DNA MİKRODİZİ ANALİZLERİ İLE MEME KANSERİ HASTALARINDA PROGNOZUN GENETİK ALGORİTMA KULLANILARAK BELİRLENMESİ

OKTAY YILDIZ

DOKTORA TEZİ
ELEKTRONİK VE BİLGİSAYAR EĞİTİMİ

GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ

MAYIS 2012
ANKARA

Oktay YILDIZ tarafından hazırlanan DNA MİKRODİZİ ANALİZLERİ İLE MEME KANSERİ HASTALARINDA PROGNOZUN GENETİK ALGORİTMA KULLANILARAK BELİRLENMESİ adlı bu tezin Doktora tezi olarak uygun olduğunu onaylarım.

Prof.Dr. İnan GÜLER
Tez Yöneticisi

Bu çalışma, jürimiz tarafından oy birliği ile ELEKTRONİK VE BİLGİSAYAR EĞİTİMİ Anabilim Dalında Doktora tezi olarak kabul edilmiştir.

Başkan: : Prof.Dr. Ömer Faruk BAY

Üye : Prof.Dr. İnan GÜLER

Üye : Prof.Dr. Hadi GÖKÇEN

Üye : Doç.Dr. Mesut TEZ

Üye : Doç.Dr. Suat ÖZDEMİR

Tarih : ...../……/……

Bu tez, Gazi Üniversitesi Bilişim Enstitüsü tez yazım kurallarına uygundur.

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Oktay YILDIZ

DNA MİKRODİZİ ANALİZLERİ İLE MEME KANSERİ HASTALARINDA PROGNOZUN GENETİK ALGORİTMA KULLANILARAK BELİRLENMESİ

(Doktora Tezi)

Oktay YILDIZ

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Mayıs 2012

ÖZET

Meme kanseri ölüme neden olabilen ciddi bir rahatsızlıktır. Erken teşhis hastalığın tedavisinde önemli rol oynar. Son yıllarda mikrodizi teknolojisi kanser teşhisinde sıklıkla kullanılmaktadır. Mikrodizi gen ifade bilgisi ölçmeye yarayan bir araçtır. Mikrodizi verisi genellikle on binlerce gen bilgisi ve az sayıda örneklem içermektedir. Ancak bu genlerin pek çoğu ilgisiz ve klinik teşhis için gereksiz bilgiden ibarettir. Yüksek boyut ve aşırı eğitim riski sebebiyle makine öğrenme teknikleri ile bu verilerin sınıflandırılması zordur. Bu sebeple nitelik seçme gen analizlerinde çok önemlidir. Nitelik seçme, sınıflandırma başarısını en iyi yapacak niteliklerin seçilmesi işlemidir.

Bu çalışmada meme kanseri hastalığında etkin rol oynayan genlerin belirlenmesi için filtreleme yöntemi ve genetik algoritma tabanlı yeni bir gen seçme metodu önerilmektedir. Yapılan çalışma iki aşamadan oluşmaktadır: İlk aşamada filtreleme yöntemi ile gen ifade verisi indirgenmiş, ikinci aşamada genetik algoritma ile meme kanserinde etkin rol alan genlerin tespiti gerçekleştirilmiştir. Destek vektör makinesi, genetik algoritma için

uygunluk fonksiyonu olarak kullanılmıştır. Yapılan çalışmada belirlenen 7 gen ile sınıflandırma doğruluğu %96,15 olarak elde edilmiştir.

Bilim Kodu : **702.1.019**
Anahtar Kelime : **veri madenciliği, biyoinformatik, nitelik seçme, veri füzyonu, genetik algoritma, meme kanseri, destek vektör makinesi**
Sayfa Adedi : **98**
Tez Yöneticisi : **Prof.Dr. İnan GÜLER**

DETERMINATION OF PROGNOSIS IN BREAST CANCER PATIENTS FROM DNA MICROARRAY ANALYSIS USING GENETIC ALGORITHM

(PhD Thesis)

Oktay YILDIZ

GAZİ UNIVERSITY
INFORMATICS INSTITUTE
May 2012

ABSTRACT

Breast cancer is a serious disease that can cause death. Early diagnosis of breast cancer has been playing very important role on treatment of the disease. Therefore, it is important to find the genes that are relevant to a diagnosis. Recently, microarray technology has been widely used in cancer diagnosis. A microarray is a tool for analyzing gene expression. Microarray data usually contain thousands of genes and a small number of samples. Although, most of them are irrelevant or insignificant to a clinical diagnosis. It is very difficult to obtain a satisfactory classification result by machine learning techniques because of both the curse-of dimensionality problem and the over-fitting problem. Feature selection plays a crucial role in microarray analysis. Feature selection is the process of choosing the most discriminative features so as to enable the classifier to perform better.

In this work, a new feature selection method for breast cancer classification based on filter method and genetic algorithm is presented. The study consists of two steps: In the first step, the dimensionality of the gene expression dataset was reduced with filter method and the second step, significant genes have been identified with genetic algorithm. SVM was used for fitness function in genetic

programming. In this study the classification accuracy rate was obtained as 96.15% when using the selected 7 genes.

Science Code : **702.1.019**
Key Words : **data mining, bioinformatics, feature selection, data fusion, genetic algorithm, breast cancer, support vector machine**
Page Number : **98**
Adviser : **Prof.Dr. İnan GÜLER**

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren danışmanım Prof. Dr. İnan GÜLER'e, Prof.Dr. M.Ali AKCAYOL'a, tez izleme komitemde yer alarak katkı sağlayan Prof.Dr. Ömer Faruk BAY ve Doç.Dr. Suat ÖZDEMİR'e, manevi destekleriyle beni hiçbir zaman yalnız bırakmayan eşime ve aileme sonsuz teşekkürler.

İÇİNDEKİLER

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

SİMGELER VE KISALTMALAR

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

Simgeler	**Açıklama**
μ	Aritmetik ortalama
σ	Standart sapma
ξ	Gevşek değişken

Kısaltmalar	**Açıklama**
DNA	Deoksiribo Nükleik Asit
cDNA	Tamamlayıcı DNA
RNA	Ribo Nükleik Asit
tRNA	Taşıyıcı RNA
rRNA	Ribozomal RNA
mRNA	Mesaj taşıyan RNA
TNM	Tümör boyutu (T), aksiler lenf nodlarına yayılım (N) ve uzak bölgelere yayılım (M)
AJCC	American Joint Committee on Cancer
FKS	Fisher Korelasyon Skorlama
WTS	Welch t-statistic
GA	Genetik Algoritma
DVM	Destek vektör makinesi
KNN	k-En Yakın Komşu
DP	Doğru Pozitif
DN	Doğru Negatif
YP	Yanlış Pozitif
YN	Yanlış Negatif

Kısaltmalar	Açıklama
ROC	Alıcı İşletim Karakteristiği (Receiver Operating Characteristics)
AUC	Eğri altında kalan alan (Area Under the Curve)

1. GİRİŞ

Modern moleküler genetiğin temelini oluşturan Deoksiribo Nükleik Asit (DNA) üzerine ilk çalışmalar Fried Miescher tarafından 1868 yılında yapılmaