve XGBoost sınıflandırıcılarında FS3 ile FS11 arasındaki performans farkları istatistiksel

olarak anlamlı bulunmuştur (sırasıyla p = 0.020899, p = 0.047156, p = 0.038782). Bu sonuçlar, FS11 yönteminin özellikle FS3'e kıyasla bazı modellerde daha düşük performans sergilediğini göstermektedir. Genel olarak, tüm sınıflandırıcılarda belirgin farklılıklar olmasa da bazı yöntem çiftlerinin uyum düzeyleri modellerin başarısını anlamlı şekilde etkileyebilmektedir.

Sonuç olarak, GSE53757 veri seti üzerinde gerçekleştirilen analizler, öznitelik seçim yöntemlerinin sınıflandırma algoritmalarının doğruluk oranları ve işlem süreleri üzerinde önemli etkiler yarattığını açıkça ortaya koymuştur. Özellikle klasik ve bilgi teorik yöntemler, yüksek doğruluk (%85–100) ve düşük işlem süresi avantajı sağlayarak hem verimlilik hem de doğruluk açısından öne çıkmıştır. GA içeren yöntemler bazı modellerde yüksek doğruluklara ulaşabilse de performans dalgalanmaları ve model uyumsuzlukları gözlemlenmiştir. Friedman testi sonuçları, kNN ve LR algoritmalarında uygulanan öznitelik seçim yöntemlerinin performans üzerinde istatistiksel olarak anlamlı bir fark yarattığını gösterirken, Nemenyi testi sadece belirli çiftler (özellikle FS3 vs FS11) arasında bu farkların anlamlı olduğunu doğrulamıştır. Diğer yandan, XGBoost, DT ve RF gibi modellerde işlem süreleri açısından büyük avantaj sağlanırken, bazı öznitelik seçim kombinasyonlarında doğruluklar düşük kalmıştır. Tüm bu bulgular, öznitelik seçimi sürecinde hem modelin doğasına hem de veri setinin yapısına uygun stratejiler belirlemenin sınıflandırma başarısını doğrudan etkilediğini göstermektedir.

GSE6919_U95v2 veri seti ile elde edilen bulgular

Prostat kanserine ait GSE6919_U95v2 veri seti toplam 124 örnekten oluşmaktadır ve her bir örnek 12,626 öznitelik içermektedir. Bu veri seti, "primary_prostate_tumor" ve "normal" olmak üzere iki ana sınıfa ayrılmıştır. Veri seti üzerinde yapılan özellik seçimi işlemi sonrası, veri setinin başlangıç boyutu 124 örnek ve 12,627 özellik iken, bu boyut 124 örnek ve 12,625 özellik olarak güncellenmiştir. Bu veri seti üzerinde de kNN sınıflandırıcısıyla yapılan özellik seçimi analizlerinde çeşitli senaryolar incelenmiştir.

FS0 kombinasyonunda doğruluk ortalama %77,7'lik bir başarı oranına ulaşmış; ancak %64’e kadar düşük oranlar ve 3,25 saniyeye ulaşan işlem süreleri, yüksek boyutlu verilerde genelleme sorunu olduğunu ve hesaplama maliyetinin yüksek olduğunu ortaya koymuştur. Klasik istatistiksel yöntemler olan FS1 ve FS2, işlem süresini belirgin biçimde azaltarak (0.13–0.23 saniye), hesaplama açısından avantaj sunmuş olsa da, doğruluk

oranlarında ortalama %62.8 ve %64.8 ile FS0'a kıyasla düşüş yaşanmıştır. FS3 yöntemi de düşük doğruluklar (%44) nedeniyle zayıf bir tablo çizerken, FS4 %84'e kadar çıkan doğruluklarla dikkat çekmiş ve hem yüksek başarı hem de işlem verimliliği sunmuştur. Bu da ReliefF'in, tekil öznitelik seçim yöntemleri arasında en dengeli ve etkili seçeneklerden biri olduğunu göstermektedir.

Bileşik özellik seçimi yöntemleri göz önüne alındığında, FS5'ten FS7'ye işlem süreleri düşük seviyelerde olsa da (örneğin, 0,12-0,16 saniye), doğrulukta dalgalanmalar görülmüştür. FS5 %48-84, FS6 %56-72 ve FS7 %52-80 çıktıları sunmuş, ancak genellenebilirlik açısından tutarsız sonuçlar üretmiştir. GA tabanlı FS8-FS11 kombinasyonları için işlem süreleri de yüksek verimlilik göstermesine rağmen, doğruluğun %32'ye düştüğü durumlarda model güvenilirliği sorgulanmıştır. Özellikle FS8'deki oldukça geniş performans aralığı ve FS11'deki çeşitlilik, GA'nın diğer yöntemlerle birleştirildiğinde her zaman başarılı olmadığını göstermektedir. Bu sonuçlar, kNN algoritmasının tek bilgi teorik tekniklerle (özellikle ReliefF) veya dikkatlice planlanmış kombinasyonlarla birleştirildiğinde yüksek doğruluk ve işlem süresi dengesi açısından en iyi sonuçları verdiğini göstermektedir; ancak GA gibi yöntemler dikkatli optimizasyon yapılmadan kullanıldığında genelleme yeteneğini azaltabilir. Bu performans ölçütlerinin ortalama değerleri Tablo 4.9'da gösterilmektedir.

Çizelge 4.9. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE6919_U95v2)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,78	0,77	0,80	0,78	1,50
FS1	0,63	0,63	0,65	0,63	0,15
FS2	0,65	0,64	0,68	0,65	0,15
FS3	0,64	0,64	0,69	0,64	0,14
FS4	0,74	0,73	0,75	0,74	0,14
FS5	0,67	0,67	0,69	0,67	0,14
FS6	0,66	0,65	0,68	0,66	0,14
FS7	0,67	0,67	0,69	0,67	0,15
FS8	0,64	0,64	0,67	0,64	0,14
FS9	0,66	0,65	0,69	0,66	0,15
FS10	0,65	0,65	0,67	0,65	0,15
FS11	0,66	0,65	0,68	0,66	0,15

kNN sınıflandırıcısında kullanılan farklı özellik seçimi yöntemlerinin performansını değerlendirmek amacıyla yapılan analizlerde, istatistiksel testler önemli bulgular ortaya

koymuştur. Gerçekleştirilen Friedman testi sonuçları, p-değeri = 0.8058 ve istatistik = 5.3177 olarak bulunmuştur. Bu durum, Friedman testine göre özellik seçimi yöntemleri arasında istatistiksel olarak anlamlı bir fark olmadığını göstermektedir. Elde edilen Nemenyi p-değerleri matrisine göre, 0.05'ten küçük p-değerleri olan çiftler arasında istatistiksel olarak anlamlı bir fark olduğu kabul edilmiş ve Çizelge 4.10'da detaylı sunulmuştur.

Çizelge 4.10. Nemenyi Testi ile Performans Farklılıkları (GSE6919_U95v2)

Yöntem Çifti	P-Değeri
FS0 vs FS7	0,0472

Ek olarak, çeşitli özellik seçimi tekniklerinin sınıflandırma algoritmalarının işlem süreleri ve performansı üzerindeki etkileri çeşitli kombinasyonlar üzerinde incelenmiştir.

FS0 kombinasyonunda %67,2'lik ortalama doğrulukla kNN, FS0 kombinasyonunda makul bir performans göstermiş, buna karşın ET'ler %68,8 ortalamayla en iyi performansa sahip olmuştur. LR, RF, AdaBoost ve XGBoost modelleri de %60 ile %76 arasında değişen doğruluklarla ortalama %64 ile %66,4 arasında bir seyir izlemiştir. Buna karşılık, DT modeli, %36 ile %68 arasında büyük dalgalanmalar göstererek ortalama %56,8 doğrulukla en düşük ve en tutarsız performansı sunmuştur. İşlem süreleri açısından ise algoritma bazında büyük farklılıklar gözlemlenmiştir: ETs ortalama 1,05 saniye ile en hızlı sınıflandırıcı olurken, diğer sınıflandırıcılar biraz daha yüksek farkla onu takip etmiştir. En yavaş algoritmalar ise ortalama 7,74 saniye ile LR ve ortalama 14,85 saniye ile AdaBoost olmuştur.

Tekil öznitelik seçimi yöntemlerinin uygulandığı FS1, FS2, FS3 ve FS4 kombinasyonları, algoritmaların performans metriklerinde belirgin değişikliklere ve işlem sürelerinde kayda değer düşüşlere neden olmuştur. Genel olarak, bu kombinasyonlarda doğruluk oranları %44 ile %88 arasında geniş bir yelpazede seyretmiştir. İşlem süreleri açısından XGBoost ve DT modelleri ortalama 0,05-0,06 saniye ile en hızlı performansı göstermiş, kNN ve LR ortalama 0,14-0,16 saniye ve AdaBoost ortalama 0,26-0,27 saniye göstermiştir. Doğruluk performansı açısından kNN ve RF FS1'de %88'e, RF FS2'de %88'e, kNN, Ets, RF ve AdaBoost FS3'te %84'e ve AdaBoost FS4'te %88'e varan yüksek doğruluk oranlarına ulaşarak öne çıkmıştır. Ancak DT modeli özellikle FS1'de (%44) tüm tek özellik seçimi

senaryolarında diğer algoritmalara kıyasla daha değişken ve genellikle daha düşük doğruluklar sergilemeye devam etmiştir.

Farklı öznitelik seçimi yöntemlerinin kombinasyonlarıyla oluşturulan FS5, FS6 ve FS7 senaryolarında, XGBoost ve DT gibi modeller en hızlı işlem sürelerini sunarken, genel olarak tüm algoritmalar öznitelik seçimi yapılmayan FS0 kombinasyonuna kıyasla önemli ölçüde hızlanmıştır; özellikle FS6'da işlem sürelerinin tutarlı bir şekilde düşük kaldığı gözlemlenmiştir. Doğruluk performansına bakıldığında, FS7 kombinasyonu, GA dışındaki yöntemlerin birleşimiyle, LR ve ETs'in %96 gibi olağanüstü doğruluk oranlarına ulaşmasına olanak tanımış ve RF da %88 ile oldukça yüksek değerler elde etmiştir. Bu durum, belirli öznitelik seçimi yöntemlerinin birleşik kullanımının, işlem süresi verimliliğini korurken sınıflandırma doğruluğunu önemli ölçüde artırabileceğini göstermektedir. Ancak, FS5 ve FS6 senaryolarında, tekli öznitelik seçimi yöntemlerine kıyasla tüm algoritmalarda tutarlı bir performans artışı sağlanamadığı ve hatta bazı durumlarda daha düşük doğrulukların elde edildiği gözlemlenmiştir.

GA'nın dahil edildiği öznitelik seçimi kombinasyonlarında sınıflandırma algoritmalarının performansı ve işlem sürelerinde dikkate değer değişiklikler ortaya koymuştur. İşlem süreleri açısından bu senaryolar, XGBoost ve DT modellerinin en hızlı olduğunu gösterirken, GA'nın dahil edilmesi öznitelik seçim süreçlerinin etkinliğini artırarak genel performansa olumlu katkı sağlamış ve işlem sürelerini makul seviyelerde tutmuştur. Doğruluk açısından, LR, kNN ve AdaBoost FS8'de yüksek doğruluk göstermiş, ancak diğer kombinasyonlarda algoritmalardaorta ile düşük düzeyde performans çeşitliliği görülmüştür. Bazı modeller için, özellikle kNN, LR ve ET'ler bazı denemelerde %80 ve üzeri doğruluklara sahip olsa da, daha sınırlı seviyelerde kalmıştır. Bu durum, GA'nın dahil edildiği kombinasyonların işlem sürelerinde tutarlılık sağlarken, doğruluk performansında modelden modele ve kombinasyondan kombinasyona önemli farklılıklar gösterdiğini, karmaşık veri setlerinde doğru öznitelik seçimi kombinasyonunun kritik önem taşıdığını ve evrimsel algoritmaların performans optimizasyonunda güçlü bir araç olabileceğini ancak her zaman tek başına yeterli olmayabileceğini vurgulamaktadır.

Sınıflandırma algoritmalarının farklı öznitelik seçimi yöntemleri altındaki performanslarını karşılaştırmak amacıyla uygulanan Friedman testi sonuçlarına göre, tüm sınıflandırıcılar (kNN: p=0.80, DT: p=0.60, LR: p=0.42, RF: p=0.61, ETs: p=0.28, AdaBoost: p=0.16,

XGBoost: p=0.33) için genel anlamda istatistiksel olarak anlamlı bir fark tespit edilmemiştir. Bu da öznitelik seçim yöntemlerinin genel model performansı üzerinde güçlü ve tutarlı bir etki göstermediğini işaret etmektedir. Ancak, Nemenyi post-hoc testleri ile yapılan ikili karşılaştırmalar bazı modellerde dikkat çekici farklılıklar ortaya koymuştur.

Özellikle RF ve ETs sınıflandırıcılarında, belirli öznitelik seçim yöntemi çiftleri arasında istatistiksel olarak anlamlı performans farkları bulunmuştur. RF için örneğin FS1 vs FS2 (p=0.0028), FS4 vs FS9 (p=0.0013); ETs için FS1 vs FS2 (p=0.0005), FS2 vs FS3 (p=0.0025) gibi sonuçlar anlamlı farklılıklara işaret etmektedir. Diğer yandan, kNN sınıflandırıcısında genel fark anlamlı bulunmasa da FS0 ile FS7 arasındaki p=0.047156 değeri, anlamlılık sınırına yakın bir farklılık olduğunu göstermektedir. AdaBoost ve XGBoost algoritmalarında ise hem genel hem de ikili testlerde anlamlı fark tespit edilmemiştir.

GSE6919_U95v2 prostat kanseri veri seti üzerinde gerçekleştirilen kapsamlı analizler, özellik seçimi yöntemlerinin sınıflandırma algoritmalarının başarısı üzerinde kritik bir rol oynadığını ortaya koymuştur. Özellikle yüksek boyutlu biyomedikal veri setlerinde, öznitelik seçim stratejileri sayesinde hem sınıflandırma doğruluğu artırılabilmekte hem de işlem süresi anlamlı ölçüde optimize edilebilmektedir. ReliefF (FS4) gibi bilgi temelli tekil yöntemler, düşük işlem süresi ve yüksek doğruluk sağlayarak en dengeli performansı sergilemiştir. Kombine yöntemler ise dikkatlice yapılandırıldığında (örneğin FS7), bazı modellerde %96 gibi üst düzey doğruluklara ulaşılmasını sağlamış, bu da doğru yöntem kombinasyonunun model başarısını ciddi şekilde iyileştirebileceğini göstermiştir. Buna karşın, GA içeren bazı kombinasyonlar (FS8–FS11), her zaman istikrarlı sonuçlar üretmemiş ve doğruluk değerlerinde dalgalanmalar gözlemlenmiştir.

İstatistiksel test sonuçları da bu bulguları desteklemektedir. Friedman testi genel olarak özellik seçim yöntemleri arasında anlamlı farklar bulunmadığını ortaya koysa da, Nemenyi post-hoc testi bazı öznitelik seçim çifti arasında anlamlı farklar olduğunu göstermiştir (örneğin, RF ve ETs modellerinde FS1 vs FS2, FS4 vs FS9 gibi). Bu durum, özellik seçimi yöntemlerinin etkisinin model türüne ve kullanılan yöntem kombinasyonlarına göre değiştiğini, yani tek bir ideal yöntemin olmadığını, ancak uygun kombinasyonlarla her model için optimal sonuçlar elde edilebileceğini göstermektedir.

GSE19804 veri seti ile elde edilen bulgular

Akciğer kanserine ait GSE19804 veri seti toplam 114 örnekten oluşmaktadır ve her bir örnek 54,676 öznitelik içermektedir. Bu veri seti, "tumoral" ve "normal" olmak üzere iki ana sınıfa ayrılmıştır. Veri seti üzerinde yapılan özellik seçimi işlemiyle, başlangıçta 114 örnek ve 54,677 özellik bulunan veri seti, 114 örnek ve 54,675 özellik olacak şekilde küçülmüştür. Bu veri seti üzerinde de kNN sınıflandırıcısıyla yapılan özellik seçimi analizlerinde çeşitli senaryolar incelenmiştir.

Herhangi bir öznitelik seçimi yapılmadan (FS0) elde edilen doğruluk oranları %65,22 ile %91,30 arasında değişmekle birlikte, bu kombinasyon bazı denemelerde dalgalı sonuçlar üretmiş ve 4,4 ila 6,5 saniye gibi daha uzun işlem süreleriyle çalışmıştır. Buna karşın, FS1, FS2, FS3 ve FS4 gibi tekil özellik seçim yöntemleri, doğruluk oranlarını birçok denemede %100'e kadar yükselterek modelin sınıflandırma başarısını anlamlı biçimde artırmıştır. Bu yöntemlerin işlem süreleri de 0,12–0,17 saniye aralığında kalarak ciddi bir hesaplama verimliliği sunmuştur. Özellikle FS4, yüksek doğruluk ve kısa süre kombinasyonuyla öne çıkmış, yalnızca bir denemede beklenmedik uzunlukta süre kaydedilmiştir.

Kombine özellik seçimi yöntemleri (FS5–FS7), tekil yöntemlerle benzer biçimde hem yüksek doğruluklar (%90–%100) hem de düşük işlem süreleri (0,12–0,20 saniye) sağlamıştır. Bu sonuçlar, iyi yapılandırılmış kombinasyonların hem performans hem de verimlilik açısından avantaj sunduğunu göstermektedir. GA'nın dahil edildiği FS8–FS11 kombinasyonlarda da genel olarak yüksek doğruluk oranları gözlemlenmiştir; özellikle FS9 kombinasyonu, üç ayrı denemede tam (%100) doğrulukla dikkat çekmiştir. Tüm GA tabanlı kombinasyonlar, işlem süresi açısından da oldukça hızlı kalmıştır (0,13–0,15 saniye). Sonuç olarak, kNN algoritmasının performansı, özellik seçimi stratejisinin kalitesine doğrudan bağlıdır ve doğru yöntem seçimi ile yüksek boyutlu veri setlerinde etkili ve verimli modeller oluşturulabilir. Bu performans metriklerinin ortalama değerleri Çizelge 4.11'de sunulmuştur.

Çizelge 4.11. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE19804)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,82	0,82	0,86	0,82	4,70
FS1	0,96	0,96	0,96	0,96	0,15
FS2	0,94	0,94	0,94	0,94	0,15
FS3	0,94	0,94	0,95	0,94	0,15
FS4	0,94	0,94	0,94	0,94	0,35
FS5	0,92	0,92	0,92	0,92	0,16
FS6	0,91	0,91	0,92	0,91	0,15
FS7	0,94	0,94	0,94	0,94	0,39
FS8	0,90	0,90	0,91	0,90	0,15
FS9	0,93	0,93	0,94	0,93	0,15
FS10	0,91	0,91	0,92	0,91	0,15
FS11	0,90	0,90	0,91	0,90	0,16

kNN sınıflandırıcısında kullanılan farklı özellik seçimi yöntemlerinin performansını değerlendirmek amacıyla yapılan analizlerde, istatistiksel testler önemli bulgular ortaya koymuştur. Friedman testi sonuçları, kNN yöntemi için öznitelik seçimi yöntemlerinin performansı üzerinde genel olarak istatistiksel açıdan anlamlı bir fark olmadığını göstermiştir (Chi^2=9.34, p=0.406). Nemenyi post-hoc testi sonuçları ise bazı yöntem çiftleri arasında belirgin performans farklılıklarının olabileceğini göstermiştir. Yapılan ikili karşılaştırmalar sonucunda, Çizelge 4.12'deki çiftler arasında anlamlı performans farklılıkları tespit edilmiştir.

Çizelge 4.12. Nemenyi Testi ile Performans Farklılıkları (GSE19804)

Yöntem Çifti	P-Değeri
FS0 vs FS1	0,0017
FS0 vs FS3	0,0053
FS0 vs FS5	0,0041
FS0 vs FS7	0,0067
FS0 vs FS9	0,0019

Çalışmanın devamında, yalnızca kNN algoritmasıyla sınırlı kalmayarak, farklı sınıflandırıcılar kullanılarak öznitelik seçim yöntemlerinin etkileri kapsamlı şekilde değerlendirilmiştir.

FS0 kombinasyonunda, herhangi bir özellik seçimi yöntemi uygulanmadan, LR, RF, ETs ve Ada Boost algoritmaları tutarlı bir şekilde %100'lük mükemmel doğruluk, F1 skoru,

kesinlik ve hassasiyet değerleri sergilemiştir. kNN %73,91-%86,96, DT %73,91-%100 ve XGBoost %91,30-%100 arasında değişen yüksek doğruluklar elde etmiştir. İşlem süreleri açısından, ETs ve RF oldukça hızlıyken, Ada Boost ve LR en yavaş kalmıştır.

Tekil öznitelik seçim yöntemlerinin uygulandığı kombinasyonlarda, algoritmaların performans metriklerinde belirgin iyileşmeler ve işlem sürelerinde kayda değer düşüşler sağlamıştır. Genel olarak, doğruluk oranları %44 ile %100 arasında geniş bir yelpazede seyretmiş, birçok algoritmada %100 doğruluklara ulaşılmıştır. FS1'de özellikle kNN, LR, RF, ETs ve Ada Boost, FS2'de kNN, LR, ETs ve XGBoost, FS3'te kNN, LR, ETs ve XGBoost benzer başarılar sergilemiştir. FS4'te ise aynı şekilde kNN, LR, RF, ETs ve XGBoost yüksek performanslarıyla öne çıkmıştır. İşlem süreleri açısından, tüm bu kombinasyonlarda XGBoost ve DT en hızlı algoritmalar olmuş, FS0'a göre tüm sürelerde belirgin optimizasyon sağlanmıştır. Ancak, DT modeli, tüm tekil özellik seçim senaryolarında, özellikle FS1'deki %44'lük düşüşle, diğer algoritmalara kıyasla daha değişken ve genellikle düşük doğruluklar sergilemeye devam etmiştir.

Farklı öznitelik seçimi yöntemlerinin kombinasyonlarıyla oluşturulan kombinasyonlar, algoritma performansı ve işlem sürelerinde dikkate değer değişiklikler sunmuştur. İşlem süreleri açısından, XGBoost ve DT modelleri yine en hızlı performansı sunarken (genellikle 0.039-0.33 saniye ve 0.050-0.064 saniye), tüm algoritmalar öznitelik seçimi yapılmayan FS0 kombinasyonuna kıyasla önemli ölçüde hızlanmıştır. İşleme süreleri, özellikle FS6 kombinasyonunda düşük kalmış, doğruluk açısından ise kNN ve LR %100 başarı oranına ulaşmıştır. FS5 kombinasyonunda ise kNN, LR, RF, ETs, AdaBoost ve XGBoost gibi bir grup algoritma çoğu denemede %100 doğruluğa ulaşmıştır. FS7 kombinasyonunda da aynı şekilde kNN, DT, LR, ET'ler, AdaBoost ve XGBoost bazı denemelerde %100 doğruluğa ulaşırken, RF yaklaşık %95,65'te biraz daha düşük bir değere ulaşmıştır.

Özellik seçimi kombinasyonuna GA eklendiğinde elde edilen sonuçlar, sınıflandırıcıların doğruluğu ve işlem süresinde de farklılıklar ortaya koymuştur. İşlem sürelerine gelince, XGBoost ve DT en hızlı modeller olarak öne çıkmıştır. Doğruluk performansına ilişkin olarak, bazı denemelerde kNN ve Ada Boost %95,65'a ulaşmış olmasına rağmen, diğer modellerde düşük ve değişen performans farkları bulunmuştur. Öte yandan, FS9'daki sınıflandırıcıların performansı oldukça iyi olmuş, birkaç algoritma için doğruluk neredeyse

%100'e ulaşmıştır. FS10'da modeller genellikle iyi performans göstermiş ve çoğu zaman yüksek ve benzer sonuçlar vermiştir. Öte yandan, FS11'de, genel olarak yüksek değerlere sahip değişken sonuçlar görülmüştür. LR bazı deneylerde tam (%100) başarım elde ederken, DT ve ET'lerin performansında bir miktar düşüş yaşanmıştır.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. Friedman testi genel olarak kNN, DT, LR, RF, AdaBoost ve XGBoost sınıflandırıcılarında istatistiksel olarak anlamlı bir fark tespit etmese de yalnızca ETs sınıflandırıcısında istatistiksel olarak anlamlı bir farklılık bulunmuştur ($Chi^2 = 18.83$, $p = 0.027$). Bu durum, öznitelik seçimi yöntemlerinin genellikle sınıflandırıcıların performansları üzerinde güçlü ve tutarlı etkiler yaratmadığını, ancak bazı özel senaryolarda anlamlı farklar doğurabileceğini göstermektedir.

Nemenyi post-hoc testi sonuçları ise daha ayrıntılı farklılıkları ortaya koymuştur. kNN için özellikle FS0 ile FS3, FS5, FS6 ve FS7 arasında; LR için FS8'in birçok yöntemle; RF için yine FS8'in FS0, FS1, FS3, FS7 gibi senaryolarla; ETs için FS7 ile FS8 arasında, AdaBoost için FS7 ile FS8 arasında; XGBoost içinse FS8'in FS0, FS1 ve FS7 ile olan karşılaştırmalarında istatistiksel olarak anlamlı farklar gözlemlenmiştir.

Genel olarak, GSE19804 veri kümesinde gerçekleştirilen analizler, özellik seçiminin sınıflandırma performansı üzerindeki etkisinin açıkça gözlemlenebilir olduğunu ortaya koymuştur. Özellikle kNN sınıflandırıcısı hem doğruluk hem de işlem süresinde ciddi kazanımlar sağlandığı görülmüştür. Tekil öznitelik seçim yöntemleri, işlem süresini önemli ölçüde azaltırken doğruluk oranlarını %100'e kadar çıkartarak oldukça güçlü sonuçlar sunmuştur. Kombine yöntemler de hem verimlilik hem de sınıflandırma başarısı açısından benzer şekilde yüksek performans göstermiştir. GA tabanlı kombinasyonlar ise bazı kombinasyonlarda tutarsızlıklar sergilese de (özellikle FS8), genellikle yüksek doğruluk oranları ve hızlı işlem süreleriyle dikkat çekmiştir; en iyi sonuçlar FS9 ve FS10 gibi kombinasyonlarda elde edilmiştir.

Diğer yandan, farklı sınıflandırıcılarla yapılan analizler, bazı algoritmaların (özellikle LR, RF, ETs, AdaBoost) belirli öznitelik seçimi stratejileriyle yüzde yüz doğruluklara ulaşabileceğini göstermiştir. Ancak özellikle DT gibi modellerin bazı senaryolarda tutarsız ve düşük doğruluklar sergilediği görülmüştür. Friedman testi genel olarak birçok

sınıflandırıcı için öznitelik seçiminin performansa etkisinin istatistiksel olarak anlamlı olmadığını söylese de Nemenyi post-hoc testi, ikili karşılaştırmalarda özellikle FS8 gibi bazı yöntemlerin modellerin performansını anlamlı şekilde etkileyebildiğini ortaya koymuştur.

GSE63270 veri seti ile elde edilen bulgular

Lösemi kanserine ait GSE63270 veri seti, toplam 101 örnekten oluşmaktadır ve her bir örnek 54.676 öznitelik içermektedir. Bu veri seti, "AML" ve "normal" olmak üzere iki ana sınıfa ayrılmıştır. Veri setinin boyutu, özellik seçimi yapılarak 101 örnek ve 54,675 özelliğe kadar düşmüştür. Veri seti üzerinde diğer veri setlerinde olduğu gibi kNN algoritması kullanılarak, çeşitli öznitelik seçim yöntemlerinin model performansı üzerindeki etkileri incelenmiştir.

FS0 kombinasyonunda doğruluk oranları %95,24 ile %100 arasında değişmiş, bu da yüksek doğruluğa rağmen 17–27 saniyeyi bulan işlem süreleriyle önemli bir zaman maliyeti oluşturmuştur. Buna karşın, tekil öznitelik seçimi yöntemleri olan FS1–FS4, hem doğruluk oranlarında benzer şekilde %95,24–%100 aralığında üstün sonuçlar sunmuş hem de işlem sürelerini 0,32–0,54 saniyeye kadar düşürerek ciddi bir verimlilik kazanımı sağlamıştır. Özellikle FS4 (ReliefF), yüksek doğruluğun yanı sıra genel olarak kısa işlem süreleriyle dikkat çekmiş, yalnızca tek bir denemede süre anomalisi gözlenmiştir.
Kombine öznitelik seçim yöntemleri olan FS5–FS7 de benzer başarıyı korumuş, birçok denemede tam (%100) doğruluk değerlerine ulaşılmış ve işlem süreleri genellikle 0,32–0,50 saniye aralığında seyretmiştir. Ancak bazı kombinasyonlarda, özellikle FS7'de tekil bir örnekte işlem süresi 2 saniyenin üzerine çıkmıştır. GA ile desteklenen kombinasyonlar ise genel olarak yüksek doğruluk değerleri üretmiş, özellikle FS10 kombinasyonu birçok denemede yüzde yüz doğrulukla öne çıkmıştır. Yine de FS8 ve FS11 gibi bazı kombinasyonlarda doğruluk %85,71'e kadar düşerek daha düşük performans sergilemiştir. Tüm GA tabanlı yöntemlerde işlem sürelerinin 0,33–0,57 saniye gibi oldukça hızlı bir aralıkta kalması, bu stratejilerin zaman açısından da verimli olduğunu göstermektedir. Bu performans metriklerinin ortalama değerleri Çizelge 4.13’de sunulmuştur.

Çizelge 4.13. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE63270)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,98	0,98	0,98	0,98	18,48
FS1	0,95	0,95	0,95	0,95	0,38
FS2	0,99	0,99	0,99	0,99	0,35
FS3	1.00	1.00	1.00	1.00	0,36
FS4	0,98	0,98	0,98	0,98	0,49
FS5	0,99	0,99	0,99	0,99	0,35
FS6	0,91	0,91	0,93	0,91	0,34
FS7	0,98	0,98	0,98	0,98	0,4
FS8	0,92	0,92	0,93	0,92	0,35
FS9	0,91	0,91	0,92	0,91	0,36
FS10	0,95	0,95	0,96	0,95	0,38
FS11	0,94	0,94	0,94	0,94	0,36

kNN sınıflandırıcısında kullanılan farklı özellik seçimi yöntemlerinin performansını değerlendirmek amacıyla yapılan analizlerde, istatistiksel testler önemli bulgular ortaya koymuştur. Friedman testi sonuçlarına göre (p-değeri = 0.6535, istatistik = 6.8423), farklı özellik seçimi yöntemlerinin performansları arasında istatistiksel olarak anlamlı bir fark bulunamamıştır. Friedman testi genel olarak anlamlı bir fark bulmasa da Nemenyi testi sonuçlarına göre anlamlı fark bulunan FS kombinasyon çiftleri Çizelge 4.14'te gösterilmiştir.

Çizelge 4.14. Nemenyi Testi ile Performans Farklılıkları (GSE63270)

Yöntem Çifti	P-Değeri
FS1 vs FS6	0,0428
FS3 vs FS6	0,0053
FS5 vs FS6	0,0151
FS3 vs FS8	0,0121
FS3 vs FS11	0,0121
FS5 vs FS8	0,0317
FS5 vs FS11	0,0317
FS6 vs FS10	0,0258

Çalışmanın ilerleyen kısmında, farklı özellik seçimi stratejilerinin sınıflandırma algoritmalarının performansı ve işlem süreleri üzerindeki etkilerini anlamak amacıyla yapılan kapsamlı analizler değerlendirilmiştir.

FS0 kombinasyonunda LR, RF, ETs, AdaBoost ve XGBoost algoritmaları, XGBoost teknikleri başlangıçta çok iyi bir performans sergilemiştir. kNN de genel olarak %100 doğruluk gösterirken, DT algoritması %85,71 ile %95,24 arasındaki doğruluk oranlarıyla daha çeşitli bir performans sunmuştur. İşlem süreleri açısından algoritmalar arasında önemli farklılıklar gözlenmiştir. AdaBoost ve LR 2,45 saniye ile 65,13 saniye arasında değişen ortalama sürelerle en yavaş algoritmalar olurken, ETs 4,26 saniye ile 6,06 saniye arasında değişen ortalama sürelerle en hızlı algoritmalardan biri olmuştur.

Tekil özellik seçimi yöntemlerinin uygulandığı kombinasyonlarda, genel olarak, doğruluk oranları %44 ile %100 arasında geniş bir yelpazede seyretmiş, birçok algoritmada %100 doğruluklara ulaşılmıştır. Bu kombinasyonlarda, sınıflandırma algoritmaları, denemelerin çoğunda %100'e varan doğruluk, F1 skoru, kesinlik ve hassasiyet değerleriyle oldukça yüksek ve tutarlı performans sergilemiştir. Özellikle XGBoost ve DT modelleri, tüm bu tekil özellik seçimi senaryolarında 0.035 ile 0.13 saniye arasında değişen sürelerle en hızlı algoritmalardan olmuş, FS0 kombinasyonuna kıyasla tüm işlem sürelerinde belirgin optimizasyon sağlanmıştır. Diğer yandan, LR, kNN ve AdaBoost da 0.048 saniye ile 1.98 saniye arasında değişen makul işlem süreleriyle oldukça hızlı sonuçlar vermiştir. RF ve ETs ise 2.72 saniye ile 7.44 saniye arasında değişen sürelerle diğerlerine göre daha uzun sürse de, yine FS0'a kıyasla önemli bir hızlanma göstermiştir.

Farklı özellik seçimi yöntemlerinin kombinasyonlarının uygulandığı kombinasyonlar, sınıflandırma algoritmalarının hem performansını hem de işlem sürelerini önemli ölçüde etkilemiştir. İşlem süreleri açısından, XGBoost (genellikle 0.035-0.452 saniye) ve DT (genellikle 0.050-0.131 saniye) gibi algoritmalar tüm bu senaryolarda en hızlı performansı sunarken, genel olarak tüm algoritmalar özellik seçimi yapılmayan FS0 kombinasyonuna kıyasla belirgin bir hızlanma göstermiştir. Özellikle FS6'da işlem sürelerinin tutarlı bir şekilde düşük kaldığı dikkat çekmiştir. Performans metrikleri açısından ise, FS5 kombinasyonunda kNN, LR, RF, ETs, AdaBoost ve XGBoost gibi birçok algoritma denemelerin çoğunda %100'lük mükemmel doğruluk, F1 skoru, kesinlik ve hassasiyet değerlerine ulaşarak oldukça üstün bir başarı sergilemiştir. FS6'da DT, ETs, AdaBoost ve XGBoost benzer şekilde %100'lük doğruluk oranlarına ulaşırken, kNN, LR ve RF %90.48 ile %95.24 arasında değişen sonuçlar göstermiştir. FS7'de ise algoritmalar çoğunlukla %100 doğruluk sergileyerek olağanüstü bir performans göstermiştir.

GA'yı içeren özellik seçimi kombinasyonları için sınıflandırma algoritmalarının işlem süreleri ve performansında dikkate değer farklılıklar gözlemlenmiştir. İşlem süreleri açısından, her deneydeki en hızlı modeller XGBoost (0,035-0,259 sn) ve DT (0,061-0,121 sn) olmuş; GA'nın eklenmesi bu işlemi hızlandırmış ve işlem sürelerini yönetilebilir hale getirmiştir. FS8 kombinasyonundaki algoritmaların çoğu, %57,14-%90,48 arasında değişen doğruluk oranları sergilemiştir. kNN, LR ve XGBoost gibi algoritmaların çoğu FS9, FS10 ve FS11 kombinasyonlarında iyi performans göstererek %100 doğruluğa ulaşırken, diğer modeller genellikle %90'ın üzerinde başarı oranları göstermiştir.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. kNN sınıflandırıcısı, Friedman testi sonucuna göre farklı özellik seçimi yöntemleri arasında istatistiksel olarak anlamlı bir fark göstermiştir ($Chi^2 = 17.04$, $p = 0.048$). Bu, kNN algoritmasının farklı öznitelik setlerine karşı istikrarsız performans sergileyebileceğini ve öznitelik seçiminin bu modelde kritik bir rol oynadığını ortaya koymuştur. Nemenyi testinde de özellikle FS0 ile FS8 ($p = 0.016$) arasında anlamlı farklılık tespit edilmiştir.

Diğer sınıflandırıcılar olan DT, LR, RF, ETs, AdaBoost ve XGBoost için Friedman testi sonuçlarında genel olarak anlamlı bir fark bulunmamıştır. Ancak, bu modeller için uygulanan Nemenyi post-hoc testi, bazı öznitelik seçimi yöntemleri arasında ikili anlamlı farklılıkların bulunduğunu göstermiştir. Özellikle FS8 yöntemi, pek çok sınıflandırıcıyla yapılan karşılaştırmalarda öne çıkmıştır. FS8'in DT ile FS1, FS5 ve FS6'ya göre; LR ile FS0 ve FS5'e göre; RF ile FS0 ve FS5'e göre; ETs ile FS0 ve FS5'e göre; AdaBoost ile FS5 ve FS6'ya göre; XGBoost ile FS0 ve FS5'e göre istatistiksel olarak anlamlı farklılıklar oluşturduğu görülmüştür. Bu bulgular, FS8 yönteminin modellerin performansı üzerinde belirgin etkiler yarattığını, pozitif veya negatif yönde önemli sonuçlar doğurabileceğini açıkça ortaya koymaktadır.

GSE63270 veri kümesindeki analiz sonuçları, özellik seçiminin kNN ve diğer sınıflandırıcıların performansına doğruluk ve hesaplama süresi açısından önemli etkilere neden olduğunu ve bunun da performans ölçütlerinin farklı değerlerine yol açtığını göstermiştir. Özellik seçilmeden elde edilen yüksek doğruluk oranları etkileyici olsa da uzun işlem süresi bu yöntemi uygulanamaz hale getirmiştir. Tekil ve kombine özellik seçimi teknikleri (FS1–FS7), doğruluğu garanti altına alırken işlem sürelerini önemli

ölçüde azaltmış ve etkili ve verimli modeller geliştirme olasılığını göstermiştir. GA içeren kombinasyonlar genel olarak başarılı performanslar üretirken, FS8 kombinasyonu belirli sınıflandırıcılarda düşük performans göstererek öne çıkmıştır. Friedman testi yalnızca kNN algoritmasında önemli değişiklikler göstermiş, Nemenyi testi ise FS8'in bazı sınıflandırıcılarda önemli farklılıklar yarattığını doğrulamıştır. Bu sonuçlar, özellikle yüksek boyutlu veri kümelerinde uygulanan özellik seçimi stratejisinin hem model performansı hem de hesaplama verimliliği için önemli olduğunu ikna edici bir şekilde göstermektedir.

GSE21510 veri seti ile elde edilen bulgular

Kolorektal kanserine ait GSE21510 veri seti, toplam 147 örnekten oluşmakta ve her bir örnek 54.676 öznitelik içermektedir. Bu veri seti, "tumoral_LCM", "normal_homogenized" ve "tumoral_homogenized" olmak üzere üç ana sınıfa ayrılmıştır. Veri setindeki sınıf dengesizliğini gidermek amacıyla SMOTE yöntemi uygulanmıştır. SMOTE, azınlık sınıfların örneklerini sentetik olarak artırarak sınıf dağılımını dengelemeye yardımcı olmuştur. Bu işlem, sınıflandırıcıların tüm sınıflardan eşit derecede öğrenmesini sağlayarak modelin genelleme yeteneğini artırmayı hedeflemiştir. SMOTE uygulaması öncesinde ve sonrasında sınıf dağılımları kontrol edilmiş ve dengeli bir veri seti elde edilmiştir. Bu dengeli veri seti üzerinde de diğer veri setlerinde olduğu gibi özellik seçimi analizleri ve sınıflandırma işlemleri gerçekleştirilmiştir. SMOTE sayesinde 312 örnek ve 54.677 özellik olarak artmış, ardından özellik seçimi ile 312 örnek ve 54.675 özellik olarak güncellenmiştir.

Özellik seçimi yapılmadan kullanılan FS0 kombinasyonunda doğruluk %95.24 ile %100 arasında yüksek seviyede seyretmiş olsa da, işlem süreleri 152 ila 161 saniye arasında oldukça yüksektir. Buna karşın, FS1–FS4 gibi tekil öznitelik seçim yöntemleri, doğruluk oranlarını benzer seviyelerde korurken (çoğu zaman %98.41 ve üzeri), işlem sürelerini 0.54–1.54 saniye aralığına kadar düşürerek büyük bir verimlilik sağlamıştır. ReliefF (FS4) yöntemi genel olarak başarılı sonuçlar sunarken, tek bir denemede süre anomalisi gözlenmiştir. Kombine öznitelik seçim yöntemleri de doğruluk oranlarında genellikle %95'in üzerinde kalmayı başarırken, işlem süreleri çoğunlukla 1 saniyeden düşük kalmış ve yine hızlı bir performans sergilemiştir. Ancak FS6'da doğruluk %66.67'ye kadar düşmüştür. GA tabanlı kombinasyonlar da benzer biçimde başarılı performanslar sergilemiş, özellikle FS9 ve FS10 birçok denemede %96.83 ve üzeri doğruluklar

üretmiştir. Ancak FS8 ve FS11 gibi bazı kombinasyonlarda doğruluk oranları %74.60'a kadar gerileyerek daha düşük başarılar kaydedilmiştir. Bu performans metriklerinin ortalama değerleri Çizelge 4.15'te sunulmuştur.

Çizelge 4.15. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE21510)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,96	0,96	0,96	0,96	157,82
FS1	0,95	0,95	0,95	0,95	0,69
FS2	0,99	0,99	0,99	0,99	0,63
FS3	0,99	0,99	0,99	0,99	0,79
FS4	0,96	0,96	0,96	0,96	1,26
FS5	0,98	0,98	0,98	0,98	0,73
FS6	0,77	0,77	0,79	0,77	0,78
FS7	1.00	1.00	1.00	1.00	0,90
FS8	0,82	0,82	0,83	0,82	0,84
FS9	0,96	0,96	0,96	0,96	0,68
FS10	0,95	0,95	0,96	0,95	0,75
FS11	0,81	0,81	0,82	0,81	0,65

Model performanslarını istatistiksel olarak karşılaştırmak amacıyla Friedman ve Nemenyi testleri uygulanmış, Friedman testi sonucunda p-değeri 0,37659 olarak bulunmuştur. Bu durum, uygulanan özellik seçimi yöntemlerinin kNN sınıflandırıcısının performansları arasında istatistiksel olarak anlamlı bir fark olmadığını göstermektedir. Friedman testinin aksine, Nemenyi çoklu karşılaştırma testi sonuçları, bazı özellik seçimi yöntem çiftleri arasında anlamlı farklılıklar olduğunu ortaya koymuştur. Çizelge 4.16'da, p-değeri 0,05'ten küçük olan ve dolayısıyla istatistiksel olarak anlamlı fark gösteren senaryo çiftleri sunulmuştur.

Çizelge 4.16. Nemenyi Testi ile Performans Farklılıkları (GSE21510)

Yöntem Çifti	P-Değeri
FS2 vs FS3	3,3e-5
FS2 vs FS5	3,3e-5
FS2 vs FS9	1,4e-5
FS3 vs FS8	7,2e-5
FS3 vs FS11	3,6e-7
FS5 vs FS8	7,2e-5
FS5 vs FS11	3,6e-7
FS8 vs FS9	3,3e-5
FS8 vs FS10	0,0096
FS9 vs FS11	1,4e-7
FS10 vs FS11	0,0002

FS0 kombinasyonunda çeşitli sınıflandırıcıların performansları değerlendirilmiştir. LR, RF, ETs ve AdaBoost gibi modellerin bazı denemelerde mükemmel %100 doğruluk sergilemiştir. Özellikle ETs bu yüksek doğruluğu 11-13 saniye gibi nispeten hızlı işlem süreleriyle elde etmiştir. Doğruluk kNN'de %93,65 ile %98,41 arasında değişirken, işlem süreleri 154-162 saniyede seyretmiştir. DT ve XGBoost da genel olarak yüksek doğruluk oranına ulaşmış, ancak bazı durumlarda küçük dalgalanmalar olmuştur. LR ve AdaBoost için her yinelemede elde edilen %100 doğruluk, bu modellerin verilerde herhangi bir özellik seçimi olmadan aslında çok başarılı olabileceğini işaret etmiştir.

Tekil özellik seçimi yöntemlerinin uygulandığı FS1'den FS4'e kadar olan kombinasyonlarda, algoritmaların performans metriklerinde belirgin iyileşmeler ve işlem sürelerinde kayda değer düşüşler sağlanmıştır. Genel olarak, doğruluk oranları geniş bir yelpazede seyretmiş ve birçok algoritmada yüksek doğruluklara ulaşılmıştır. Ki-kare (FS1) kullanımıyla tüm sınıflandırıcıların işlem sürelerinde, özellik seçimi uygulanmayan FS0 kombinasyonuna kıyasla önemli azalmalar görülmüş; örneğin kNN 2.17 saniyeden 0.55 saniyeye düşerken, doğruluk oranları %90.47 ile %98.41 arasında korunmuştur. FS2 ve FS3 kombinasyonları ise çok daha etkileyici sonuçlar vererek, kNN, LR, RF, ETs, AdaBoost ve XGBoost gibi birçok sınıflandırıcının birden fazla denemede mükemmel %100 doğruluk oranlarına ulaşmasını sağlamıştır. Bu kombinasyonlarda işlem süreleri de oldukça düşük olmuştur; XGBoost 0.08-0.11 saniye gibi çok hızlı sürelerde sonuç vermiştir. FS4 kombinasyonu da genel olarak yüksek doğruluk oranları elde etmiş, çoğu model için %100'e yakın doğruluklar sağlamıştır.

Farklı özellik seçimi yöntemlerinin kombinasyonlarının uygulandığı kombinasyonlarda ise FS5 kombinasyonu önceki tekil özellik seçimi senaryolarına kıyasla daha dalgalı performanslar sunsa da, kNN, RF, ETs, AdaBoost ve XGBoost gibi modeller yüksek doğruluk oranlarını korumuştur. Özellikle kNN %93.65 ile %100, RF ve AdaBoost da %92.06 ile %100 arasında doğruluklar sergilemiştir. İşlem süreleri çoğunlukla verimli sonuçlar göstermiş; XGBoost 0,08-0,34 saniye arasında değişmiştir. Buna karşın, FS6 kombinasyonunda performansta düşüşler yaşanmış, çoğu modelin doğruluğu azalmış (RF %63.49'a kadar düşüş göstermiştir), işlem süreleri genel olarak daha tutarsız hale gelmiştir. Tersine, FS7 kombinasyonu, GA ve K yöntemlerinin devre dışı bırakılmasıyla olağanüstü başarıya ulaşmıştır. Bu kombinasyonda kNN, LR, RF, ETs ve AdaBoost gibi birçok model %100 doğruluk oranına ulaşırken, işlem süreleri de oldukça verimli kalmıştır.

GA tabanlı özellik seçimi kombinasyonları değerlendirildiğinde, sınıflandırma algoritmalarının doğruluk ve işlem süreleri açısından farklı performanslar sergilediği gözlemlenmiştir. FS8 kombinasyonu, bazı modellerde (örneğin RF) düşük doğruluk (%61.90) göstererek zayıf performans sergilerken, FS9 ve FS10 kombinasyonları genel olarak yüksek doğruluk oranları ve hızlı işlem süreleri ile öne çıkmıştır. Özellikle FS10'da kNN, LR ve AdaBoost gibi modellerde %98.41'e varan doğruluk elde edilmiştir; bu durum IG ve GA kombinasyonunun etkili bir seçim olduğunu göstermektedir. Buna karşın, FS11 kombinasyonu genel olarak daha düşük doğruluk oranları sunmuş, kNN, DT ve RF gibi algoritmalarda performans düşüşleri yaşanmıştır. Bu, her GA tabanlı kombinasyonun aynı derecede etkili olmadığını ve özellikle ReliefF ile GA'nın birlikte kullanıldığında her zaman ideal sonuçlar vermediğini ortaya koymaktadır.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. Özellikle kNN sınıflandırıcısı için uygulanan Friedman testi sonucunda p-değerinin 0.011 olarak bulunması, bu modelin performansında öznitelik seçimi stratejilerine bağlı olarak istatistiksel olarak anlamlı farklar oluştuğunu göstermektedir. Nemenyi testi, bu farkların kaynağını daha ayrıntılı olarak belirlemiş ve özellikle FS2 yönteminin birçok kombinasyona kıyasla anlamlı düzeyde üstün performans sunduğunu ortaya koymuştur (örneğin FS2 ile FS3, FS5 ve FS11 arasında $p<0.001$). Benzer şekilde, ETs algoritması da 0.0082 p-değeri ile öznitelik seçiminin performans üzerinde anlamlı bir etkisi olduğunu gösteren ikinci model olmuş, FS2'nin bu modelde de diğer yöntemlerle karşılaştırıldığında istatistiksel olarak daha başarılı sonuçlar verdiği belirlenmiştir.

Diğer sınıflandırıcılar olan DT, LR, RF, AdaBoost ve XGBoost için Friedman testleri genel anlamda anlamlı farklar göstermemiş olsa da Nemenyi testi bu modellerde de FS2'nin öne çıktığını ve FS0, FS3, FS5, FS8, FS9 ve FS11 gibi yöntemlerle karşılaştırıldığında anlamlı performans farkları yarattığını ortaya koymuştur. Bu durum, FS2'nin sadece tek bir modele özgü olmadığını, aksine birçok farklı sınıflandırıcıda istikrarlı ve güçlü sonuçlar verdiğini göstermektedir. Sonuç olarak, genel analiz bulguları, öznitelik seçimi stratejilerinin sınıflandırma başarısına doğrudan etki ettiğini, özellikle FS2 gibi yöntemlerin doğru kullanıldığında modellerin doğruluk, verimlilik ve istikrar açısından önemli ölçüde geliştirilebileceğini açıkça ortaya koymuştur.

GSE21510 veri setinin kapsamlı incelemesi, özellik seçimi tekniklerinin sınıflandırma performansını önemli ölçüde ve doğrudan etkilediğini ortaya koymuştur. Özellik seçimi olmadan yüksek doğruluk oranlarının elde edilmesine rağmen aşırı uzun işlem sürelerinin modelin verimliliğini önemli ölçüde düşürdüğü görülmüştür. Buna karşılık olarak, tekil ve uygun kombinasyonlu özellik seçimi yöntemleri, doğruluğu koruyarak işlem sürelerini büyük oranda düşürmüş ve verimliliği artırmıştır. Özellikle FS2 yöntemi hem doğruluk hem de hız açısından neredeyse tüm sınıflandırıcılarda istikrarlı şekilde üstün performans göstermiştir. FS7 gibi bazı kombinasyonlar da GA ve K gibi yöntemler dışlandığında oldukça başarılı sonuçlar sunmuştur.

GA tabanlı yöntemlerin etkisi ise karışıktır: FS9 ve FS10 gibi iyi yapılandırılmış kombinasyonlar yüksek performans sergilerken, FS8 ve FS11 gibi bazıları yetersiz doğruluk oranları ile dikkat çekmiştir. Friedman ve Nemenyi test sonuçları, FS2'nin istatistiksel olarak anlamlı şekilde öne çıktığını doğrulamış ve bu yöntemin sınıflandırma modelleri için güvenilir ve etkili bir öznitelik seçimi stratejisi olduğunu ortaya koymuştur. Sonuç olarak, doğru öznitelik seçimi yöntemi ile hem model doğruluğu artırılabilir hem de işlem süreleri makul seviyelere çekilerek genel sınıflandırma başarısı optimize edilebilir.

GSE50161 veri seti ile elde edilen bulgular

GSE50161 beyin kanseri veri kümesi toplamda 130 örnekten ve 54.676 öznitelikten oluşmaktadır. Bu veri kümesini oluşturan beş sınıf "ependymoma", "glioblastoma", "medulloblastoma", "normal" ve "pilocytic_astrocytoma"dır. Diğer veri kümesinde olduğu gibi, bu veri kümesinde de SMOTE yöntemi kullanılmış, ardından kNN algoritması farklı özellik seçimi tekniklerinin modelin performansını nasıl etkilediğini analiz etmek için kullanılmıştır. Veri setinde yapılan SMOTE ve özellik seçimi işlemi sonucunda veri seti boyutu 230 örnek ve 54,675 özellik olarak güncellenmiştir.

Özellik seçimi yapılmadan kullanılan FS0 kombinasyonunda model, %91.3–%100 arasında yüksek doğruluklar elde etmiş ancak 43.76–54.10 saniyelik uzun işlem süreleriyle dikkat çekmiştir. Tekil özellik seçimi yöntemleri olan FS1–FS4, doğruluk oranlarını bazı durumlarda düşürse de işlem süresinde büyük verimlilik sağlamıştır. Özellikle FS2 ve FS3, %95'in üzerine çıkan doğruluklar ve 0.13–0.18 saniye arası kısa işlem süreleriyle öne çıkarken, FS1 ve FS4 daha düşük doğruluklara ulaşmışlardır.

Kombine yöntemlerin değerlendirildiği FS5–FS7 kombinasyonlarda, FS5 orta seviye bir başarı sergilemiş, FS6 ise en düşük doğruluklara (%32.6–%52.2) sahip olarak öne çıkmıştır. GA'nın çıkarıldığı FS7 kombinasyonu ise yüksek doğruluk (%84.8–%100) ve hızlı işlem süreleriyle dikkat çekmiştir. GA içeren FS8–FS11 kombinasyonları genel olarak düşük doğruluk oranları üretmiş; FS8 ve FS9'da doğruluk %60'ın altında kalmış, FS11 ise en düşük başarıyı göstermiştir. Bu durum, GA'nın bazı yöntemlerle birleştiğinde etkisiz hatta zararlı olabileceğini ortaya koymaktadır. İşlem sürelerinin kısa kalmasına rağmen doğruluk düşüşü, bu kombinasyonların genel başarıya katkı sağlamadığını göstermektedir. Bu performans metriklerinin ortalama değerleri Çizelge 4.17'de sunulmuştur.

Çizelge 4.17. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE50161)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,96	0,96	0,96	0,96	45,69
FS1	0,81	0,84	0,81	0,82	0,15
FS2	0,86	0,88	0,86	0,87	0,15
FS3	0,94	0,94	0,94	0,94	0,15
FS4	0,80	0,82	0,80	0,80	0,7
FS5	0,76	0,77	0,76	0,76	0,18
FS6	0,40	0,44	0,40	0,40	0,16
FS7	0,93	0,93	0,93	0,93	0,55
FS8	0,55	0,56	0,57	0,54	0,17
FS9	0,55	0,56	0,58	0,55	0,17
FS10	0,74	0,74	0,74	0,73	0,15
FS11	0,49	0,49	0,49	0,49	0,17

K-Neighbors sınıflandırıcısıyla elde edilen sonuçların istatistiksel olarak karşılaştırılması amacıyla diğer veri setlerindeki gibi Friedman ve Nemenyi testleri uygulanmıştır. Elde edilen Friedman Chi^2 testi sonuçlarına göre, p-value değeri 0,5305 ve test istatistiği 8,0362 olarak bulunmuştur. Bu durumda, p-value değerinin 0,05'ten büyük olması, uygulanan öznitelik seçimi yöntemleri arasında istatistiksel olarak anlamlı bir fark olmadığını göstermektedir. Ancak, Nemenyi testinde göre anlamlı farklılıklar oluşmaktadır. Çizelge 4.18'de, p-değeri 0,05'ten küçük olan ve dolayısıyla istatistiksel olarak anlamlı fark gösteren kombinasyon çiftleri sunulmuştur.

Çizelge 4.18. Nemenyi Testi ile Performans Farklılıkları (GSE50161)

Yöntem Çifti	P-Değeri
FS0 vs FS2	4,0e-6
FS0 vs FS6	2,4e-5
FS0 vs FS8	4,5e-5
FS0 vs FS11	6,6e-8
FS2 vs FS5	0,0041
FS3 vs FS2	2,0e-5
FS3 - FS6	0,0001
FS3 - FS8	0,0002
FS3 - FS11	4,4e-7
FS5 – FS6	0,0151
FS5 – FS8	0,0232
FS9 - FS2	3,3e-5
FS9 - FS6	0,0002
FS9 - FS8	0,0003
FS9 - FS11	7,6e-7

Farklı özellik seçimi stratejilerinin, sınıflandırma algoritmalarının hem performansı hem de işlem süreleri üzerindeki etkilerini anlamak amacıyla kapsamlı bir analiz yapılmış ve analizde, FS0'dan FS11'e kadar çeşitli kombinasyonlar değerlendirilmiştir.

Genel olarak, FS0'da sınıflandırıcılar iyi performans göstermiş ve birçok tekrarlamada en iyi doğruluk puanları LR, RF ve ET sınıflandırıcılarında %100'e kadar ulaşmıştır. kNN ve XGBoost sınıflandırıcıları %91 ile %100 arasında doğruluklarla yüksek sonuçlar elde etmiş, DT ve AdaBoost sınıflandırıcılarının doğrulukları ise %71,73 ile %100 arasında değişmiştir. İşlem süreleri açısından ise önemli farklılıklar gözlemlenmiştir: ETs (yaklaşık 5.9-6.6 saniye) ve RF (yaklaşık 13.7-18.1 saniye) oldukça hızlı çalışırken, kNN (yaklaşık 42.3-44.3 saniye) ve XGBoost (yaklaşık 98-111 saniye) orta düzeyde, LR (yaklaşık 221-236 saniye) ve AdaBoost (yaklaşık 165-180 saniye) ise nispeten daha uzun sürelerde sonuç vermiştir.

Tekil özellik seçimi yöntemlerinin uygulandığı kombinasyonlarda, algoritmaların işlem sürelerinde çarpıcı düşüşler gözlemlenmiştir, çoğu model artık 0.03 ila 2.5 saniye gibi oldukça hızlı sürelerde tamamlanmaktadır. Ancak, bu hızlanma genellikle doğruluk performansında bir düşüşle sonuçlanmıştır. FS1 kombinasyonunda Ki-kare uygulandığında, FS0'da sıklıkla %100 doğruluk sergileyen modellerin doğrulukları %60.86

ile %86.95 aralığına gerilemiş, DT ise en belirgin düşüşü yaşayarak bazı denemelerde %54.34 gibi düşük doğruluklara inmiştir. FS2 kombinasyonunda, sınıflandırıcıların performansları da FS1'e göre genel bir iyileştirme görülmüş, ancak hesaplama süresi düşük kalmıştır. Özellikle kNN (%82,60-%93,47) ve ETs (%82,60-%93,47) modelleri iyi performans göstermiştir. FS3 kombinasyonunda sınıflandırıcılar doğruluk açısından yüksek ve işlem süresi açısından düşük seyretmiştir. Özellikle kNN, RF, ETs ve XGBoost modelleri, çoğu denemede %90'ın, hatta sıkça %95'in üzerinde ve hatta bazı durumlarda %100'e varan doğruluklar sergilemiştir. Özellikle AdaBoost modeli, doğruluk metriklerinde oldukça düşük sonuçlar (%34.78 - %60.86) sergileyerek bu senaryoda en zayıf performans gösteren sınıflandırıcı olmuştur.

Farklı özellik seçimi yöntemlerinin kombinasyonlarının uygulandığı kombinasyonlar, sınıflandırma algoritmalarının hem performansını hem de işlem sürelerini önemli ölçüde etkilemiştir. FS5 senaryosunda elde edilen sonuçlar genel olarak karmaşık bir tablo çizmektedir. kNN, RF ve ETs modelleri en iyi performansı sergilerken, AdaBoost tüm öznitelik seçimi kombinasyonlarında olduğu gibi bu kombinasyonda da en düşük doğruluk oranlarıyla dikkat çekmiştir. FS6 kombinasyonunda, modellerin performansında ciddi bir düşüş gözlemlenmiştir. Çoğu sınıflandırıcının doğruluk oranları %30 ve %50 arasındaki düşük değerlere gerilemiştir. FS7 kombinasyonunda ise sınıflandırma performansında çarpıcı bir iyileşme gözlemlenmiş; özellikle kNN, LR, RF ve ETs modelleri %97.82 ile %100 arasında değişen olağanüstü doğruluk oranlarına ulaşmıştır. Bu sonuçlar, bu dört öznitelik seçimi yönteminin birlikte son derece etkili olduğunu kanıtlamıştır.

Sınıflandırma algoritmalarının doğruluk ve işlem sürelerinde GA gibi özellik seçimli kombinasyonları incelendiğinde önemli farklılıklar tespit edilmiştir. İşlem sürelerine bakıldığında genel olarak en hızlı modeller XGBoost (0,035-0,259 sn) ve DT (0,061-0,121 sn) olmuş ve GA'nın dahil edilmesi bu süreçlerin verimliliğini artırarak süreleri makul seviyelerde tutmuştur. Doğruluk performansına bakıldığında, FS8 kombinasyonunda algoritmaların performansında belirgin bir düşüş yaşanmış, çoğu model %30 ile %60 arasında değişken doğruluk oranları göstermiştir. FS9 kombinasyonunda, genel olarak orta düzeyde, hatta bazı durumlarda düşük performansa işaret eden sonuçlar elde edilmiş, modellerin doğruluk oranları genellikle %30 ile %65 arasında değişmiştir. Bununla birlikte, özellikle bazı modeller FS10 kombinasyonunda çok yüksek sınıflandırma başarımları ortaya koymuş ve %60'ların üzerinde tutarlı doğruluk oranlarına ulaşmıştır.

Buna karşılık, FS11 kombinasyonu çoğu sınıflandırma modeli için genel olarak zayıf bir performans göstermiş; doğruluk oranları %30 ile %55 arasında değişmiştir.

Sınıflandırma işleminin ardından, Friedman ve Nemenyi post-hoc testleri uygulanmıştır. Yedi farklı sınıflandırma algoritması için yapılan Friedman testi sonuçlarına göre, yalnızca kNN algoritmasında ($p=0.1658$) anlamlılığa yakın bir fark gözlemlenmiş ancak bu da %5'lik anlamlılık düzeyinin altında kalmamıştır. Diğer algoritmalarda (DT, LR, RF, ETs, AdaBoost, XGBoost) p-değerleri sırasıyla 0.2744, 0.8464, 0.3255, 0.5969, 0.1828 ve 0.3549 olarak belirlenmiş ve bu değerler, farklı öznitelik seçimi yöntemlerinin sınıflandırıcı performansları üzerinde genel olarak istatistiksel olarak anlamlı farklar yaratmadığını göstermiştir.

Buna rağmen, Nemenyi post-hoc testleri detaylı karşılaştırmalar sunarak bazı öznitelik seçimi kombinasyonlarının belirgin şekilde daha iyi ya da daha kötü performans sergilediğini ortaya koymuştur. Özellikle FS2'nin neredeyse tüm algoritmalar için FS0, FS4, FS6 ve FS11 gibi yöntemlere kıyasla anlamlı düzeyde daha başarılı olduğu tespit edilmiştir. Ayrıca FS0'ın genel olarak düşük performans sergilediği ve FS2 dışındaki bazı kombinasyonlarla da anlamlı farklara sahip olduğu görülmüştür.

GSE50161 veri kümesinde yapılan kapsamlı analizlerde, özellik seçimi yöntemlerinin sınıflandırma performansı ve hesaplama süresi üzerinde önemli bir etkiye sahip olduğu net bir şekilde görülmüştür. FS0 kombinasyonunda yüksek doğruluklar olmasına rağmen, uzun bir işlem süresi olmuştur. Tekil öznitelik seçimi yöntemlerinden özellikle FS2 ve FS3, hem doğruluk hem de işlem süresi açısından dengeli ve başarılı sonuçlar sunmuştur. Kombine yöntemlerden FS7, GA'nın çıkarılmasıyla belirgin bir performans artışı sağlarken, GA içeren kombinasyonlar (özellikle FS8, FS9 ve FS11) genel olarak düşük doğruluklar sergilemiş, bu da GA'nın bazı yöntemlerle birlikte kullanımında dikkatli olunması gerektiğini göstermiştir.

İstatistiksel analizler doğrultusunda, Friedman testi genel olarak öznitelik seçimi yöntemleri arasında anlamlı fark olmadığını öne sürse de Nemenyi post-hoc testi FS2 gibi belirli yöntemlerin diğerlerine kıyasla istatistiksel olarak anlamlı şekilde daha başarılı olduğunu açıkça ortaya koymuştur. FS2 yöntemi hem farklı sınıflandırıcılar hem de farklı veri setleri üzerinde tutarlı ve güçlü performans sergileyerek en öne çıkan öznitelik seçimi

yöntemi olmuştur. Bu bağlamda, FS2 gibi bilgi temelli yöntemlerin, genetik algoritma gibi hesaplama açısından ağır tekniklerle kıyaslandığında daha pratik ve etkili bir çözüm sunduğu söylenebilir.

GSE45827 veri seti ile elde edilen bulgular

Meme kanserine ait GSE45827 veri seti, toplam 151 örnekten oluşmaktadır ve her bir örnek tıpkı 54.676 öznitelik içermektedir. Bu veri seti, "basal", "HER", "cell_line", "normal", "luminal_A" ve "luminal_B" olmak üzere altı farklı sınıfa ayrılmıştır. Bu veri setinde de diğer veri setinde olduğu gibi SMOTE yöntemi uygulanmış, daha sonra kNN algoritması kullanılarak, çeşitli öznitelik seçim yöntemlerinin model performansı üzerindeki etkileri incelenmiştir. Veri setinde yapılan SMOTE ve özellik seçimi işlemi sonrasında veri seti, 246 örnek ve 54,675 özellik olarak güncellenmiştir.

Özellik seçimi uygulanmadığı durumda, model yüksek doğruluk (%84–%94) ve F1 skoru (%84.2–%93.9) değerlerine ulaşsa da, işlem sürelerinin 48.31–57.54 saniye arasında değişerek ciddi bir dezavantaj oluşturmuştur. Klasik öznitelik seçimi yöntemleri (FS1–FS4), doğrulukta farklılıklar gösterse de işlem sürelerinde belirgin düşüş sağlamıştır. Özellikle FS2 ve FS3 gibi yöntemler, %86'ya varan doğruluklar ve 0.14–0.20 saniyelik işlem süreleriyle hem hızlı hem de etkili sonuçlar sunmuştur. FS4, bazı denemelerde %96'ya varan doğruluklara ulaşsa da nadiren işlem süresinde anomali yaşanmıştır.

Kombine özellik seçimlerinin değerlendirildiği FS5–FS7 kombinasyonlarda, FS5 ve FS6 genel olarak düşük doğruluklar üretmiş, özellikle FS6 %30–%54 aralığıyla en zayıf performansı göstermiştir. Buna karşılık FS7, %94'e kadar ulaşan doğrulukları ve düşük işlem süreleriyle dikkat çekmiştir. GA içeren FS8–FS11 kombinasyonlarında ise genellikle %42–%70 aralığında düşük doğruluklar elde edilmiş, işlem süreleri hızlı kalsa da model başarısı anlamında yetersiz kalınmıştır. Bu performans metriklerinin ortalama değerleri Çizelge 4.19'da sunulmuştur.

Çizelge 4.19. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE45827)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,91	0,91	0,92	0,91	49,85
FS1	0,74	0,73	0,78	0,74	0,21
FS2	0,78	0,78	0,79	0,78	0,19
FS3	0,81	0,81	0,82	0,81	0,17
FS4	0,92	0,92	0,93	0,92	0,84
FS5	0,64	0,64	0,67	0,64	0,16
FS6	0,42	0,43	0,47	0,42	0,16
FS7	0,91	0,91	0,92	0,91	0,64
FS8	0,56	0,55	0,58	0,56	0,18
FS9	0,56	0,55	0,58	0,56	0,18
FS10	0,62	0,62	0,65	0,62	0,16
FS11	0,60	0,58	0,61	0,60	0,20

Özellik seçimi yöntemlerinin performanslarını karşılaştırmak amacıyla, önceki veri setlerinde olduğu gibi Friedman testi ve Nemenyi testi uygulanmıştır. Friedman testi sonuçlarına göre, elde edilen p-değeri 0,9277 ve istatistik değeri Chi^2 = 3,7393 olarak bulunmuştur. Bu, farklı özellik seçimi senaryolarının K-Neighbors sınıflandırıcısının performansı üzerinde istatistiksel olarak anlamlı bir fark yaratmadığını göstermektedir. Ancak uygulanmış Nemenyi post-hoc testine göre bazı çiftler arasında anlamlı fark olmuştur. Nemenyi test sonuçlarına göre, α=0,05 anlamlılık düzeyinde anlamlı fark bulunan yöntem çiftleri Çizelge 4.20'de sunulmuştur:

Çizelge 4.20. Nemenyi Testi ile Performans Farklılıkları (GSE45827)

Yöntem Çifti	P-Değeri
FS0 vs FS2	0,0022
FS0 vs FS4	0,0108
FS0 vs FS6	0,0002
FS0 vs FS8	0,0002
FS0 vs FS11	3,6e-7
FS1 vs FS2	0,0010
FS1 vs FS4	0,0053
FS1 vs FS6	0,0001
FS1 vs FS8	7,2e-5
FS1 vs FS11	1,2e-7
FS2 vs FS11	0,0009
FS3 vs FS6	0,0317
FS3 vs FS8	0,0258
FS3 vs FS11	0,0002
FS5 vs FS11	0,0041
FS6 vs FS11	0,0099
FS7 vs FS11	0,0085
FS8 vs FS11	0,0099
FS9 vs FS6	0,0002
FS9 vs FS8	0,0001
FS9 vs FS11	2,5e-7
FS10 vs FS11	0,0066

Çalışmanın ilerleyen kısmında, farklı özellik seçimi kombinasyonlarında sınıflandırma algoritmalarının performansı ve işlem süresi üzerindeki etkisini araştırmak için kapsamlı bir analiz yapılmış ve analizde FS0'dan FS11'e kadar farklı kombinasyonlar değerlendirilmiştir.

FS0 kombinasyonunda, RF ve ETs, sırasıyla %97 doğruluk ve hızlı işlem süreleri (18-20 saniye ve 7-8 saniye) ile öne çıkarken, kNN %91 doğrulukla yüksek performans sergilemiş ancak daha uzun işlem süreleri (47-49 saniye) kaydetmiştir. LR, %92-%100 doğrulukla dikkat çekmiş ancak 282-297 saniye ile en uzun işlem süresine sahip olmuştur. XGBoost için sınıflandırma doğruluğu %94 ve ortalama işlem süresi 134-137 saniye, DT için doğruluk %90 ve verileri işlemesi 31-36 saniye sürmüştür. Doğruluk açısından en zayıf algoritma ise AdaBoost (%60 ve 193-195 saniye) olmuştur.

Tekil özellik seçimi yöntemleri uygulandığında, algoritmaların işlem sürelerinde belirgin bir hızlanma sağlanmış ve çoğu model 0.03 ila 2.9 saniye arasında tamamlanmıştır. Ancak bu hızlanma, doğrulukta bir miktar düşüşe yol açmıştır. FS1 kombinasyonunda doğruluklar %60.86 ile %86.95 arasında değişirken, özellikle DT %54.34 ile önemli bir performans kaybı yaşamıştır. FS2 ise doğrulukları iyileştirerek işlem sürelerini düşük tutmuş ve kNN (%73.2) ile ETs (%76.6) tutarlı performans sergilemiştir. FS3 kombinasyonunda, genel olarak daha yüksek doğruluklar ve düşük işlem süreleri elde edilmiştir; ETs (%85.6) ve RF (%83.4) öne çıkmıştır. FS4 ise en yüksek doğruluk oranlarını (%95.6) elde ederken, işlem sürelerini makul seviyelerde tutmayı başarmıştır; ETs ve kNN bazı denemelerde %98 doğruluğa ulaşmıştır. AdaBoost, tüm tekil özellik seçimi kombinasyonlarında olduğu gibi en düşük performansı (%46) sergilemiştir.

Farklı özellik seçimi kombinasyonları, sınıflandırma algoritmalarının performansını ve işlem sürelerini önemli ölçüde etkilemiştir. FS5 kombinasyonunda, tüm özellik seçim yöntemlerinin bir arada kullanılması, beklenen performans artışının aksine doğruluk oranlarında belirgin bir düşüşe yol açmıştır. RF (ortalama %42.3) ve ETs (ortalama %42.2) en yüksek doğrulukları elde ederken, diğer algoritmalar %36.2 ile %40.7 arasında düşük performans sergilemiştir. FS6 kombinasyonunda doğruluk oranları daha da düşerek en düşük seviyelere inmiştir; kNN en yüksek doğruluğu gösterse de, diğer algoritmalar %42.2 ile %50.2 arasında çok düşük performans sergilemiştir. FS7 kombinasyonunda ise performans önemli ölçüde iyileşmiş, özellikle ETs ve kNN ortalama %94.8 doğrulukla en iyi sonuçları vermiş, bazı denemelerde %98'e kadar ulaşmıştır. LR (%93.4) ve RF (%91.6) de yüksek doğruluklar elde etmiştir. Ancak AdaBoost, ortalama %39.6 doğrulukla en düşük performansı sergilemeye devam etmiştir.

Özellik seçiminin çeşitli kombinasyonları, sınıflandırma algoritmalarının verimliliği ve işlem süreleri üzerinde önemli etkilere sahip olmuştur. FS5 kombinasyonu için, tüm özellik seçimi yöntemleri birleştirildiğinde beklendiği gibi performans artışı yerine doğruluk oranlarında bir performans düşüşü elde edilmiş; RF'nin (ortalama; %42,3) ve ETs'in (ortalama; %42,2) doğruluğu nispeten yüksek olmasına rağmen, algoritmaların çoğu düşük doğrulukla sonuçlanmıştır (%36,2 ile %40,7 arasında). FS6 kombinasyonu için doğruluk, kNN hariç diğer tüm algoritmalar %42,2 ile %50,2 arasında değişen oldukça düşük bir doğruluğa sahip performans göstermiştir. Bu performans FS7 kombinasyonunda artmış, ETs ve kNN, %94,8'lik ortalama doğrulukla en iyi sonuçları vermiş, bazı denemelerde ise

%98'e kadar başarı elde etmiştir. LR (%93,4) ve RF (%91,6) ile sınıflandırma da iyi doğruluklar elde etse de AdaBoost yine de ortalama %39,6 doğrulukla en kötü sonuçları vermiştir.

GA'yı içeren özellik seçimi kombinasyonlarında, sınıflandırma algoritmalarının performanslarının yanı sıra işlem süreleri de farklılıklar göstermiştir. İşlem süreleri arasında, XGBoost (0,11-0,27 sn) ve DT (0,03-0,08 sn) en hızlı modeller olmuş ve GA'nın dahil edilmesi işlem sürelerini makul seviyelerde tutmuştur. kNN (ortalama %54,7) ve XGBoost (ortalama %52,9) en yüksek doğrulukları gösterirken, AdaBoost (%43,7) FS8 kombinasyonunda en düşük performansı göstermiştir. FS9 kombinasyonunda ise kNN (ortalama %56,3) ve XGBoost (ortalama %52,9) yüksek doğruluklar göstermiştir. FS10 kombinasyonu ise önceki kombinasyonlara göre daha yüksek doğruluk oranları sunmuş; kNN (ortalama %62.6), LR (ortalama %61.7) ve XGBoost (ortalama %59.8) başarılı performanslar sergilemiş, AdaBoost (%46.5) yine en düşük ortalamayı kaydetmiştir. Son olarak, FS11 kombinasyonunda doğruluk oranları düşmüş; kNN (ortalama %46.5) ve DT (ortalama %44.5) en yüksek doğrulukları elde ederken, LR (%32.4) ve AdaBoost (%34.3) düşük sonuçlar sergilemiştir.

Sınıflandırma sonrası kullanılan Friedman ve Nemenyi post-hoc testleri uygulanmış, 7 sınıflandırıcı arasında yapılan Friedman testleri, sınıflandırıcıların performansı arasındaki farkların istatistiksel olarak anlamlı olmadığını göstermiştir. Tüm sınıflandırıcıların p-değerleri (kNN: 0.784, DT: 0.884, LR: 0.615, RF: 0.947, ETs: 0.998, AdaBoost: 0.629, XGBoost: 0.816) $\alpha = 0.05$'ten yüksek olmuştur. Ancak Nemenyi post-hoc testleri, bu genel sonuçların ötesine geçerek belirli öznitelik seçimi kombinasyonlarının bazı modellerin performansı üzerinde anlamlı etkiler yarattığını göstermiştir. Özellikle FS0 ile FS2, FS0 ile FS10 ve FS1 ile FS2 gibi çiftler arasında çok düşük p-değerleri elde edilmiştir. Bu da bazı öznitelik seçim kombinasyonlarının, genel test sonuçlarında fark edilmeyen ancak ikili karşılaştırmalarda ortaya çıkan belirgin performans farklılıkları yarattığını göstermektedir.

GSE45827 veri setine yönelik yapılan analizler, öznitelik seçiminin hem model doğruluğu hem de işlem süresi üzerinde önemli etkiler yarattığını açıkça ortaya koymuştur. Özellik seçimi yapılmadığında modeller yüksek doğruluklar sergilese de, işlem sürelerinin aşırı uzun olması nedeniyle büyük veri setlerinde bu durum önemli bir dezavantaja dönüşmektedir. Klasik öznitelik seçim yöntemleri (özellikle FS2, FS3 ve FS4) hızlı ve

yüksek doğruluk içeren sonuçlar sunmuştur. Buna karşılık, GA içeren kombinasyonlar (FS8–FS11), düşük doğruluk değerleriyle genellikle yetersiz kalmıştır. Ayrıca, tüm sınıflandırıcılar için Friedman testleri genel bir anlamlı fark göstermemiş olsa da Nemenyi testi belirli kombinasyonlar arasında istatistiksel olarak anlamlı farkların bulunduğunu ortaya koymuştur.

GSE28497 veri seti ile elde edilen bulgular

Lösemi kanserine ait GSE28497 veri seti, toplam 281 örnekten oluşmaktadır ve her bir örnek 22.284 öznitelik içermektedir. Bu veri seti, "B-CELL_ALL", "B-CELL_ALL_TCF3-PBX1"," B-CELL_ALL_HYPERDIP", "B-CELL_ALL_HYPO", "B-CELL_ALL_MLL", "B-CELL_ALL_T-ALL" ve "B-CELL_ALL_ETV6-RUNX1" olmak üzere yedi farklı sınıfa ayrılmıştır. Bu veri seti üzerinde de bir önceki analizde olduğu gibi SMOTE yöntemi uygulanmış, kNN algoritması kullanılarak, çeşitli öznitelik seçim yöntemlerinin model performansı üzerindeki etkileri incelenmiştir. SMOTE uygulaması ile 518 örnek ve 22,285 özellik olarak artırılmış, ardından özellik seçimi işlemi ile 518 örnek ve 22,283 özellik olarak güncellenmiştir.

FS0 durumunda, model %79.81 ile %87.5 arasında yüksek doğruluklara ulaşabilmiş ancak işlem süreleri 60–90 saniye arasında değişerek ciddi bir hesaplama maliyeti doğurmuştur. Buna karşılık, tekil öznitelik seçimi yöntemlerinden FS1 ve FS3 hem doğruluk açısından tatmin edici (%78.3–%83.1) hem de işlem süresi açısından oldukça verimli (0.17–0.19 saniye) sonuçlar sunmuştur. Her ne kadar IG ile yapılan seçimler %90,4'lük başarı oranıyla dikkat çekse de, FS2 daha düşük doğruluk oranı (%69.9) ile bu veri seti için zayıf bir tercih olduğunu göstermiştir. FS4 %79,8'lik iyi bir ortalama doğruluk oranıyla makul bir başarı elde etmiş ancak yine de yüksek bir işlem süresine (0,5 saniye) sahip olmuştur.

Kombine öznitelik seçim yöntemlerine bakıldığında, FS5 ve FS6 kombinasyonları %33–%36 doğruluk ve benzer F1 skorları ile tüm yöntemler arasında en düşük performansı sergilemiş, işlem süresi avantajına rağmen model başarısını düşürmüştür. Buna karşın FS7, %88.9 ortalama doğruluk ve %88.4 F1 skoru ile hem yüksek doğruluk hem de düşük işlem süresi (ortalama 0.6 saniye) sunarak en başarılı kombinasyon olmuştur. GA içeren FS8 ve FS9 kombinasyonlarında doğruluklar %32.8–%36.8 arasında kalmış ve etkisiz performans göstermiştir. FS10 ve FS11 ise doğrulukları %44.8–%46.9 aralığına çıkararak GA ile daha

uyumlu kombinasyonlar oluşturmuş ancak yine de orta seviyede kalmıştır. Bu performans metriklerinin ortalama değerleri Çizelge 4.21'de sunulmuştur.

Çizelge 4.21. Özellik Seçimi Yöntemlerinin kNN Performansı (GSE28497)

FS	Accuracy	Precision	Recall	F1-skor	Ortalama Süre (sn)
FS0	0,83	0,84	0,83	0,80	68,96
FS1	0,78	0,79	0,78	0,77	0,19
FS2	0,70	0,71	0,70	0,69	0,19
FS3	0,83	0,84	0,83	0,82	0,17
FS4	0,80	0,80	0,80	0,79	0,69
FS5	0,33	0,34	0,33	0,33	0,2
FS6	0,36	0,36	0,36	0,35	0,22
FS7	0,89	0,90	0,89	0,88	0,89
FS8	0,33	0,34	0,33	0,33	0,24
FS9	0,37	0,38	0,37	0,37	0,27
FS10	0,47	0,47	0,47	0,46	0,19
FS11	0,45	0,45	0,45	0,44	0,25

Özellik seçimi yöntemlerinin performanslarını karşılaştırmak amacıyla, önceki veri setlerinde olduğu gibi Friedman testi ve Nemenyi testi uygulanmıştır. K-Neighbors sınıflandırıcısının farklı özellik seçimi yöntemleri altındaki performansını değerlendirmek amacıyla yapılan Friedman Chi^2 testi, p-değeri 0.507294 olarak bulunmuştur. Bu durum, test edilen özellik seçimi yöntemlerinin kNN performansları arasında istatistiksel olarak anlamlı bir fark olmadığını göstermektedir. Friedman testi genel bir anlamlı fark göstermese de Nemenyi test sonuçlarına göre bazı yöntem çiftleri arasında istatistiksel olarak anlamlı farklar gözlemlenmiştir. Çizelge 4.22'de bu anlamlı farklar detaylandırılmıştır:

Çizelge 4.22. Nemenyi Testi ile Performans Farklılıkları (GSE28497)

Yöntem Çifti	P-Değeri
FS0 vs FS 6	0,0011
FS0 vs F8	1,0e-5
FS0 vs 10	3,3e-5
FS vs FS	0,0168
FS1 vs FS8	0,0003
FS1 vs FS10	0,0009
FS3 vs FS6	0,0013
FS vs FS8	1,2e-5
FS3 vs FS10	3,8e-5
FS5 vs FS8	0,0388
FS6 vs FS9	7,0e-6
FS7 vs FS8	0,0007
FS7 vs FS10	0,0017
FS8 vs FS9	2,5e-8
FS8 vs FS10	9,7e-8
FS9 vs FS11	5,0e-6
FS10 vs FS9	9,7e-8
FS11 vs FS9	5,0e-6

Diğer veri setlerindeki gibi farklı öznitelik seçimi yöntemlerinin sınıflandırma algoritmalarının performansı ve işlem süreleri üzerindeki etkileri çeşitli kombinasyonlar üzerinden incelenerek değerlendirilmiştir.

FS0 kombinasyonunda algoritmalar arasındaki farklar yalnızca performans açısından değil, aynı zamanda işlem süresi açısından da farklılıklar göstermiştir. LR'nin ortalama doğruluğu %96,15 ile en yüksek performans göstermiş, ardından sırasıyla XGBoost, RF, ETs ve DT takip etmiştir. AdaBoost (%59,61) en kötü performansı gösterirken, kNN (%82,5) orta seviyede kalmıştır. İşlem sürelerinde önemli farklılıklar kaydedilmiş; kNN, XGBoost, AdaBoost ve özellikle LR algoritmaları daha uzun işlem süreleri nedeniyle dikkat çekerken, diğer algoritmalar, özellikle ET'ler, RF ve DT, 10,3-40,2 saniye ile hızlı performans göstermiştir. Bu algoritmaların işlem süreleri ortalama 63.5-213.8 saniye olarak kaydedilmiştir.

Tekil özellik seçimi yöntemlerinin uygulandığı FS1, FS2, FS3 ve FS4 kombinasyonlarda algoritmaların performans metriklerinde belirgin değişiklikler olmuştur. Genel olarak,

doğruluk oranları %44 ile %88 arasında değişirken, işlem süreleri önemli ölçüde optimize edilmiştir. 0,05-0,09 saniyelik ortalamayla XGBoost ve DT modelleri en hızlı performansı gösterirken, diğer algoritmalar da önemli ölçüde daha hızlı sonuçlar vermiştir. Doğruluk performansı açısından özellikle dikkat çekenler FS1'de ETs (%81,42) ve XGBoost (%79,69), FS2'de ETs (%72,9), FS3'te RF (%85,87) ve LR (%83,43) ve FS4'te RF (%82,68) ve ETs (%82,89) olmuştur. Ancak, AdaBoost tüm tekil senaryolarda nispeten düşük performans sergilemeye devam ederken, DT de değişken ve genellikle düşük doğruluklar göstermiştir.

Birden fazla özellik seçimi yönteminin kombinasyonlarında algoritmaların performanslarında karışık sonuçlar elde edilmiştir. FS5 kombinasyonunda genel doğruluk oranları düşüş göstererek, birden fazla özellik seçiminin her zaman avantaj sağlamayabileceğini ortaya koymuştur; burada en yüksek doğruluklar RF (%53.79) tarafından kaydedilmiştir. FS6 kombinasyonunda, tüm algoritmaların performansı önemli ölçüde düşmüş; RF'nin en iyi performansı bile yalnızca %37,28 doğrulukta olmuştur. Bununla birlikte, LR (%88,08), ET'ler (%90,33), RF (%90,41) ve XGBoost (%91,13) gibi algoritmalar, FS7 kombinasyonu sayesinde olağanüstü yüksek doğruluklara ulaşabilmiştir. Tüm algoritmalar genel olarak FS0 kombinasyonundan çok daha hızlı olmuştur, ancak DT, işlem süreleri açısından tüm bu kombinasyonlar arasında en hızlı algoritma olmuştur.

Sınıflandırma algoritmalarının işlem süreleri ve performansı, GA'yı (FS8, FS9, FS10 ve FS11) içeren özellik seçimi kombinasyonları tarafından önemli ölçüde dikkate değer değişiklikler ortaya koymuştur. XGBoost ve DT modelleri genellikle en hızlı işlem sürelerine sahip olmuş ve GA'nın eklenmesi işlem sürelerini makul seviyelerde tutmuştur. Ancak doğruluk performansında kombinasyondan kombinasyona önemli farklılıklar gözlemlenmiştir: FS8 ve FS9 kombinasyonları genel olarak düşük doğruluk oranları sergileyerek bu kombinasyonların yetersiz kaldığını göstermiştir. Buna karşın, FS10 ve FS11 kombinasyonları, özellikle kNN (%45.06 ve %45.98), LR (%42.27 ve %45.98) ve XGBoost (%42.25 ve %43.95) için önceki bazı kombinasyonlara kıyasla önemli ölçüde daha yüksek doğruluk oranları sunmuştur.

Sınıflandırma algoritmalarının farklı öznitelik seçimi yöntemleri altındaki performanslarını karşılaştırmak amacıyla uygulanan Friedman testleri, öznitelik seçimi yöntemlerinin genel performans üzerindeki etkisinin istatistiksel olarak anlamlı bir farklılık yaratmadığını

göstermiştir. kNN (p≈0.168), DT (p≈0.186), LR (p≈0.246), RF (p≈0.410), ET (p≈0.309), AdaBoost (p≈0.599) ve XGBoost (p≈0.464) için elde edilen p-değerleri, 0.05'lik anlamlılık düzeyinin üzerinde kalmış ve genel olarak öznitelik seçiminin sınıflandırma başarısı üzerinde belirgin bir fark oluşturmadığını ortaya koymuştur. Ancak bu genel tabloya rağmen, Nemenyi post-hoc testleri daha ayrıntılı farklılıkları açığa çıkarmıştır.

Nemenyi test sonuçları, bazı özellik seçimi yöntem çiftleri arasında anlamlı performans farkları olduğunu göstermiştir. Özellikle FS2, FS4, FS6, FS8, FS9 ve FS11 gibi yöntemlerin birçok sınıflandırıcıda, FS0, FS1 ve FS3 gibi yöntemlere kıyasla daha iyi sonuçlar ürettiği dikkat çekmiştir. Örneğin, kNN için FS0 ile FS4 ve FS6 ile FS7 arasında, DT için FS2 ve FS4'ün FS0 ve FS3'e göre üstün olduğu; LR, RF, ET, AdaBoost ve XGBoost için ise FS9 ve FS11 gibi yöntemlerin bazı geleneksel yöntemlere kıyasla daha anlamlı performans artışları sağladığı gözlemlenmiştir. Bu bulgular, genel farklılık olmasa bile, belirli özellik seçimi stratejilerinin bazı modellerde performansı anlamlı şekilde etkileyebileceğini göstermektedir.

GSE28497 veri seti üzerinde yapılan analizler sonucunda, öznitelik seçiminin sınıflandırma performansı ve işlem süresi üzerinde belirgin etkileri olduğu görülmüştür. Genel olarak, Friedman testleri sınıflandırıcıların performansları arasında istatistiksel olarak anlamlı farklar olmadığını gösterse de, Nemenyi testleri bazı özellik seçimi yöntemleri arasında anlamlı farklılıklar bulunduğunu ortaya koymuştur. Özellikle FS7 kombinasyonu, hem doğruluk (%88.9) hem de F1 skoru açısından en yüksek performansı sergilemiş ve düşük işlem süresiyle en verimli yöntemlerden biri olmuştur. FS1, FS3 ve FS4 gibi klasik yöntemler de işlem süresi avantajları ve makul doğruluk değerleri ile dengeli sonuçlar sunmuştur. Buna karşılık, FS5, FS6, FS8 ve FS9 gibi kombinasyonlar hem doğruluk hem de F1 skoru açısından düşük performans göstermiştir.

Özellikle GA'nın dahil edildiği kombinasyonlar, işlem süresi açısından avantaj sağlasa da, genellikle düşük doğruluk ve F1 skorları üretmiştir. Bununla birlikte, FS10 ve FS11 gibi daha uyumlu kombinasyonlarda GA'nın katkısı kısmen olumlu olmuş ve bazı algoritmalar için önceki kombinasyonlara kıyasla daha yüksek doğruluklar elde edilmiştir. Analizin bulguları, özellik seçimi yöntemlerinin GSE28497 veri kümesinde başarıya ulaşmasında kritik rol oynadığını ortaya koymuş; uygun teknikler seçilerek işlem sürelerini büyük ölçüde azaltabileceğini ve sınıflandırma doğruluğunu artırabileceğini göstermiştir. Bu,

büyük ve çok sınıflı biyomedikal verilerde özellik seçiminin hesaplama ve modelleme başarıları için önemli bir adım olduğunu göstermiştir.

Bu çalışmada uygulanan 11 farklı öznitelik seçim modeli, 7 farklı sınıflandırma yaklaşımı ile değerlendirilmiş ve elde edilen sonuçlar Çizelge 4.23-4.26'da sunulmuştur. Çizelgeler kullanılan öznitelik seçim kombinasyonlarına (*komb.*) göre dört alt bölüme ayrılmıştır: Çizelge 4.23 (FS0-FS2), Çizelge 4.24 (FS3-FS5), Çizelge 4.25 (FS6-FS8) ve Çizelge 4.26 (FS9-FS11).

Çizelge 4.23, dört temel özellik seçim yöntemi (FS0-2) kullanılarak çeşitli veri setleri üzerinde sınıflandırma algoritmalarının test sonuçlarını detaylı bir şekilde sunmaktadır. Bu analizde, algoritmaların genel performansları ve özellik seçim yöntemlerinin etkileri incelenmiştir.

Çizelge 4.23. Özellik Seçim Yaklaşımları (FS0-3) ve Sınıflandırma Algoritmalarının Test Sonuçları

		K-Neighbors				Decision Tree				Logistic Regression				Random Forest				Extra Tree				Ada Boost				XGBoost			
veri seti	komb.	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec
GSE70947	F0	0.85	0.85	0.87	0.85	0.78	0.78	0.78	0.78	0.92	0.92	0.92	0.92	0.89	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.89	0.89	0.90	0.89	0.89	0.89	0.90	0.89
GSE14520_U133A		0,94	0,94	0,94	0,94	0,92	0,92	0,92	0,92	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,96	0,96	0,97	0,96	0,96	0,96	0,96	0,96	0,97	0,97	0,97	0,97
GSE44076		0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99
GSE53757		0.84	0.85	0.85	0.84	0.72	0.71	0.75	0.72	0.87	0.87	0.87	0.87	0.85	0.85	0.86	0.85	0.87	0.86	0.87	0.87	0.83	0.83	0.84	0.83	0.86	0.86	0.87	0.86
GSE6919_U95Av2		0.67	0.67	0.70	0.67	0.56	0.55	0.59	0.56	0.64	0.64	0.67	0.64	0.66	0.65	0.68	0.66	0.67	0.67	0.69	0.67	0.64	0.64	0.66	0.64	0.63	0.63	0.65	0.63
GSE19804		0.83	0.82	0.85	0.83	0.87	0.86	0.87	0.87	0.97	0.97	0.98	0.97	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.95
GSE63270		1.00	1.00	1.00	1.00	0.93	0.93	0.93	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
GSE21510		0.97	0.96	0.97	0.97	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.99	0.98
GSE50161		0.96	0.96	0.96	0.96	0.94	0.95	0.94	0.96	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.87	0.87	0.91	0.87	0.95	0.95	0.95	0.95
GSE45827		0.87	0.86	0.91	0.87	0.91	0.91	0.92	0.91	0.96	0.96	0.97	0.96	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.57	0.53	0.54	0.57	0.94	0.94	0.95	0.94
GSE28497		0.83	0.79	0.85	0.83	0.86	0.86	0.87	0.86	0.96	0.96	0.96	0.96	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.62	0.60	0.68	0.62	0.93	0.93	0.94	0.93
GSE70947	F1	0.83	0.83	0.84	0.83	0.81	0.81	0.81	0.81	0.84	0.84	0.85	0.84	0.86	0.86	0.87	0.86	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.84	0.84	0.84	0.84
GSE14520_U133A		0,97	0,97	0,97	0,97	0,93	0,93	0,94	0,93	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,95	0,95	0,95	0,95	0,96	0,96	0,96	0,96
GSE44076		0.99	0.99	0.99	0.99	0.97	0.97	0.98	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98
GSE53757		0.82	0.81	0.82	0.82	0.73	0.73	0.76	0.73	0.81	0.81	0.82	0.81	0.81	0.80	0.82	0.81	0.82	0.82	0.83	0.82	0.80	0.80	0.80	0.80	0.79	0.78	0.80	0.79
GSE6919_U95Av2		0.62	0.62	0.64	0.62	0.62	0.62	0.65	0.62	0.70	0.70	0.73	0.70	0.69	0.69	0.71	0.69	0.67	0.67	0.70	0.67	0.67	0.67	0.70	0.67	0.66	0.66	0.68	0.66
GSE19804		0.99	0.99	0.99	0.99	0.93	0.93	0.94	0.93	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.96
GSE63270		0.94	0.94	0.95	0.94	0.96	0.96	0.96	0.96	0.99	0.99	0.99	0.99	0.95	0.95	0.95	0.95	0.96	0.96	0.97	0.96	0.94	0.94	0.95	0.94	0.95	0.95	0.96	0.95
GSE21510		0.95	0.95	0.95	0.95	0.90	0.90	0.90	0.90	0.91	0.91	0.92	0.91	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.91	0.91	0.91	0.91	0.93	0.93	0.93	0.93
GSE50161		0.79	0.79	0.81	0.79	0.70	0.70	0.72	0.70	0.74	0.74	0.75	0.74	0.77	0.77	0.78	0.77	0.80	0.80	0.81	0.80	0.68	0.67	0.70	0.68	0.77	0.77	0.79	0.77
GSE45827		0.74	0.73	0.77	0.74	0.75	0.75	0.77	0.75	0.73	0.71	0.75	0.73	0.81	0.81	0.82	0.81	0.79	0.79	0.82	0.79	0.62	0.59	0.62	0.62	0.79	0.78	0.81	0.79
GSE28497		0.78	0.77	0.79	0.78	0.72	0.72	0.73	0.72	0.68	0.66	0.68	0.68	0.80	0.80	0.80	0.80	0.81	0.80	0.81	0.81	0.57	0.52	0.56	0.57	0.79	0.79	0.80	0.79
GSE70947	F2	0.88	0.88	0.89	0.88	0.79	0.79	0.80	0.79	0.90	0.90	0.91	0.90	0,84	0,87	0,88	0,86	0.88	0.88	0.88	0.88	0.87	0.87	0.88	0.87	0.87	0.87	0.88	0.87
GSE14520_U133A		0.96	0.96	0.96	0.96	0,94	0,94	0,94	0,94	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,96	0,95	0,96	0,96	0,96	0,96
GSE44076		1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98
GSE53757		0.84	0.85	0.85	0.84	0.79	0.78	0.80	0.79	0.83	0.83	0.83	0.83	0.82	0.82	0.83	0.82	0.85	0.85	0.86	0.85	0.84	0.84	0.85	0.84	0.82	0.82	0.84	0.82
GSE6919_U95Av2		0.70	0.70	0.74	0.70	0.62	0.61	0.64	0.62	0.71	0.71	0.75	0.71	0.67	0.67	0.72	0.67	0.70	0.69	0.73	0.70	0.64	0.63	0.68	0.64	0.59	0.58	0.63	0.59
GSE19804		0.95	0.95	0.96	0.95	0.93	0.93	0.94	0.93	0.94	0.94	0.95	0.94	0.93	0.93	0.94	0.93	0.94	0.94	0.95	0.94	0.92	0.92	0.93	0.92	0.94	0.94	0.95	0.94
GSE63270		1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
GSE21510		1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GSE50161		0.88	0.87	0.88	0.88	0.81	0.80	0.81	0.81	0.83	0.83	0.86	0.83	0.85	0.84	0.86	0.85	0.88	0.87	0.89	0.88	0.76	0.76	0.80	0.76	0.85	0.84	0.86	0.85
GSE45827		0.72	0.72	0.74	0.72	0.68	0.68	0.71	0.68	0.73	0.73	0.75	0.73	0.74	0.74	0.77	0.74	0.76	0.76	0.78	0.76	0.56	0.53	0.57	0.56	0.72	0.73	0.74	0.72
GSE28497		0.71	0.70	0.70	0.71	0.66	0.66	0.67	0.66	0.62	0.60	0.59	0.62	0.70	0.69	0.69	0.70	0.73	0.72	0.72	0.73	0.58	0.56	0.59	0.58	068	0.68	0.68	0.68

Çizelge 4.23’e genel olarak bakıldığında, farklı öznitelik seçim yöntemlerinin (FS0-FS2) ve sınıflandırma algoritmalarının (K-Nearest Neighbors, Decision Tree, Logistic Regression, Random Forest, Extra Tree, Ada Boost, XGBoost) performansları detaylı bir şekilde analiz edilmiştir. Veriler, doğruluk, F1 skoru, kesinlik ve hassasiyet gibi metrikler bazında değerlendirilmiş olup, her bir algoritmanın başarıları incelenmiştir. Çeşitli veri setleri üzerinde elde edilen sonuçlar, hangi algoritmaların daha etkili olduğunu ve hangi veri setlerinde hangi algoritmaların daha iyi performans gösterdiğini ortaya koymaktadır.

İlk olarak FS0 kombinasyonunda, LR, RF, ETs ve XGBoost algoritmalarının çoğu veri setinde oldukça yüksek performans göstermiştir. Özellikle GSE63270 ve GSE21510 veri setlerinde bu algoritmaların %100 doğruluk ve F1 skoru değerlerine ulaştığı görülmüştür. Bu durum, temel özelliklerin bile belirli veri setleri için sınıflandırma görevinde son derece yeterli olabileceğini göstermektedir. Ancak, DT ve kNN gibi algoritmalar, özellikle GSE6919_U95Av2 ve GSE53757 gibi veri kümelerinde diğer veri setlerinden daha kötü performans göstermiştir. Bu da bu algoritmaların gürültülü verilere veya karmaşık veri yapılarına karşı daha hassas olabileceğini gösteriyor.

FS1 öznitelik seçim yöntemi incelendiğinde, genel performansın FS0 ile benzer eğilimler gösterdiği ancak bazı veri setlerinde küçük farklılıklar ortaya çıktığı gözlemlenmiştir. Örneğin, GSE19804 veri setinde RF ve ETs algoritmaları FS0'a göre daha yüksek doğruluk ve F1 skoru elde etmiştir (her ikisi de %0.99). Bu, FS1'in bu veri kümesi için daha uygun bir özellik alt kümesi sağladığını göstermektedir. Yine de, FS1 daha düşük veri kümelerinde, örneğin GSE6919_U95Av2 ve genel olarak GSE50161'de performans artışı sağlamaz ve bazı durumlarda FS0'a göre daha kötü sonuçlar bile sunar (örneğin DT'den GSE50161'e). Bu gözlem, özellik seçimi yöntemlerinin performansının veriye bağlı olduğunu ve bir veri dağıtımı için en uygun yöntemin değişebileceğini vurgular.

Son olarak, FS2 öznitelik seçim kombinasyonuna bakıldığında, genel olarak performansın diğer iki temel özellik seçimi (FS0 ve FS1) ile tutarlı olduğu görülmektedir. Yine LR, RF, ETs ve XGBoost algoritmaları, özellikle GSE44076, GSE63270 ve GSE21510 gibi veri setlerinde mükemmel veya mükemmele yakın sonuçlar elde etmiştir. Bu algoritmalar bu veri kümeleri üzerinde %100 doğruluk ve F1 puanı ile üstünlüklerini korumuşlardır. Ancak, GSE6919_U95Av2 veri setinde birkaç algoritma (kNN ve XGBoost) FS0 ve FS1'den önemli ölçüde daha kötü performans göstermiş ve bu da FS2'nin tüm veri

kümeleri için en iyi çözüm olmadığını göstermiştir. Özellikle, GSE45827 ve GSE28497'de AdaBoost ve XGBoost sınıflandırıcısının düşüşleri, bu özellik seçiminin bazı veri kümelerinde modelin genelleştirilmesini daha da kötüleştireceğini göstermektedir.

Algoritmalar özelinde değerlendirme yapıldığında, LR, RF ve ETs algoritmalarının, kullanılan FS0-FS2 öznitelik seçim yöntemlerine bakılmaksızın, genel olarak tüm veri setlerinde tutarlı ve yüksek bir performans sergilediği göze çarpmaktadır. Bu algoritmalar, özellikle yüksek boyutlu gen ekspresyonu verilerinde güçlü bir sınıflandırma yeteneği göstermiştir. Örneğin, GSE14520_U133A veri setinde bu algoritmalar %0.96 ve üzeri doğruluk oranlarına ulaşmıştır.

Buna karşılık, DT algoritması genel olarak diğer algoritmalardan (örn. GSE6919_U95Av2, GSE53757) daha kötü performans göstermiştir. Bu, karar ağaçlarının bu veri kümelerinin karmaşıklığını idare edemediğini ve aşırı uyum sağlama olasılıklarının daha yüksek olduğunu gösteriyor. Benzer şekilde, kNN bazı veri kümelerinde (örneğin GSE6919_U95Av2) en azından orta düzeyde performans gösterse de özellikle GSE63270 ve GSE21510 gibi yüksek performanslı veri kümelerinde %100 doğruluk elde etmiştir. Bu da veri setinin özelliklerine bağlı olarak performansının büyük ölçüde değişebildiğini göstermektedir.

Ada Boost ve XGBoost gibi topluluk öğrenme yöntemleri ise genellikle yüksek performans göstermiştir. Özellikle XGBoost'un, LR, RF ve ETs ile birlikte birçok veri setinde en iyi veya en iyiye yakın sonuçlar elde ettiği görülmüştür. Ancak, GSE45827 ve GSE28497 veri setlerinde Ada Boost ve XGBoost'un diğer bazı algoritmalara göre düşüş yaşadığı ve özellikle FS2 öznitelik seçiminde performanslarının belirgin şekilde azaldığı dikkat çekmektedir. Bu durum, bu güçlü algoritmaların bile belirli öznitelik seçim stratejileri ve veri setlerinin özellikleri karşısında hassasiyet gösterebileceğini ortaya koymaktadır.

Çizelge 4.24, FS3, FS4 gibi tekil özellik seçim yöntemleri ve bu yöntemlerin GA ile birlikte kullanıldığı FS5 yaklaşımının sınıflandırma performansına etkisini detaylıca incelemektedir.

Çizelge 4.24. Özellik Seçim Yaklaşımları (FS3-5) ve Sınıflandırma Algoritmalarının Test Sonuçları

		K-Neighbors				Decision Tree				Logistic Regression				Random Forest				Extra Tree				Ada Boost				XGBoost			
veri seti	komb.	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec
GSE70947	F3	0.86	0.86	0.87	0.86	0.81	0.81	0.82	0.81	0.86	0.86	0.88	0.86	0.87	0.87	0.88	0.87	0.86	0.86	0.87	0.86	0.87	0.87	0.88	0.87	0.87	0.87	0.88	0.87
GSE14520_U133A		0,97	0,97	0,97	0,97	0,93	0,93	0,94	0,93	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
GSE44076		1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.97	0.97	0.98	0.97	0.98	0.98	0.99	0.98
GSE53757		0.88	0.88	0.89	0.88	0.81	0.81	0.82	0.81	0.85	0.85	0.86	0.85	0.89	0.89	0.9	0.89	0.89	0.89	0.9	0.89	0.88	0.88	0.89	0.88	0.89	0.89	0.9	0.89
GSE6919_U95Av2		0.72	0.72	0.75	0.72	0.61	0.61	0.64	0.61	0.70	0.70	0.73	0.70	0.73	0.73	0.75	0.73	0.72	0.72	0.74	0.72	0.71	0.71	0.73	0.71	0.70	0.70	0.73	0.70
GSE19804		0.97	0.97	0.97	0.97	0.89	0.89	0.90	0.89	0.97	0.97	0.97	0.97	0.94	0.94	0.95	0.94	0.97	0.97	0.97	0.97	0.93	0.93	0.93	0.93	0.94	0.94	0.95	0.94
GSE63270		1.00	1.00	1.00	1.00	0.93	0.93	0.94	0.93	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
GSE21510		1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GSE50161		0.95	0.95	0.96	0.95	0.92	0.92	0.93	0.92	0.92	0.92	0.93	0.92	0.94	0.94	0.94	0.94	0.95	0.95	0.96	0.95	0.91	0.91	0.92	0.91	0.95	0.95	0.95	0.95
GSE45827		0.83	0.83	0.85	0.83	0.78	0.78	0.80	0.78	0.82	0.82	0.84	0.82	0.83	0.82	0.84	0.83	0.85	0.85	0.86	0.85	0.60	0.54	0.54	0.60	0.82	0.82	0.84	0.82
GSE28497		0.82	0.81	0.83	0.82	0.78	0.78	0.79	0.78	0.84	0.83	0.85	0.84	0.84	0.84	0.85	0.84	0.85	0.84	0.85	0.84	0.56	0.52	0.53	0.56	0.83	0.83	0.84	0.83
GSE70947	F4	0.88	0.88	0.89	0.88	0.81	0.81	0.81	0.81	0.87	0.87	0.88	0.87	0.86	0.86	0.86	0.86	0.88	0.88	0.88	0.88	0.85	0.85	0.86	0.85	0.86	0.86	0.87	0.86
GSE14520_U133A		0,96	0,96	0,96	0,96	0,92	0,92	0,92	0,92	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95
GSE44076		1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.99	0.98	0.99	0.99	0.99	0.99
GSE53757		0.86	0.86	0.88	0.86	0.73	0.73	0.74	0.73	0.87	0.87	0.88	0.87	0.86	0.86	0.87	0.86	0.87	0.87	0.88	0.87	0.84	0.84	0.86	0.84	0.86	0.86	0.88	0.86
GSE6919_U95Av2		0.72	0.72	0.73	0.72	0.67	0.66	0.68	0.67	0.74	0.73	0.75	0.74	0.72	0.72	0.73	0.72	0.72	0.72	0.73	0.72	0.75	0.75	0.76	0.75	0.69	0.69	0.70	0.69
GSE19804		0.95	0.95	0.95	0.95	0.93	0.93	0.93	0.93	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96
GSE63270		0.99	0.99	0.99	0.99	0.97	0.97	0.97	0.97	0.94	0.94	0.94	0.94	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.95	0.95	0.96	0.95
GSE21510		0.96	0.96	0.96	0.96	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.97	0.97	0.97	0.97	0.99	0.99	0.99	0.99	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
GSE50161		0.80	0.80	0.83	0.80	0.74	0.74	0.75	0.74	0.75	0.74	0.78	0.75	0.79	0.79	0.81	0.79	0.86	0.86	0.88	0.86	0.47	0.42	0.43	0.47	0.76	0.75	0.77	0.76
GSE45827		0.94	0.94	0.95	0.94	0.89	0.89	0.90	0.89	0.90	0.90	0.92	0.90	0.92	0.92	0.93	0.92	0.95	0.95	0.96	0.95	0.45	0.34	0.34	0.45	0.91	0.91	0.92	0.91
GSE28497		0.81	0.80	0.82	0.81	0.76	0.76	0.77	0.76	0.79	0.78	0.80	0.79	0.82	0.82	0.84	0.82	0.83	0.82	0.84	0.83	0.48	0.42	0.49	0.48	0.81	0.81	0.82	0.81
GSE70947	F5	0.91	0.91	0.91	0.91	0.86	0.86	0.86	0.86	0.90	0.90	0.90	0.90	0.92	0.92	0.92	0.92	0.90	0.90	0.91	0.90	0.89	0.89	0.89	0.89	0.91	0.91	0.91	0.91
GSE14520_U133A		0.95	0.95	0.95	0.95	0,91	0,91	0,91	0,91	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,94	0,94	0,94	0,94	0,95	0,95	0,95	0,95
GSE44076		0.99	0.99	1.00	0.99	0.98	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
GSE53757		0.87	0.87	0.87	0.87	0.80	0.81	0.83	0.80	0.86	0.86	0.87	0.86	0.86	0.86	0.87	0.86	0.86	0.86	0.87	0.86	0.86	0.86	0.87	0.86	0.86	0.86	0.87	0.86
GSE6919_U95Av2		0.68	0.68	0.69	0.68	0.65	0.65	0.66	0.65	0.71	0.71	0.72	0.71	0.66	0.65	0.67	0.66	0.68	0.68	0.69	0.68	0.65	0.65	0.66	0.65	0.66	0.66	0.66	0.66
GSE19804		0.97	0.97	0.97	0.97	0.94	0.94	0.94	0.94	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.95	0.95	0.96	0.95
GSE63270		1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.96	0.96	0.97	0.96	0.95	0.95	0.96	0.95
GSE21510		0.97	0.97	0.98	0.97	0.94	0.94	0.94	0.94	0.95	0.95	0.96	0.95	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.95	0.95	0.96	0.95	0.94	0.94	0.95	0.94
GSE50161		0.88	0.88	0.90	0.88	0.82	0.83	0.85	0.82	0.82	0.81	0.84	0.82	0.87	0.86	0.88	0.87	0.85	0.85	0.87	0.85	0.70	0.68	0.74	0.70	0.82	0.82	0.84	0.82
GSE45827		0.40	0.40	0.43	0.40	0.41	0.41	0.45	0.41	0.37	0.32	0.31	0.37	0.42	0.42	0.45	0.42	0.41	0.41	0.45	0.41	0.41	0.38	0.41	0.41	0.40	0.39	0.43	0.40
GSE28497		0.53	0.52	0.54	0.53	0.50	0.50	0.51	0.50	0.44	0.42	0.42	0.44	0.53	0.52	0.53	0.53	0.52	0.52	0.53	0.52	0.47	0.45	0.49	0.47	0.51	0.51	0.52	0.51

Çizelge 4.24’e genel olarak bakıldığında, FS3 kombinasyonu bazı veri setlerinde oldukça yüksek ve istikrarlı performanslar gözlemlenmektedir. Özellikle GSE14520_U133A, GSE44076, GSE63270 ve GSE21510 veri setlerinde, LR, RF ETs ve XGBoost algoritmaları %0.97 ile %1.00 arasında doğruluk ve F1 skoru değerlerine ulaşmıştır. Bu durum, IG'in bu veri setleri için oldukça bilgilendirici ve ayırıcı özellikler seçtiğini göstermektedir. kNN algoritması da bu başarılı veri setlerinde yüksek performans sergilerken, DT özellikle GSE6919_U95Av2 ve GSE50161 gibi daha karmaşık veya gürültülü veri setlerinde diğer algoritmalara kıyasla daha düşük sonuçlar vermiştir. Bu, DT'nin bu tür veri yapılarını işleme kapasitesine sahip olmadığını göstermektedir. Ada Boost ve XGBoost'un başarısına rağmen, GSE45827 ile GSE28497 veri kümeleri için F1 skoru ve hassasiyet açısından belirgin düşüşler yaşaması ise dikkat çekici olmuştur.

FS4 öznitelik seçim yöntemi, performans açısından FS3'e benzer eğilimler sergilemiştir. GSE44076, GSE63270 ve GSE21510 veri setlerinde yine LR, RF ETs ve kNN gibi algoritmalar %0.99-%1.00 aralığında mükemmel sonuçlar elde etmiştir. Bu da ReliefF'in bu veri setleri için etkili bir özellik seçimi sağladığını kanıtlamaktadır. Ancak GSE50161, GSE45827 ve GSE28497 gibi veri setlerinde Ada Boost ve XGBoost algoritmalarının performansı önemli ölçüde azalmıştır (örneğin, Ada Boost’un F1 puanı GSE50161 için 0,42, GSE45827 için 0,34). Bu, ReliefF'nin özellik seçiminin AdaBoost ve XGBoost'un bu veri kümeleri için genelleme yeteneğini olumsuz etkileyebileceğini düşündürmektedir. Dikkat çekici bir şekilde, daha zorlu veri kümelerinde kNN ve ETs (FS4 ile) herhangi bir performans düşüşü göstermemiştir.

FS5 kombinasyonu, kapsamlı bir hibrit yaklaşım olarak dikkat çekmektedir. Bu kombinasyonun, özellikle GSE70947 veri setinde diğer FS kombinasyonlarına göre daha iyi veya benzer performanslar sergilediği görülmüştür. Bu veri setinde RF ve ETs algoritmaları %0.92 doğruluk oranına ulaşarak dikkat çekici sonuçlar vermiştir. Ancak, bazı veri setlerinde (örn. GSE45827 ve GSE28497) bu karmaşık kombinasyonun dahi sınıflandırma performansını beklenen düzeyde artıramadığı, hatta bazı algoritmalar için önceki FS kombinasyonlarından daha düşük sonuçlar verdiği gözlemlenmiştir. Bu durum, daha fazla özellik seçimi tekniğinin eklenmesinin her zaman daha iyi sonuçlar vermediğini ve en iyi kombinasyonun veri setinin belirli yapısına göre değişebileceğini göstermektedir. Ayrıca, algoritmaların GSE45827 ve GSE28497 veri setlerindeki genel zayıf performansı,

bu veri setlerinin doğası gereği daha zor olma olasılığını ortaya koymakta ve daha karmaşık veri ön işleme veya model optimizasyonu gerektirebilmektedir.

Genel bir değerlendirme yapıldığında, LR, RF ve ETs algoritmaları, özellik seçim yaklaşımları (FS3-FS5) altında da geniş bir veri seti yelpazesinde güçlü ve tutarlı bir performans sergilemiştir. Bu algoritmalar, özellikle yüksek boyutlu gen ekspresyonu verilerinde güçlü sınıflandırma yeteneklerini sürdürmüşlerdir. XGBoost da genellikle iyi performans göstermiş, ancak özellikle GSE45827 ve GSE28497 veri setlerinde bazı düşüşler yaşamıştır. kNN ve DT'nin performansı ise veri setine ve özellik seçim yöntemine göre daha değişkenlik göstermiştir. Özellikle DT, bu tabloda da diğerlerine kıyasla genellikle daha düşük performanslı sonuçlar vermiştir.

Sonuç olarak, Çizelge 4.24, gelişmiş özellik seçim yöntemlerinin, özellikle GA gibi optimizasyon teknikleriyle birleştiğinde, sınıflandırma performansını önemli ölçüde artırma potansiyeline sahip olduğunu göstermektedir. Ancak, en iyi sonucun elde edilmesi, özellik seçim stratejisinin ve sınıflandırma algoritmasının veri setinin özellikleriyle uyumlu bir şekilde seçilmesine bağlıdır.

Çizelge 4.25, özellik seçim yaklaşımları (FS6-8) ile çeşitli sınıflandırma algoritmalarının farklı veri setleri üzerindeki test sonuçlarını sunmaktadır. Bu analizde, önceki tablolarla karşılaştırmalı olarak genetik algoritma içeren daha spesifik kombinasyonların performansları değerlendirilmektedir.

Çizelge 4.25. Özellik Seçim Yaklaşımları (FS6-8) ve Sınıflandırma Algoritmalarının Test Sonuçları

		K-Neightbors				Decision Tree				Logistic Regression				Random Forest				Extra Tree				Ada Boost				XGBoost			
veri seti	komb.	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec	acc	f1	pre	rec
GSE70947	F6	0.77	0.77	0.78	0.77	0.74	0.74	0.75	0.74	0.78	0.78	0.79	0.78	0.75	0.75	0.76	0.75	0.74	0.74	0.75	0.74	0.76	0.76	0.77	0.76	0.76	0.76	0.76	0.76
GSE14520_U133A		0,86	0,86	0,87	0,86	0,80	0,80	0,80	0,80	0,86	0,86	0,87	0,86	0,80	0,80	0,81	0,80	0,80	0,80	0,81	0,80	0,84	0,84	0,86	0,84	0,82	0,82	0,83	0,82
GSE44076		0.93	0.93	0.94	0.93	0.92	0.92	0.92	0.92	0.93	0.93	0.94	0.93	0.92	0.92	0.93	0.92	0.92	0.92	0.92	0.92	0.93	0.93	0.94	0.93	0.93	0.93	0.93	0.93
GSE53757		0.84	0.84	0.86	0.84	0.79	0.79	0.80	0.79	0.86	0.86	0.86	0.86	0.78	0.78	0.79	0.78	0.79	0.79	0.80	0.79	0.86	0.86	0.87	0.86	0.76	0.76	0.78	0.76
GSE6919_U95Av2		0.73	0.73	0.74	0.73	0.64	0.64	0.65	0.64	0.69	0.69	0.72	0.69	0.65	0.65	0.66	0.65	0.64	0.63	0.65	0.64	0.73	0.72	0.75	0.73	0.65	0.65	0.67	0.65
GSE19804		0.94	0.94	0.95	0.94	0.90	0.90	0.92	0.90	0.95	0.95	0.96	0.95	0.92	0.92	0.93	0.92	0.91	0.91	0.92	0.91	0.90	0.90	0.92	0.90	0.94	0.94	0.95	0.94
GSE63270		0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96	0.96
GSE21510		0.78	0.78	0.79	0.78	0.74	0.74	0.75	0.74	0.80	0.80	0.82	0.80	0.75	0.75	0.76	0.75	0.74	0.74	0.75	0.74	0.74	0.73	0.81	0.74	0.73	0.73	0.75	0.73
GSE50161		0.49	0.48	0.49	0.49	0.45	0.44	0.46	0.45	0.54	0.50	0.57	0.54	0.45	0.45	0.47	0.45	0.44	0.44	0.46	0.44	0.53	0.48	0.48	0.53	0.44	0.44	0.46	0.44
GSE45827		0.54	0.52	0.55	0.54	0.49	0.49	0.52	0.49	0.45	0.37	0.39	0.45	0.50	0.49	0.52	0.50	0.50	0.49	0.52	0.50	0.48	0.39	0.37	0.48	0.51	0.50	0.53	0.51
GSE28497		0.39	0.38	0.40	0.39	0.36	0.35	0.37	0.36	0.24	0.16	0.14	0.24	0.35	0.35	0.37	0.35	0.35	0.35	0.36	0.35	0.31	0.22	0.20	0.31	0.35	0.35	0.36	0.35
GSE70947	F7	0.88	0.88	0.89	0.88	0.83	0.83	0.84	0.83	0.90	0.90	0.90	0.90	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89
GSE14520_U133A		0,97	0,97	0,97	0,97	0,95	0,95	0,95	0,95	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97	0,97
GSE44076		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GSE53757		0.84	0.84	0.85	0.84	0.79	0.79	0.80	0.79	0.83	0.83	0.84	0.83	0.85	0.85	0.85	0.85	0.86	0.86	0.86	0.86	0.84	0.84	0.85	0.84	0.85	0.85	0.85	0.85
GSE6919_U95Av2		0.72	0.71	0.73	0.72	0.67	0.66	0.70	0.67	0.75	0.75	0.77	0.75	0.77	0.77	0.79	0.77	0.75	0.75	0.76	0.75	0.69	0.69	0.71	0.69	0.72	0.72	0.74	0.72
GSE19804		0.96	0.96	0.96	0.96	0.91	0.91	0.92	0.91	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.94	0.94	0.95	0.94	0.91	0.91	0.93	0.91
GSE63270		1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.97	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.96
GSE21510		1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98
GSE50161		0.94	0.94	0.95	0.94	0.92	0.92	0.93	0.92	0.95	0.95	0.96	0.95	0.96	0.96	0.96	0.96	0.98	0.98	0.98	0.98	0.89	0.89	0.91	0.89	0.95	0.95	0.95	0.95
GSE45827		0.93	0.93	0.94	0.93	0.87	0.87	0.89	0.87	0.94	0.94	0.95	0.94	0.93	0.93	0.94	0.93	0.95	0.95	0.95	0.95	0.43	0.37	0.38	0.43	0.89	0.89	0.90	0.89
GSE28497		0.88	0.88	0.89	0.88	0.83	0.83	0.84	0.83	0.88	0.87	0.88	0.88	0.89	0.89	0.90	0.89	0.91	0.91	0.92	0.91	0.71	0.70	0.75	0.71	0.89	0.89	0.90	0.89
GSE70947	F8	0.85	0.85	0.85	0.85	0.76	0.76	0.76	0.76	0.82	0.81	0.84	0.82	0.77	0.77	0.77	0.77	0.76	0.76	0.77	0.76	0.84	0.84	0.85	0.84	0.80	0.80	0.81	0.80
GSE14520_U133A		0,78	0,77	0,78	0,78	0,73	0,73	0,73	0,73	0,81	0,81	0,82	0,81	0,73	0,73	0,74	0,73	0,73	0,73	0,73	0,73	0,81	0,80	0,82	0,81	0,76	0,76	0,76	0,76
GSE44076		0.94	0.94	0.95	0.94	0.94	0.94	0.95	0.94	0.92	0.92	0.93	0.92	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.95	0.94
GSE53757		0.84	0.84	0.85	0.84	0.78	0.78	0.79	0.78	0.83	0.84	0.84	0.83	0.79	0.79	0.80	0.79	0.78	0.78	0.78	0.78	0.86	0.86	0.86	0.86	0.77	0.77	0.78	0.77
GSE6919_U95Av2		0.67	0.67	0.69	0.67	0.60	0.59	0.63	0.60	0.67	0.65	0.73	0.67	0.58	0.58	0.61	0.58	0.59	0.59	0.62	0.59	0.66	0.66	0.70	0.66	0.60	0.60	0.62	0.60
GSE19804		0.84	0.84	0.85	0.84	0.81	0.81	0.82	0.81	0.83	0.83	0.85	0.83	0.79	0.79	0.80	0.79	0.81	0.81	0.82	0.81	0.88	0.88	0.89	0.88	0.81	0.81	0.82	0.81
GSE63270		0.86	0.86	0.87	0.86	0.72	0.71	0.73	0.72	0.79	0.78	0.80	0.79	0.73	0.73	0.74	0.73	0.72	0.71	0.73	0.72	0.80	0.80	0.82	0.80	0.77	0.76	0.77	0.77
GSE21510		0.78	0.78	0.78	0.78	0.75	0.75	0.76	0.75	0.78	0.78	0.79	0.78	0.75	0.75	0.75	0.75	0.74	0.74	0.75	0.74	0.76	0.74	0.82	0.76	0.75	0.75	0.76	0.75
GSE50161		0.49	0.49	0.52	0.49	0.52	0.53	0.55	0.52	0.41	0.35	0.35	0.41	0.52	0.52	0.55	0.52	0.52	0.52	0.55	0.52	0.46	0.42	0.44	0.46	0.48	0.48	0.52	0.48
GSE45827		0.53	0.53	0.55	0.53	0.50	0.49	0.52	0.50	0.47	0.41	0.49	0.47	0.51	0.50	0.53	0.51	0.50	0.50	0.54	0.50	0.51	0.44	0.43	0.51	0.53	0.53	0.55	0.53
GSE28497		0.37	0.36	0.39	0.37	0.34	0.34	0.35	0.34	0.34	0.27	0.26	0.34	0.34	0.34	0.35	0.34	0.34	0.34	0.35	0.34	0.33	0.23	0.23	0.33	0.35	0.34	0.36	0.35

Çizelge 4.25’e bakıldığında, FS6 kombinasyonu, tabloya göre genel olarak orta düzeyde performanslar sergilemiştir. Özellikle GSE70947, GSE14520_U133A ve GSE44076 gibi veri setlerinde algoritmaların performansları %0.74 ile %0.93 arasında değişmektedir. Bu durum, bu karmaşık hibrit modelin her veri setinde en üst düzeyde sonuç vermediğini göstermektedir. Örneğin, GSE44076 veri kNN, LR, Ada Boost ve XGBoost %0.93 doğruluk oranına ulaşırken, RF ve ETs %0.92'de kalmıştır. Daha zorlu veri setlerinde, örneğin GSE50161, GSE45827 ve GSE28497'de performanslar belirgin şekilde düşmektedir. GSE28497 veri setinde tüm algoritmalar %0.30'lar seviyesinde performans göstermiştir ki bu oldukça düşüktür. Bu, FS6 kombinasyonunun bu tür zorlu veri setlerinin içerdiği karmaşıklıkla başa çıkmada yetersiz kalabileceğine işaret etmektedir. DT algoritması, genel olarak FS6 ile diğerlerine kıyasla daha düşük performans sergilemeye devam etmiştir (%0.64 - %0.92 arası).

FS7 kombinasyonu, FS6'ya benzer ancak GA'nın farklı bir kullanımı ile dikkat çekmektedir. Bu yaklaşım, birçok veri setinde kayda değer ölçüde daha iyi performanslar sunmuştur. Özellikle GSE14520_U133A, GSE44076, GSE63270 ve GSE21510 veri setlerinde, LR, RF, ETs ve XGBoost algoritmaları %0.97 ile %1.00 arasında mükemmel doğruluk ve F1 skoru değerlerine ulaşmıştır. Bu durum, FS7'nin bu veri setleri için oldukça ayırıcı ve bilgilendirici özellik setleri ürettiğini göstermektedir. Ancak, GSE6919_U95Av2 gibi orta zorluktaki veri setlerinde algoritmalar %0.70'ler seviyesinde kalmıştır. En büyük iyileşme ise GSE50161 ve GSE45827 veri setlerinde gözlemlenmektedir. Özellikle GSE50161'de RF, ETs ve XGBoost %0.95'in üzerinde performans gösterirken, GSE45827'de de benzer şekilde yüksek sonuçlar elde edilmiştir. Bu, FS7'nin, önceki FS kombinasyonlarına göre daha geniş bir yelpazedeki veri setlerinde daha güçlü ve genellenebilir özellik setleri sağlamada başarılı olduğunu göstermektedir. Ada Boost algoritması, GSE45827 ve GSE28497 veri setlerinde hala düşük performans sergilese de, genel olarak önceki FS6'ya göre daha iyi bir tablo çizmektedir.

FS8 kombinasyonu, GA'nın Ki-kare ile birlikte daha odaklı kullanımını temsil etmektedir. Genel olarak bakıldığında, FS8'in performansının FS6'dan daha iyi, ancak FS7'den biraz daha düşük olduğu söylenebilir. GSE44076, GSE63270 ve GSE21510 gibi veri setlerinde yine yüksek performanslar (%0.94 - %1.00) gözlemlenmiştir. Bu çıktılar, Ki-kare ve

Genetik Algoritmanın bu veri setleri için etkili özellikler bulabildiğini göstermektedir. Ancak, GSE14520_U133A veri setinde FS8'in performansı (%0.70'ler) FS7'ye göre belirgin şekilde düşerken, GSE6919_U95Av2 gibi veri setlerinde FS8 ile bazı algoritmalar (örn. LR, kNN) FS7'ye kıyasla daha düşük sonuçlar vermiştir. Özellikle GSE50161, GSE45827 ve GSE28497 gibi zorlu veri setlerinde FS8'in performansı, FS7'ye göre daha düşüktür. Örneğin, GSE28497'de LR'nin performansı %0.24'e kadar düşerken, Ada Boost ve XGBoost da düşük sonuçlar vermiştir. Bu durum, GA'nın tek başına Ki-kare ile birleşiminin, tüm veri setlerinde en iyi sonuçları garanti etmediğini ve diğer temel özellik seçim yöntemlerinin (FKs, IG, ReliefF) eklenmesinin (FS7'de olduğu gibi) genel performansı artırabildiğini ortaya koymaktadır.

Bu çizelgeye göre, LR, RF ve ET algoritmaları, özellikle FS7 kombinasyonu altında en tutarlı ve yüksek performansları sergilemişlerdir. Bu durum, bu algoritmaların farklı özellik seçim stratejilerine adaptasyon yeteneklerinin yüksek olduğunu ve gen ekspresyonu verilerindeki sınıflandırma görevleri için güvenilir seçenekler olduğunu pekiştirmektedir. XGBoost genel olarak üstün performans gösterse de zorlu veri kümelerine karşı, özellikle FS6, FS8 kombinasyonlarında düşüşler gözlemlenmiştir. kNN'nin performansı veri kümesine göre değişse de, DT ortalamada daha da kötü ve diğerlerinden daha az tutarlı bir performans göstermiştir.

Sonuç olarak, Çizelge 4.25, hibrit özellik seçim yöntemlerinin karmaşıklıklarının ve bileşenlerinin sınıflandırma performansı üzerindeki etkilerini açıkça ortaya koymaktadır. FS7 kombinasyonu, bu tablodaki diğer hibrit yaklaşımlara göre genel olarak daha başarılı ve genellenebilir sonuçlar üretmiştir. Bu da, farklı özellik seçim algoritmalarının uyumlu bir şekilde birleştirilmesinin, sadece GA'ya odaklanmaktan daha verimli olabileceğini düşündürmektedir.

Çizelge 4.26, özellik seçim yaklaşımları (FS9-FS11) ile çeşitli sınıflandırma algoritmalarının farklı veri setleri üzerindeki test sonuçlarını sunmaktadır. Bu tablo, genetik algoritmaların yanı sıra diğer spesifik filtre tabanlı ve hibrit kombinasyonların sınıflandırma performansı üzerindeki etkilerini daha kapsamlı bir şekilde incelemektedir.

Çizelge 4.26. Özellik Seçim Yaklaşımları (FS9-11) ve Sınıflandırma Algoritmalarının Test Sonuçları

		K-Neighbors				Decision Tree				Logistic Regression				Random Forest				Extra Tree				Ada Boost				XGBoost			
veri seti	komb.	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec	acc	fl	pre	rec
GSE70947	F9	0.81	0.81	0.82	0.81	0.73	0.73	0.74	0.73	0.82	0.82	0.84	0.82	0.74	0.74	0.75	0.74	0.73	0.73	0.74	0.73	0.83	0.83	0.84	0.83	0.77	0.77	0.78	0.77
GSE14520_U133A		0,92	0,92	0,92	0,92	0,90	0,90	0,90	0,90	0,93	0,93	0,94	0,93	0,89	0,89	0,90	0,89	0,90	0,90	0,90	0,90	0,94	0,94	0,94	0,94	0,90	0,90	0,91	0,90
GSE44076		0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.97	0.98	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
GSE53757		0.89	0.89	0.90	0.89	0.83	0.83	0.85	0.83	0.88	0.88	0.89	0.88	0.83	0.83	0.84	0.83	0.82	0.82	0.84	0.82	0.87	0.87	0.89	0.87	0.83	0.83	0.85	0.83
GSE6919_U95Av2		0.62	0.62	0.66	0.62	0.63	0.63	0.68	0.63	0.68	0.68	0.72	0.68	0.63	0.63	0.67	0.63	0.64	0.64	0.69	0.64	0.60	0.60	0.63	0.60	0.60	0.60	0.63	0.60
GSE19804		0.96	0.96	0.97	0.96	0.92	0.92	0.92	0.92	0.97	0.97	0.97	0.97	0.93	0.93	0.93	0.93	0.91	0.91	0.92	0.91	0.92	0.92	0.92	0.92	0.96	0.96	0.97	0.96
GSE63270		0.97	0.97	0.98	0.97	0.96	0.96	0.97	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.97	0.96	0.94	0.94	0.95	0.94	0.96	0.96	0.97	0.96	0.97	0.97	0.97	0.97
GSE21510		0.88	0.88	0.90	0.88	0.86	0.86	0.86	0.86	0.84	0.84	0.85	0.84	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.89	0.88	0.91	0.89	0.85	0.85	0.86	0.85
GSE50161		0.53	0.53	0.55	0.53	0.51	0.51	0.54	0.51	0.54	0.51	0.52	0.54	0.51	0.51	0.54	0.51	0.52	0.52	0.55	0.52	0.57	0.55	0.59	0.57	0.53	0.53	0.54	0.53
GSE45827		0.57	0.57	0.59	0.57	0.51	0.51	0.54	0.51	0.53	0.51	0.52	0.53	0.52	0.52	0.56	0.52	0.51	0.51	0.54	0.51	0.54	0.50	0.51	0.54	0.53	0.51	0.56	0.53
GSE28497		0.34	0.34	0.36	0.34	0.33	0.33	0.34	0.33	0.32	0.26	0.25	0.32	0.33	0.33	0.34	0.33	0.33	0.33	0.34	0.33	0.33	0.29	0.28	0.33	0.33	0.33	0.34	0.33
GSE70947	F10	0.81	0.81	0.82	0.81	0.78	0.78	0.79	0.78	0.83	0.83	0.84	0.83	0.78	0.78	0.79	0.78	0.77	0.77	0.78	0.77	0.82	0.82	0.83	0.82	0.80	0.80	0.81	0.80
GSE14520_U133A		0,95	0,95	0,95	0,95	0,91	0,91	0,92	0,91	0,95	0,95	0,95	0,95	0,91	0,91	0,92	0,91	0,91	0,91	0,91	0,91	0,95	0,95	0,95	0,95	0,92	0,92	0,92	0,92
GSE44076		0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
GSE53757		0.90	0.90	0.90	0.90	0.82	0.82	0.84	0.82	0.87	0.87	0.88	0.87	0.82	0.82	0.84	0.82	0.82	0.82	0.83	0.82	0.87	0.87	0.88	0.87	0.86	0.86	0.86	0.86
GSE6919_U95Av2		0.66	0.65	0.69	0.66	0.58	0.58	0.60	0.58	0.63	0.62	0.67	0.63	0.58	0.57	0.60	0.58	0.58	0.58	0.60	0.58	0.63	0.61	0.67	0.63	0.64	0.63	0.67	0.64
GSE19804		0.92	0.92	0.93	0.92	0.91	0.91	0.92	0.91	0.93	0.93	0.94	0.93	0.91	0.91	0.93	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.92	0.91
GSE63270		0.95	0.95	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.94	0.94	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.95	0.94
GSE21510		0.94	0.94	0.94	0.94	0.91	0.92	0.92	0.91	0.93	0.93	0.93	0.93	0.92	0.92	0.92	0.92	0.91	0.91	0.92	0.91	0.93	0.93	0.94	0.93	0.92	0.92	0.93	0.92
GSE50161		0.65	0.65	0.67	0.65	0.64	0.65	0.67	0.64	0.53	0.47	0.46	0.53	0.65	0.65	0.67	0.65	0.63	0.64	0.67	0.63	0.56	0.52	0.55	0.56	0.62	0.62	0.65	0.62
GSE45827		0.61	0.61	0.66	0.61	0.56	0.56	0.58	0.56	0.61	0.60	0.64	0.61	0.57	0.57	0.59	0.57	0.56	0.56	0.58	0.56	0.52	0.45	0.45	0.52	0.58	0.58	0.62	0.58
GSE28497		0.46	0.45	0.48	0.46	0.39	0.38	0.39	0.39	0.43	0.38	0.38	0.43	0.40	0.40	0.41	0.40	0.38	0.37	0.38	0.38	0.43	0.38	0.42	0.43	0.43	0.42	0.44	0.43
GSE70947	F11	0.79	0.79	0.81	0.79	0.75	0.75	0.76	0.75	0.82	0.82	0.85	0.82	0.75	0.75	0.76	0.75	0.76	0.76	0.76	0.76	0.81	0.80	0.84	0.81	0.77	0.77	0.78	0.77
GSE14520_U133A		0.94	0.94	0.94	0.94	0.92	0.92	0.92	0.92	0.94	0.94	0.94	0.94	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.94	0.94	0.94	0.94	0.93	0.93	0.94	0.93
GSE44076		0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.94	0.94	0.94	0.94	0.95	0.95	0.96	0.95	0.95	0.95	0.96	0.95	0.96	0.96	0.96	0.96
GSE53757		0.82	0.82	0.84	0.82	0.79	0.79	0.80	0.79	0.85	0.85	0.86	0.85	0.78	0.78	0.79	0.78	0.79	0.79	0.79	0.79	0.80	0.80	0.82	0.80	0.79	0.79	0.80	0.79
GSE6919_U95Av2		0.60	0.59	0.62	0.60	0.55	0.55	0.56	0.55	0.68	0.67	0.73	0.68	0.57	0.57	0.58	0.57	0.53	0.53	0.54	0.53	0.62	0.60	0.68	0.62	0.61	0.61	0.62	0.61
GSE19804		0.91	0.91	0.91	0.91	0.89	0.89	0.90	0.89	0.92	0.92	0.93	0.92	0.90	0.90	0.92	0.90	0.88	0.88	0.90	0.88	0.89	0.89	0.91	0.89	0.92	0.92	0.92	0.92
GSE63270		0.94	0.94	0.95	0.94	0.93	0.93	0.94	0.93	0.94	0.94	0.95	0.94	0.92	0.92	0.93	0.92	0.92	0.92	0.93	0.92	0.93	0.93	0.94	0.93	0.92	0.92	0.93	0.92
GSE21510		0.84	0.84	0.85	0.84	0.79	0.79	0.80	0.79	0.87	0.86	0.87	0.87	0.79	0.79	0.80	0.79	0.79	0.79	0.80	0.79	0.83	0.83	0.85	0.83	0.80	0.80	0.81	0.80
GSE50161		0.47	0.47	0.50	0.47	0.49	0.49	0.51	0.49	0.49	0.45	0.46	0.49	0.47	0.47	0.50	0.47	0.48	0.48	0.50	0.48	0.45	0.36	0.38	0.45	0.43	0.42	0.44	0.43
GSE45827		0.44	0.43	0.45	0.44	0.45	0.45	0.48	0.45	0.34	0.28	0.29	0.34	0.45	0.45	0.48	0.45	0.45	0.45	0.48	0.45	0.37	0.29	0.32	0.37	0.42	0.42	0.45	0.42
GSE28497		0.47	0.46	0.50	0.47	0.43	0.42	0.44	0.43	0.47	0.46	0.51	0.47	0.42	0.42	0.43	0.42	0.42	0.42	0.43	0.42	0.35	0.26	0.27	0.35	0.43	0.42	0.45	0.43

Çizelge 4.26'ya bakıldığında, FS9 kombinasyonun genel performansı veri setine göre değişkenlik göstermektedir. GSE14520_U133A (%0.92-0.94), GSE44076 (%0.97-0.98), GSE19804 (%0.91-0.97) ve GSE63270 (%0.94-0.97) gibi veri setlerinde oldukça yüksek doğruluk oranları elde edilmiştir. Özellikle GSE44076 veri setinde, çoğu sınıflandırma algoritması çok başarılı sonuçlar vermiştir. Bu durum, FKs ve GA'nın bu veri setlerindeki ayırt edici özellikleri bulmada etkili olduğunu göstermektedir. Ancak, GSE6919_U95Av2 (%0.60-0.68) gibi orta zorluktaki veri setlerinde performans düşerken, GSE50161 (%0.51-0.57), GSE45827 (%0.51-0.57) ve GSE28497 (%0.32-0.34) gibi daha zorlu veri setlerinde algoritmaların doğruluk oranları belirgin şekilde düşmektedir. Özellikle DT ve Ada Boost, bu zorlu veri setlerinde en düşük performansları sergilemiştir. Bu da FKs ve GA birleşiminin her veri setinin karmaşıklığıyla başa çıkmakta zorlandığını ortaya koymaktadır.

FS10 kombinasyonu, IG ve GA'nın bir araya gelmesiyle oluşmuştur. Bu yaklaşım, genel olarak FS9'a kıyasla daha istikrarlı ve bazı veri setlerinde daha iyi sonuçlar sunmuştur. GSE14520_U133A (%0.91-0.95), GSE44076 (%0.98-0.99), GSE19804 (%0.91-0.93) ve GSE63270 (%0.94-0.95) veri setlerinde önceki yaklaşımlara benzer şekilde yüksek doğruluk oranları korunmuştur. LR ve XGBoost bu kombinasyonla sıklıkla en yüksek performansları gösteren algoritmalar arasında yer almıştır. Önceki tablolarda düşük performans gösteren veri setlerinde de önemli iyileşmeler gözlemlenmiştir. Örneğin, GSE50161 veri setinde kNN ve RF %0.65'e kadar yükselirken, GSE45827'de de benzer bir eğilim görülmüştür. Ancak, GSE6919_U95Av2 (%0.58-0.66) ve GSE28497 (%0.38-0.46) veri setlerinde performanslar hala düşük kalmaktadır. Bu durum, IG ve GA'nın birleşiminin genel performansı artırdığını, ancak hala bazı veri setleri için yetersiz kalabildiğini göstermektedir.

FS11 kombinasyonu, Relief ve GA'nın bir araya gelmesiyle oluşturulmuştur. Bu yaklaşımın performansı, genel olarak FS10'a oldukça yakın, ancak bazı veri setlerinde hafif farklılıklar göstermektedir. GSE14520_U133A (%0.92-0.94), GSE44076 (%0.94-0.96), GSE19804 (%0.88-0.92) ve GSE63270 (%0.92-0.94) gibi veri setlerinde yüksek ve tutarlı sonuçlar elde edilmiştir. Özellikle LR ve XGBoost bu kombinasyonla da güçlü performanslarını sürdürmüştür. Ancak, GSE6919_U95Av2 (%0.53-0.68) gibi veri setlerinde performans hala ortalama seviyededir. En zorlu veri setleri olan GSE50161 (%0.36-0.49), GSE45827 (%0.28-0.48) ve GSE28497 (%0.26-0.47)'de ise performanslar

hala düşüktür. Özellikle Ada Boost ve XGBoost'un bu veri setlerindeki düşük performansları dikkat çekicidir. Bu durum, Relief ve GA'nın birleşiminin de tüm veri setlerinde yüksek başarı garantilemediğini ve özellikle gürültülü veya karmaşık düşük boyutlu veri setlerinde zorlanabileceğini düşündürmektedir.

Performanslarına göre, en istikrarlı ve en yüksek doğruluk oranları LR ve XGBoost yöntemleri tarafından (GSE14520_U133A, GSE44076, GSE19804, GSE63270) elde edilmiştir. Genel olarak, kNN, LR, RF ve ETs algoritmaları dengeli performans gösterse de bazı durumlarda düşüşlere neden olmuştur. Ancak DT ve Ada Boost, diğer algoritmalara kıyasla daha düşük sonuçlar vermiştir

Sonuç olarak, Çizelge 4.26, genetik algoritmaların FKs, IG ve Relief gibi farklı filtre tabanlı özellik seçim yöntemleriyle kombinasyonlarının sınıflandırma performansı üzerindeki çeşitli etkilerini göstermektedir. FS10 kombinasyonu, genel olarak en istikrarlı ve başarılı sonuçları sunarken, FS9 ve FS11'in performansı veri setine göre daha fazla değişkenlik göstermiştir. Bazı çok zorlu veri setlerinin sınıflandırılması ise hala önemli bir zorluk teşkil etmektedir. Bu analiz, özellik seçim stratejilerinin veri setinin doğasına göre özelleştirilmesinin önemini bir kez daha vurgulamaktadır.

Sonuç olarak, bu çalışmada uygulanan çeşitli öznitelik seçim yöntemlerinin, farklı sınıflandırma algoritmalarıyla birlikte kullanılarak, çeşitli gen ekspresyonu veri setleri üzerindeki performansları kapsamlı bir şekilde değerlendirilmiştir. Elde edilen sonuçlar, kullanılan öznitelik seçim stratejilerinin ve sınıflandırma algoritmalarının veri setine özgü olarak farklı performanslar sergileyebileceğini ortaya koymuştur. Genel olarak, LR, RF, ETs ve XGBoost gibi algoritmalar, FS0-FS2 arası temel öznitelik seçim kombinasyonları altında çoğu veri setinde yüksek doğruluk ve F1 skoru değerleri elde etmiştir. Özellikle yüksek boyutlu gen ekspresyonu verilerinde güçlü sınıflandırma yetenekleri göstermişlerdir. Bununla birlikte, bazı veri setlerinde DT ve kNN algoritmalarının performansları daha değişken olmuş, özellikle daha karmaşık veya gürültülü veri setlerinde bu algoritmaların performansları düşmüştür. Öznitelik seçim kombinasyonları arasında, FS3-FS5 gibi yöntemlerin daha başarılı sonuçlar sunduğu, özellikle GA içeren kombinasyonların sınıflandırma performansını artırdığı gözlemlenmiştir. Ancak, bazı veri setlerinde, özellikle GSE45827 ve GSE28497 gibi daha zorlu veri setlerinde, her öznitelik seçim kombinasyonu yüksek performans gösterememiştir. Bu, her veri kümesinin kendine

özgü özellikleri olduğunu ve dolayısıyla özellik seçim stratejisinin bu zorluklara uyum sağlaması gerektiğini vurgulamaktadır. Son olarak, çalışma, özellik seçimi yöntemleri ve sınıflandırma algoritmaları kombinasyonunun gen ekspresyonu gibi yüksek boyutlu veri setlerinde sınıflandırma başarısı için önemli belirleyicilerden biri olduğunu ve veri kümesine bağlı olarak dikkatli bir şekilde seçilmesi gerektiğini göstermektedir.

4.2. Literatürel Karşılaştırma

Önerilen tekniklerin etkinliğini ve yeni yaklaşımların performansa katkısını değerlendirmek için, bu çalışmanın bulguları literatürdeki çalışmaların bulgularıyla karşılaştırılmıştır. Karmaşık biyolojik süreçleri analiz etmek için literatürde çok sayıda teknik geliştirilmiştir. Bu çalışmaların çoğunda farklı özellik seçim teknikleri ve sınıflandırma algoritmaları test edilmiştir. Bununla birlikte, veri setinin özelliklerine, ele alınan sınıflandırma probleminin türüne ve yöntemlerin kurulumuna dayanarak, her çalışma farklı bir metodoloji göstermiştir.

Çizelge 4.27. GSE70947: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
ESA - DL [79]	0.73	-	-	-
FS [80]	0.92	-	-	-
DE [81]	0.94	0.84	0.8	0.90
PSO [81]	0.94	0.84	0.81	0.90
GA [81]	0.93	0.84	0.81	0.91
SVM-RFE – BDF [82]	0.89	-	-	-
NSGA-II Optimizasyon [83]	0.96	-	-	-
FCBF [84]	0.96	0.96	0.95	0.97
SpAM, M=3 [85]	0.95	-	-	-
Önerilen yöntem (FS5)	0.92	0.92	0.92	0.92

Çizelge 4.27'de görüldüğü üzere, GSE70947 meme kanseri veri kümesi üzerinde gerçekleştirilen literatür çalışmaları, farklı öznitelik seçimi ve sınıflandırma yaklaşımlarıyla %89 ile %96 arasında değişen doğruluk oranları elde etmiştir. Örneğin, Diferansiyel Evrim (DE), PSO ve GA gibi evrimsel algoritmalar %93–%94 doğruluk elde ederken, NSGA-II (Non-Dominated Sorting Genetic Algorithm 2) ve FCBF (Fast correlation-based filter) yöntemleri %96 doğrulukla öne çıkmaktadır. Önerilen FS5-RF yöntemi ise, %92 doğrulukla birlikte tüm metriklerde dengeli ve yüksek performans

göstermiştir. Bu da çoklu öznitelik seçim stratejilerinin birleşimiyle elde edilen kapsamlı öznitelik kümesinin, literatürdeki tekil yaklaşımlara kıyasla daha tutarlı sonuçlar verdiğini ortaya koymaktadır.

Literatürde GSE70947 veri kümesi üzerinde yapılan farklı çalışmalard%89 ila %96 doğruluk oranları göstermiştir. Örneğin, GA, DE, PSO tabanlı yöntemler %93-%94 doğruluk ve %84 F1-Skoru ve SVM-RFE yöntemi için %89 doğruluk elde etmiştir. NSGA-II tabanlı optimizasyon yaklaşımı %96 doğrulukla dikkat çekerken, FCBF yöntemi %96 doğruluk gibi yüksek değerler sunmuştur. Ancak bu yöntemlerin birçoğu ya yalnızca tek bir öznitelik seçimi yaklaşımına dayalıdır. Çalışmada önerilen FS5-RF yöntemi ise, %92 doğruluk, %92 F1-Skor, %92 kesinlik ve %92 hassasiyet gibi dengeli ve tutarlı sonuçlar sunarak hem genel başarı hem de sınıf dağılımındaki denge açısından literatürdeki bazı yöntemleri geride bırakmaktadır. Çoklu filtre ve sarmal öznitelik seçimi stratejilerinin birleşimiyle oluşturulan FS5 yöntemi, sadece doğruluk değil, tüm sınıflandırma metriklerinde istikrarlı bir performans sağlayarak literatüre önemli bir katkı sunmaktadır.

Çizelge 4.28. GSE14520_U133A: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
TL [86]	0.97	-	-	-
Stacking ensemble [87]	0.97	0.97	0.96	0.99
Önerilen yöntem (FS1, FS3, FS7)	0.97	0.97	0.97	0.97

Çizelge 4.28'de görüldüğü üzere, Karaciğer kanseri sınıflandırması için kullanılan GSE14520_U133A veri kümesi üzerinde yapılan çalışmalar, yüksek doğruluk oranları elde eden farklı yaklaşımlar önermiştir. Literatürde, Transfer Learning (TL) tabanlı yöntem ve Stacking ensemble yaklaşımı %97 doğruluk sunmuştur. Bu çalışmada önerilen FS1, FS3 ve FS7 öznitelik seçim yöntemleri kullanılarak elde edilen sonuçlar ise %97 doğruluk, %97 F1-skor, %97 kesinlik ve %97 duyarlılık değerleriyle hem doğruluk açısından eşdeğer hem de tüm metriklerde dengeli ve tutarlı bir performans sergilemiştir. Literatürdeki bazı yöntemler belirli metriklerde yüksek değerlere ulaşsa da önerilen yöntem, tüm temel sınıflandırma ölçütlerinde kararlı (tutarlı) başarım çıktıları sunmuştur. Ayrıca, öznitelik seçimi tekniklerinin akıllı kombinasyonları sayesinde modelin hem öğrenme başarısı hem de genellenebilirliği artırılmış, bu da önerilen yaklaşımı literatürdeki yöntemlerle rekabet edebilir ve pratikte uygulanabilir hale getirmiştir.

Çizelge 4.29. GSE6919_U95B: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
NNE [88]	0.74	-	-	-
ESA with DL [79]	0.88	-	-	-
QMFOA [89]	0.96	-	-	-
Trend Filtering, k=1 [85]	0.77	-	-	-
Önerilen yöntem (FS7)	0.77	0.77	0.79	0.77

Çizelge 4.29'da görüldüğü üzere, GSE6919_U95B veri seti üzerinde gerçekleştirilen literatür çalışmaları, farklı yaklaşımlar kullanılarak çeşitli doğruluk seviyelerine ulaşmıştır. Nöral ağ temelli NNE (Neural Network Ensemble) yöntemi %74 doğruluk sunarken, derin öğrenme ile zenginleştirilmiş ESA yöntemi %88 doğruluk sağlamıştır. QMFOA (Quantum Moth Flame Optimization Algorithm) algoritması ise %96 doğrulukla literatürde en yüksek performansı sergileyen yöntem olarak öne çıkmaktadır. Buna karşın, önerilen FS7 öznitelik seçim yöntemi %77 doğruluk, %77 F1-Skor, %79 kesinlik ve %77 duyarlılık değerleri ile dengeli bir performans ortaya koymuştur.

Her ne kadar önerilen yöntemin doğruluk oranı bazı literatür yaklaşımlarının gerisinde olsa da, tüm sınıflandırma metriklerinde sağladığı denge, modelin tutarlılığını ve sınıflar arası ayrımı ne derece başarılı gerçekleştirdiğini göstermektedir. Ayrıca, öznitelik seçiminin akıllı kombinasyonuna dayanan bu yaklaşım, sınıflandırma başarısını artırırken modelin genellenebilirliğini de desteklemektedir. Bu bağlamda, FS7 yöntemi GSE6919_U95B veri seti üzerinde istikrarlı ve uygulanabilir bir çözüm olarak değerlendirilmektedir.

Çizelge 4.30. GSE19804: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
FCBF [90]	0.95	0.96	0.95	0.97
L_1 [91]	0.81	-	-	0.87
L_{EN} [91]	0.81	-	-	0.92
$L_{1/2}$ [91]	0.87	-	-	0.92
FGS without FS [92]	0.95	0.95	0.96	0.95
FGS [92]	0.96	0.96	0.96	0.96
Önerilen yöntem (FS1)	0.99	0.99	0.99	0.99

Çizelge 4.30'da görüldüğü üzere, GSE19804 veri seti üzerinde gerçekleştirilen çalışmalarda, çeşitli öznitelik seçimi ve sınıflandırma yaklaşımlarıyla yüksek doğruluk

oranları elde edilmiştir. Literatürde, FCBF yöntemi %95 doğruluk ve %96 F1-skor ile başarılı sonuçlar sunarken, L_1, L_{EN} ve $L_{1/2}$ gibi yöntemler %81–%87 arasında doğruluk sağlamıştır. FGS (Fuzzy Gene Selection) yöntemi ise %96 doğruluk ve tüm sınıflandırma metriklerinde dengeli değerlerle dikkat çekmiştir.

Bu çalışmada önerilen FS1 yöntemi, %99 doğruluk, %99 F1-Skor, %99 kesinlik ve %99 duyarlılık değerleriyle hem doğruluk oranı hem de sınıflandırma metrikleri açısından literatürdeki tüm yöntemleri geride bırakmaktadır. Elde edilen bu üstün performans, öznitelik seçim sürecinin sınıflandırma başarısına olan etkisini açıkça ortaya koymakta ve önerilen yöntemin hem öğrenme kabiliyeti hem de genellenebilirlik açısından güçlü bir alternatif sunduğunu göstermektedir.

Çizelge 4.31. GSE44076: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
NN/SVC [93]	0.99	-	-	-
LASSO [94]	-	0.88	0.87	0.88
COL10A1 [95]	-	0.70	0.78	0.63
Önerilen yöntem (FS2, FS3, FS4, FS7)	1.00	1.00	1.00	1.00

Çizelge 4.31’de görüldüğü üzere, GSE44076 veri seti üzerinde gerçekleştirilen çeşitli çalışmalar, kolorektal kanserin sınıflandırılmasına yönelik farklı yöntemlerin performansını ortaya koymaktadır. Literatürde yer alan NN/SVC tabanlı yöntem %99 doğruluk oranıyla oldukça başarılı bir sonuç sunarken, LASSO temelli yaklaşım %0.88 F1-Skor, %0.87 kesinlik ve %0.88 duyarlılık değerleri elde etmiştir. Kan bazlı bir biyobelirteç olan COL10A1'e dayalı model ise %0.63 duyarlılık, %0.78 kesinlik ve %0.69 F1-Skor ile orta düzeyde performans göstermektedir.

Bu çalışmada önerilen FS2, FS3, FS4 ve FS7 öznitelik seçim stratejilerinin kullanıldığı model; doğruluk, F1-Skor, kesinlik ve duyarlılık metriklerinde tam doğruluk (%100) ile tüm sınıflandırma çıktılarında maksimum başarıya ulaşmıştır. Bu sonuç, önerilen çoklu öznitelik seçimi stratejilerinin, veriye en uygun ve ayrım gücü yüksek öznitelikleri belirleyerek sınıflandırma performansını üst düzeye taşıdığını göstermektedir. Literatürdeki yöntemlerle karşılaştırıldığında, önerilen yaklaşım yalnızca doğruluk

açısından değil, tüm metriklerde tam puan alarak modelin hem güvenilirliğini hem de genellenebilirliğini güçlü biçimde ortaya koymaktadır.

Çizelge 4.32. GSE53757: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
FGS without FS [92]	0.96	0.96	0.95	0.97
FGS [92]	0.98	0.98	0.97	0.99
Önerilen yöntem (FS10)	0.90	0.90	0.90	0.90

Çizelge 4.32'de görüldüğü üzere, GSE53757 veri seti üzerinde yapılan çalışmalarda farklı öznitelik seçimi stratejileri ve sınıflandırma yöntemleri karşılaştırılmıştır. Literatürde yer alan FGS without FS yöntemi %0.96 doğruluk, %0.96 F1-Skor, %0.95 kesinlik ve %0.97 duyarlılık ile yüksek performans göstermektedir. Aynı çalışmada öznitelik seçimi entegre edilmiş FGS yöntemi ise %0.98 doğruluk, %0.98 F1-Skor, %0.97 kesinlik ve %0.99 duyarlılık ile en yüksek başarıyı elde etmiştir.

Bu çalışmada önerilen FS10 yöntemi, GSE53757 veri seti üzerinde %90 doğruluk, %90 F1-Skor, %90 kesinlik ve %90 duyarlılık ile dengeli ve tatmin edici bir performans sergilemiştir. Literatürdeki yöntemlere göre doğruluk oranı daha düşük olsa da, önerilen yöntemin istikrarlı metrikleri, özellikle sınıf dengesine duyarlı durumlar için uygulanabilir bir çözüm sunduğunu göstermektedir.

Çizelge 4.33. GSE63270: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
Multi Model Integration [96]	97.44	-	-	-
Önerilen yöntem (FS0, FS2, FS3, FS5, FS7)	1.00	1.00	1.00	1.00

Çizelge 4.33'de görüldüğü üzere, GSE63270 veri seti üzerinde gerçekleştirilen sınıflandırma çalışmaları, farklı yöntemlerin başarı oranlarını ortaya koymaktadır. Literatürde yer alan Multi Model Integration tabanlı yaklaşım, %97.44 doğruluk oranı ile dikkat çekici bir performans sergilemiştir. Ancak bu yöntemde diğer temel sınıflandırma metrikleri (F1-Skor, kesinlik, duyarlılık) raporlanmamıştır.

Bu çalışmada önerilen ve FS0, FS2, FS3, FS5 ve FS7 öznitelik seçim yöntemlerinin birleşiminden oluşan yaklaşım, tüm ölçütlere göre maksimum sınıflandırma başarısına ulaşmıştır. Tüm metriklerde elde edilen tam skorlar, önerilen yöntemin yalnızca doğruluk değil, aynı zamanda model dengesini ve güvenilirliğini de üst düzeye taşıdığını göstermektedir. Ayrıca, çoklu öznitelik seçim stratejilerinin sinerjik etkisi, modelin genellenebilirliğini artırarak literatürdeki yöntemlere kıyasla üstünlük sağlamıştır.

Çizelge 4.34. GSE45827: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
FGS without FS [92]	0.95	0.96	0.96	0.96
FGS [92]	0.97	0.98	0.98	0.98
Önerilen yöntem (FS0)	0.98	0.98	0.98	0.98

Çizelge 4.34'de, GSE45827 veri seti üzerinde yapılan sınıflandırma çalışmalarının performans karşılaştırması sunulmuştur. Literatürde yer alan FGS without FS ve FGS yaklaşımları sırasıyla %0.95 ve %0.97 doğruluk oranlarına ulaşmıştır. Bu yöntemler F1-skor, kesinlik ve duyarlılık gibi metriklerde de oldukça yüksek sonuçlar vermiştir. Özellikle FGS yöntemi, %0.98 F1-skor, %0.98 kesinlik ve %0.98 duyarlılık ile dikkat çekmektedir.

Bu çalışmada önerilen FS0 öznitelik seçimi yöntemi ile elde edilen model ise %0.98 doğruluk, %90.8 F1-skor, %0.98 kesinlik ve %0.98 duyarlılık değerleriyle, literatürdeki en başarılı yöntemle kıyaslanabilir hatta bazı yönlerden daha dengeli bir performans sergilemiştir. Bu sonuçlar, FS0 yönteminin hem öğrenme başarımı hem de genellenebilirliği açısından güçlü bir seçenek olduğunu ortaya koymaktadır.

Çizelge 4.35. GSE28497: Yöntem ve Literatür Karşılaştırması

Yöntem	Acc	F1-skor	Pre	Rec
WVCNN with Chi2+ SMOTE-Tomek [97]	0.99	0.99	0.99	0.99
FSVM with PCA [98]	0.97	-	-	-
FSVM without PCA [98]	0.88	-	-	-
Threshold segmentation [99]	0.85	-	-	-
Önerilen yöntem (FS0)	0.96	0.96	0.96	0.96

Çizelge 4.35'de, GSE28497 veri seti üzerinde yapılan sınıflandırma çalışmalarına ait performans karşılaştırmaları sunulmuştur. Literatürde yer alan yöntemler arasında en yüksek doğruluk oranı, Chi2 öznitelik seçimi ve SMOTE-Tomek örnekleme yöntemi ile desteklenen WVCNN (Weighted CNN) modeli tarafından elde edilmiştir. Bu yöntem %0.99 doğruluk, %0.99 F1-skor, %0.99 kesinlik ve %0.99 duyarlılık değerleriyle dikkat çekici bir başarı sergilemiştir. Bir diğer yöntem olan FSVM (Fuzzy support vector machine) modeli, PCA (Principal Component Analysis) kullanıldığında %0.97 doğruluk ile başarılı bir performans göstermiş, PCA kullanılmadan uygulandığında ise doğruluk oranı %0.88'ya düşmüştür. Ayrıca, eşik değeri tabanlı görüntü bölütleme teknikleri ile yapılan çalışmalarda ise yalnızca %0.85 doğruluk elde edilebilmiştir.

Bu çalışmada önerilen FS0 öznitelik seçimi yöntemi, %96 doğruluk, %96 F1-skor, %96 kesinlik ve %96 duyarlılık değerleriyle, özellikle FSVM ve eşik bölütleme yöntemlerinden daha yüksek ve dengeli sonuçlar ortaya koymuştur. WVCNN kadar yüksek olmasa da, önerilen yöntemin bu başarıyı çok daha sade ve klasik makine öğrenimi modelleriyle elde etmesi, yönteminin pratik uygulanabilirliğini ve etkili bir öznitelik seçimi sunduğunu göstermektedir. Ayrıca metriklerin birbirine yakın ve yüksek değerlerde olması, modelin dengeli ve kararlı bir sınıflandırma yeteneğine sahip olduğunu ortaya koymaktadır.

5. SONUÇ VE ÖNERİLER

Bu tez çalışması, gen ekspresyon verileri üzerinde makine öğrenimi tabanlı sınıflandırma modellerinin performansını artırmak amacıyla farklı öznitelik seçim yöntemleri ve sınıflandırma algoritmalarının etkinliğini derinlemesine incelemiştir. Özellikle yüksek boyutlu biyomedikal verilerde öznitelik seçimi sürecinin sınıflandırma başarısını nasıl etkilediği analiz edilmiş ve bu doğrultuda en etkili yöntemlerin belirlenmesi hedeflenmiştir. Çalışmada Ki-kare, Fisher Skor, Bilgi Kazancı, ReliefF ve Genetik Algoritma gibi çeşitli öznitelik seçim yöntemleri uygulanmış ve bu yöntemlerin başarımı farklı sınıflandırma algoritmaları ile birlikte istatistiksel testler yardımıyla karşılaştırılmıştır.

Gerçekleştirilen deneysel çalışmalar sonucunda elde edilen bulgular, önerilen öznitelik seçim yöntemlerinin sınıflandırma başarımına olan katkısını hem doğruluk oranları hem de diğer temel performans metrikleri açısından çarpıcı bir şekilde ortaya koymuştur. Analiz edilen 11 farklı gen ekspresyon veri setinde, uygulanan yöntemlerin çoğunluğu, literatürde daha önce raporlanmış tekniklere kıyasla üstün veya rekabetçi sonuçlar üretmiştir.

Özellikle GSE21510 veri setinde uygulanan FS0 (özellik seçimi uygulanmadan), FS2 (FKs), FS3 (IG) ve FS7 (Ki-kare + FKs + IG + ReliefF + B) yöntemlerinin her biri %100 doğruluk, %100 duyarlılık, %100 kesinlik ve %100 F1-Skor gibi ideal performans metriklerine ulaşmış, önerilen yöntemin sınıflandırma başarısının üst düzeyde olduğunu açıkça göstermiştir. Benzer şekilde, GSE44076 ve GSE63270 veri setlerinde de tüm temel metriklerde %100'lük başarı sağlanmış; bu sonuçlar, önerilen çoklu öznitelik seçim stratejilerinin yalnızca teorik olarak değil, aynı zamanda pratikte de son derece etkili ve güvenilir bir yapı sunduğunu ortaya koymuştur. GSE19804 veri setinde önerilen FS1 (Ki-kare) yöntemiyle %99 doğruluk ve tüm metriklerde %99 başarı elde edilmiş, bu değerlerle literatürde %95-96 doğruluk oranlarına ulaşabilmiş FCBF ve FGS gibi yöntemlerin de ötesine geçilmiştir. Sadece doğruluk oranı değil; aynı zamanda diğer performans metrikleri açısından da elde edilen bu yüksek ve dengeli değerler, önerilen yöntemin sadece tahmin gücünü değil, aynı zamanda sınıflar arası ayrım yeteneğini de son derece geliştirdiğini kanıtlamaktadır.

GSE14520_U133A veri setinde önerilen FS1, FS3 ve FS7 yöntemlerinin kombinasyonu, transfer learning ve Stacking ensemble gibi güncel ve güçlü yöntemlerle aynı doğruluk oranına (%97) ulaşmakla kalmamış, bu yöntemleri F1-Skor, duyarlılık ve kesinlik gibi diğer metriklerde de dengeleyerek daha istikrarlı bir performans ortaya koymuştur. Bu durum, önerilen yöntemin yalnızca literatürle rekabet etmekle kalmayıp, farklı veri setleri üzerinde genel geçer başarım sağlayan bir yapı sunduğunu göstermektedir. GSE70947 veri setinde literatürde %96 doğrulukla öne çıkan NSGA-II ve FCBF gibi tekniklere karşın, önerilen FS5 (Ki-kare + FKs + IG + ReliefF + GA + B) yöntemi %92 doğruluk ve %92 değerinde tüm metriklerle oldukça dengeli ve sağlam bir yapı sunmuştur. Bu durum, evrimsel ya da tekil yaklaşımların ötesinde, çoklu filtre ve sarmal stratejilerin birleşiminden oluşan yöntemlerin veriye daha uygun ve genellenebilir öznitelik alt kümeleri sunduğunu desteklemektedir.

Önerilen yöntem, yalnızca yüksek başarı gösterdiği veri setlerinde değil, bazı durumlarda literatürün gerisinde kaldığı veri setlerinde dahi metrik dengesi ve istikrar açısından dikkat çekici bir yapı ortaya koymuştur. Örneğin, GSE53757 veri setinde önerilen FS10 (IG + GA + B) yöntemi %90 doğruluk ile FGS yöntemine (%97.8) göre daha düşük görünse de, F1-Skor, duyarlılık ve kesinlik metriklerinin tümünde aynı seviyede (%90) denge sağlamış; bu da önerilen yöntemin sınıf dengesine duyarlı durumlar için uygun ve güvenilir bir alternatif sunduğunu ortaya koymuştur. GSE6919_U95B veri setinde ise önerilen FS7 yöntemi %77 doğruluk ve %79 kesinlik ile segmentasyon ya da derin öğrenme temelli modellere göre daha sade fakat tutarlı bir başarı sağlamış, özellikle düşük doğruluğun yüksek istikrarla dengelendiği durumlarda öznitelik seçim sürecinin model üzerindeki pozitif etkisini göstermiştir. Ayrıca GSE28497 veri setinde FS0 yöntemi, %96 doğruluk ile WVCNN gibi karmaşık modellerin ardından en iyi sonucu vermiş; buna rağmen bu başarıyı çok daha sade ve klasik makine öğrenimi algoritmalarıyla gerçekleştirmesi, önerilen yöntemin hesaplama verimliliği ve uygulama sadeliği açısından ciddi bir avantaj sunduğunu kanıtlamıştır.

Bulguların bir diğer dikkat çekici yönü, bilgi teorisi temelli yöntemlerin (özellikle IG ve ReliefF) çok sayıda veri setinde en başarılı veya en istikrarlı performansı sergilemesi olmuştur. Bu yöntemlerin tek başına veya çoklu kombinasyonlarla kullanılması, öznitelik sayısını azaltırken sınıflandırma performansını maksimum seviyede koruma avantajı sağlamıştır. Ayrıca, öznitelik seçim sürecinde dikkatli yapılandırılmayan bazı GA tabanlı

kombinasyonların (örneğin FS8 (Ki-kare + GA + B)), beklenildiği gibi yüksek performans göstermemesi, sarmal yaklaşımlarda parametrik uyumun ve optimizasyon stratejisinin ne kadar kritik olduğunu bir kez daha ortaya koymuştur.

Tüm bu sonuçlar ışığında, önerilen çoklu öznitelik seçim yaklaşımı, yüksek boyutlu ve dengesiz yapıya sahip biyomedikal veri setleri üzerinde sınıflandırma doğruluğunu, model genellenebilirliğini ve hesaplama verimliliğini birlikte maksimize eden bütüncül ve etkili bir çözüm sunmaktadır. Yöntemin sadece belirli veri kümelerinde değil, farklı kanser türlerine ait çok sayıda veri setinde tutarlı bir şekilde başarı göstermesi, bu tezin en güçlü bilimsel çıktılarından biri olarak değerlendirilmektedir.

Bu tez çalışmasında elde edilen bulgular, gen ekspresyon verileri üzerinde farklı öznitelik seçim yöntemlerinin ve sınıflandırma algoritmalarının başarılarını artırabileceğini göstermektedir. Ancak, bazı veri setlerinde elde edilen daha düşük doğruluk oranları ve belirli öznitelik seçim yöntemlerinin sınırlamaları göz önünde bulundurulduğunda, gelecekteki çalışmalarda aşağıdaki önerilere odaklanmak faydalı olacaktır. Öncelikle, evrensel ve genelleştirilebilir yaklaşımların oluşturulması önemlidir. Çalışma, belirli yöntemlerin bazı veri kümelerinde daha iyi çalıştığını göstermiştir. Bu nedenle, çoklu öznitelik seçim yöntemlerinin daha geniş bir yelpazede test edilmesi, genellenebilirliği artırarak daha sağlam ve evrensel çözümler sunabilir. Ayrıca, önerilen yöntemlerin daha büyük veri tabanlarında daha iyi değerlendirilmesi, sonuçların doğruluğunu iyileştirmeye katkıda bulunabilir. İkinci olarak, derin öğrenme ve diğer gelişmiş yöntemlerle entegrasyon sağlanabilir. Gelecekte, makine öğrenmesi ve derin öğrenme tabanlı yöntemlerin birleşimi, öznitelik seçim sürecini daha verimli hale getirebilir. Derin öğrenme algoritmalarının, özellikle sınıf dengesizliği olan veri setlerinde, daha iyi sonuçlar verebileceği düşünülmektedir. Bu nedenle, derin öğrenme tekniklerinin öznitelik seçim süreciyle entegrasyonu, model başarısını daha da artırabilir. Bir diğer öneri, genetik algoritmaların iyileştirilmesi konusudur. GA tabanlı yöntemlerin bazı veri setlerinde beklenen başarıyı göstermemesi, bu yöntemlerin parametre uyumunun ve optimizasyon süreçlerinin daha dikkatli bir şekilde ele alınması gerektiğini göstermektedir. Gelecekteki çalışmalar, genetik algoritmaların daha sofistike yöntemlerle optimize edilmesiyle, bu tür yaklaşımların performansını artırabilir.

Son olarak, otomatik öznitelik seçimi ve model geliştirme alanında da ilerlemeler sağlanabilir. Bu tezde elde edilen bulgular, öznitelik seçim sürecinin manuel olarak yapılmasının yanı sıra, bu sürecin otomatikleştirilmesinin önemini de ortaya koymaktadır. Otomatik öznitelik seçimi araçlarının geliştirilmesi, modelin başarısını artırırken, veri analizi sürecini de hızlandırabilir. Ayrıca, bu otomatik sistemlerin yapay zeka ve derin öğrenme algoritmaları ile entegrasyonu, daha verimli ve hızlı sonuçlar alınmasını sağlayabilir.

KAYNAKLAR

1. İnternet: Cancer. Web: https://www.who.int/news-room/fact-sheets/detail/cancer, Son Erişim Tarihi: 15.04.2025.

2. İnternet: What Is Cancer?. Web: https://www.cancer.gov/about-cancer/understanding/what-is-cancer, Son Erişim Tarihi: 22.04.2025.

3. İnternet: Cancer Today. Web: https://gco.iarc.fr/today/en/dataviz/maps-ranking?mode=ranking&key=total&cancers=15, Son Erişim Tarihi: 22.04.2025.

4. İnternet: Cancer Stat Facts: Common Cancer Sites. Web: https://seer.cancer.gov/statfacts/html/common.html, Son Erişim Tarihi: 22.04.2025.

5. İnternet: Cancer Classification. Web: https://training.seer.cancer.gov/disease/categories/classification.html, Son Erişim Tarihi: 22.04.2025.

6. Shen, L., Er, M. J., ve Yin, Q. (2022). Classification for high-dimension low-sample size data. *Pattern Recognition,* 130, 108828.

7. Yıldız, O., Tez, M., Bilge, H. Ş., Akcayol, M. A., ve Güler, İ. (2012). *Gene selection for breast cancer*. 20. Sinyal İşleme ve İletişim Uygulamaları Konferansı, Muğla, Türkiye.

8. Yıldız, O., Tez, M., Bilge, H. Ş., Akcayol, M. A., ve Güler, İ. (2013). Meme kanseri sınıflandırması için veri füzyonu ve genetik algoritma tabanlı gen seçimi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi,* 27(3), 21062.

9. Shen, Q., Shi, W. M., ve Kong, W. (2008). Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. *Computational Biology and Chemistry*, 32(1), 53-60.

10. Chandra, B., ve Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, 44(4), 529-535.

11. Hsu, H. H., Hsieh, C. W., ve Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, 38(7), 8144-8150.

12. Opitz, D., ve Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, 169-198.

13. Lorena, A. C., Costa, I. G., Spolaôr, N., ve de Souto, M. C. P. (2012). Analysis of complexity indices for classification problems: Cancer gene expression data. *Neurocomputing*, 75(1), 33-42.

14. Kaya, M. (2014). *Gen ifade verilerinde öznitelik seçimi ve sınıflandırma.* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

15. Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., Yang, T. S., Teng, N. C., Tan, K. P., ve Chang, K. S. (2014). Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinformatics*, 15(49), 24555567.

16. Shanab, A. A., Khoshgoftaar, T. M., ve Wald, R. (2014). *Evaluation of wrapper-based feature selection using hard, moderate, and easy.* IEEE Uluslararası Biyoinformatik ve Biyomühendislik Konferansı, Boca Raton, Florida, ABD.

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

18. Taşçı, A. (2016). *A comparative evaluation of feature selection algorithms for cancer classification through gene expression data.* Yüksek Lisans Tezi, İzmir Ekonomi Üniversitesi, İzmir.

19. Sserwadda, A. (2018). *Investigating the effect of different feature selection strategies for classification of gene expression signatures of tumor cells.* Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

20. Alanni, R., Hou, J., Azzawi, H., ve Xiang, Y. (2019). A novel gene selection algorithm for cancer classification using microarray datasets. *BMC Medical Genomics*, 12(10), 1755-8794.

21. Özkan, G. Y. (2019). *Meme kanseri hastalarının sağ kalım tahmini için öznitelik seçimi ve sınıflandırma algoritmalarının karşılaştırılması.* Yüksek Lisans Tezi, Eskişehir Teknik Üniversitesi, Eskişehir.

22. Candan, H., Durmuş, A., ve Harman, G. (2019). Genetik algoritma ve sınıflandırıcı yöntemler ile kanser tahmini. *Veri Bilimi*, 2(1), 30-34.

23. Arık, Ö. (2020). *Mikrodizi gen ifade verilerinde farklı öznitelik seçim yöntemleri ile sınıflama yöntemlerinin performanslarının değerlendirilmesi.* Doktora Tezi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

24. Emanet, S., Baydoğmuş, G. K., ve Demir, Ö. (2021). Öznitelik seçme yöntemlerinin makine öğrenmesi tabanlı saldırı tespit sistemi. *DÜMF MD*, 12(5), 743–755.

25. Wang, A., Liu, H., Yang, J., ve Chen, G. (2022). Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data. *Computers in Biology and Medicine*, 142(C), 105208.

26. Akcan, U. (2022). *Karınca koloni algoritması ile öznitelik seçimi.* Yüksek Lisans Tezi, Bursa Uludağ Üniversitesi, Bursa.

27. Rezaee, K., Jeon, G., Khosravi, M. R., Attar, H. H., ve Sabzevari, A. (2022). Deep learning-based microarray cancer classification and ensemble gene selection approach. *IET Systems Biology*, 16(3-4), 120-131.

28. Simangunsong, A. J. B., ve Hidayat, V. T. (2023). Feature reduction of lung cancer microarray data using mutual information selection and PyCaret-supported recursive feature elimination. *Nusantara Science and Technology Proceedings*, 2023(36), 1-6.

29. Hassan, M., ve Kaabouch, N. (2024). Exploring the impact of feature selection techniques on machine learning performance. *Applied Sciences*, 14(22), 10532.

30. Kılıç, M. (2023). *Makine öğrenmesi özellik seçimi (anova-boruta) ve sınıflandırma yaklaşımlarıyla pan-kanserde potansiyel mikroRNA biyobelirteçlerinin belirlenmesi.* Yüksek Lisans Tezi, Ege Üniversitesi, İzmir.

31. Shah, M., Gandhi, K., Patel, K. A., Kantawala, H., Patel, R., ve Kothari, A. (2023). *Theoretical evaluation of ensemble machine learning techniques.* 5. Uluslararası Akıllı Sistemler ve Yenilikçi Teknolojiler Konferansı, Tirunelveli, Hindistan.

32. Baltacıoğlu, B. N. (2024). *An investigation of feature selection methods for machine learning.* Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi, İzmir.

33. Yaqoob, A., Verma, N. K., ve Aziz, R. M. (2024). Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm. *Journal of Medical Systems*, 48(10), 38193948.

34. İnternet: CuMiDa. Web: https://sbcb.inf.ufrgs.br/cumida, Son Erişim Tarihi: 15.05.2025.

35. Dong, G., ve Liu, H. (2018). *Feature engineering for machine learning and data analytics*. (1. Baskı). Boca Raton: CRC Press, 418.

36. Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X., ve Dehmer, M. (2018). Feature selection of gene expression data for cancer classification using double RBF-kernel. *BMC Bioinformatics*, 19(396), 30373514.

37. Prastyo, P. H., Ardiyanto, I., ve Hidayat, R. (2020). *A review of feature selection techniques in sentiment analysis.* 6. Uluslararası Bilim ve Teknoloji Konferansı, Yogyakarta, Endonezya.

38. Kaya, A. (2015). *Bilgisayar destekli tanı sistemi ile akciğer nodüllerinin nitelendirilmesi.* Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

39. Almugren, N., ve Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*, 7, 78533-78548.

40. Dat, T. H., and Guan, C. (2007). *Feature selection based on Fisher ratio and mutual information analyses for robust brain computer interface.* IEEE Uluslararası Akustik, Konuşma ve Sinyal İşleme Konferansı, Honolulu, Hawaii, ABD.

41. Han, J., Kamber, M., ve Pei, J. (2013). *Data mining: Concepts and techniques* (3. Baskı). Waltham: Morgan Kaufmann.

42. Ladha, L., ve Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering (IJCSE)*, 3(5), 1787-1797.

43. Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F. X., ve Mewes, H. W. (2005). Gene selection from microarray data for cancer classification—a machine learning approach. *Computational Biology and Chemistry*, 29(1), 37-46.

44. Wah, Y. B., Ibrahim, N., Hamid, H. A., Rahman, S. A., ve Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Journal of Science and Technology*, 26(1), 329-340.

45. Bahassine, S., Madani, A., Al-Sarem, M., ve Kissi, M. (2020). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences*, 32(2), 225-231.

46. Kira, K., ve Rendell, L. A. (1992). *The feature selection problem: Traditional methods and a new algorithm.* 10. Ulusal Yapay Zeka Konferansı, Atlanta, ABD, 129-134.

47. Hung, L. P., Alfred, R., ve Hijazi, M. H. A. (2015). A review on feature selection methods for sentiment analysis. *Journal of Computational and Theoretical Nanoscience*, 21(10), 2952-2956.

48. Katoch, S., Chauhan, S. S., ve Kumar, V. (2021). A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 80, 8091–8126.

49. Babaoglu, İ., Findik, O., ve Ülker, E. (2010). A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. *Expert Systems with Applications*, 37(4), 3177-3183.

50. Kennedy, J., andve Eberhart, R. (1995). *Particle swarm optimization.* IEEE Uluslararası Sinir Ağları Konferansı, 4, 1942-1948.

51. Jiménez-Cordero, A., Morales, J. M., ve Pineda, S. (2021). A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. *European Journal of Operational Research*, 293(1), 24-35.

52. Jovic, A., Brkić, K., ve Bogunovic, N. (2015). A review of feature selection methods with applications. 38. Uluslararası Bilgi ve İletişim Teknolojileri Konvansiyonu, Opatija, Hırvatistan.

53. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-288.

54. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2. Baskı). New York: Springer.

55. Zou, H., ve Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301–320.

56. Sanz, H., Valim, C., Vegas, E., Oller, J. M., ve Reverter, F. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics*, 19(432), 30453885.

57. Guyon, I., Weston, J., Barnhill, S., ve Vapnik, V. (2002). Gene selection for cancer classification using Support Vector Machines. *Machine Learning*, 46, 389–422.

58. Anwar, H., Qamar, U., ve Qureshi, A. W. M. (2014). Global optimization ensemble model for classification methods. *The Scientific World Journal*, 2014, 313164.

59. Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40, 139–157.

60. Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123-140.

61. Bühlmann, P., ve Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4), 927–961.

62. Freund, Y., ve Schapire, R. E. (1996). *Experiments with a new boosting algorithm.* 13. Uluslararası Makine Öğrenmesi Konferansı, Bari, İtalya, 148-156.

63. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

64. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241-259.

65. Smyth, P., ve Wolpert, D. H. (1997). Stacked density estimation. *NIPS'97: Proceedings of the 11th International Conference on Neural Information Processing Systems*, 668-674.

66. Ma, Z., Wang, P., Gao, Z., Wang, R., ve Khalighi, K. (2018). Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. *PLoS One*, 13(10), e0205872.

67. Xiong, Y., Ye, M., ve Wu, C. (2021). Cancer classification with a cost-sensitive naive Bayes stacking ensemble. *Computational and Mathematical Methods in Medicine*, 2021, 5556992.

68. Soofi, A. A., ve Awan, A. (2017). Classification techniques in machine learning: Applications and issues. *Journal of Basic & Applied Sciences*, 13, 459-465.

69. Alnuaimi, A. F. A. H., ve Albaldawi, T. H. K. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, 97(4), 00133.

70. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

71. Bénard, C., Biau, G., da Veiga, S., ve Scornet, E. (2021). *Interpretable Random Forests via Rule Extraction.* 24. Uluslararası Yapay Zeka ve İstatistik Konferansı, 130, 937-945.

72. Geurts, P., Ernst, D., ve Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3-42.

73. Chen, T., ve Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. 22. ACM SIGKDD Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansı, New York, ABD.

74. Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using extreme gradient boosting (XGBoost). *Educational and Psychological Measurement*, 79(5), 931-961.

75. Freund, Y., ve Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society of Artificial Intelligence*, 14(5), 771-780.

76. Wang, F., Li, Z., He, F., ve Nie, F. (2019). Feature learning viewpoint of AdaBoost and a new algorithm. *IEEE Access*, 7, 149890-149899.

77. Siegel, S. (1988). *Nonparametric statistics for the behavioral sciences* (2. Baskı). New York: McGraw-Hill.

78. Jia, B. B., Liu, J. Y., ve Zhang, M. L. (2025). Pairwise statistical comparisons of multiple algorithms. *Frontiers of Computer Science*, 19, 1912372.

79. De Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J., ve Moguerza, J. M. (2022). General Performance Score for classification problems. *Applied Intelligence*, 52, 12049–12063.

80. Saito, T., ve Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 10(3), e0118432.

81. İnternet: scikit-learn. Web: https://scikit-learn.org/stable/, Son Erişim Tarihi: 15.04.2025.

82. İnternet: imbalanced-learn documentation. Web: https://imbalanced-learn.org/stable/, Son Erişim Tarihi: 15.04.2025.

83. İnternet: Matplotlib: Visualization with Python. Web: https://matplotlib.org/, Son Erişim Tarihi: 15.04.2025.

84. İnternet: seaborn: statistical data visualization. Web: https://seaborn.pydata.org/, Son Erişim Tarihi: 15.04.2025.

85. İnternet: scikit-posthocs. Web: https://scikit-posthocs.readthedocs.io/en/latest/, Son Erişim Tarihi: 15.04.2025.

86. İnternet: Yellowbrick: Machine Learning Visualization. Web: https://www.scikit-yb.org/en/latest/, Son Erişim Tarihi: 15.04.2025.

87. İnternet: skrebate. Web: https://pypi.org/project/skrebate/, Son Erişim Tarihi: 15.04.2025.

88. Panda, M. (2020). Elephant search optimization combined with deep neural network for microarray data analysis. *Journal of King Saud University - Computer and Information Sciences*, 32(8), 940-948.

89. Cilia, N. D., De Stefano, C., Fontanella, F., Raimondo, S., ve Scotto di Freca, A. (2019). An experimental comparison of feature-selection and classification methods for microarray datasets. *Information*, 10(3), 109.

90. Barbachan e Silva, M., Narloch, P. H., Dorn, M., ve Ó Broin, P. (2021). *Optimisation of cancer status prediction pipelines using bio-inspired computing.* IEEE Evrimsel Hesaplama Kongresinde sunuldu. Kraków, Polonya.

91. Medjahed, S. A., Saadi, T. A., Benyettou, A., ve Ouali, M. (2017). Kernel-based learning and feature selection analysis for cancer diagnosis. *Applied Soft Computing*, 51, 39-48.

92. Bohrer, J. da S., ve Dorn, M. (2024). Enhancing classification with hybrid feature selection: A multi-objective genetic algorithm for high-dimensional data. *Expert Systems with Applications*, 255(Part A), 124518.

93. Ead, W. M., Abdelazim, M. A., ve Nasr, M. M. (2022). Feedforward Deep Learning Optimizer-based RNA-Seq Women's cancers Detection with a hybrid Classification Models for Biomarker Discovery. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 13(12), 1-10.

94. Haris, A., Simon, N., ve Shojaie, A. (2022). Generalized Sparse Additive Models. *Journal of Machine Learning Research*, 23(70), 1-56.

95. Alrefai, N., Ibrahim, O., Shehzad, H. M. F., Altigani, A., Abu-ulbeh, W., Alzaqebah, M., ve Alsmadi, M. K. (2023). An integrated framework based deep learning for cancer classification using microarray datasets. *Journal of Ambient Intelligence and Humanized Computing,* 14, 2249–2260.

96. Mahmoud, A., ve Takaoka, E. (2025). An enhanced machine learning approach with stacking ensemble learner for accurate liver cancer diagnosis using feature selection and gene expression data. *Healthcare Analytics*, 7, 100373.

97. Wolin, K. Y., Carson, K., ve Colditz, G. A. (2010). Obesity and cancer. *The Oncologist*, 15(6), 556-565.

98. Seitz, H. K., ve Stickel, F. (2007). Molecular mechanisms of alcohol-mediated carcinogenesis. *Nature Reviews Cancer*, 7(8), 599-612.



Gazili olmak ayrıcalıktır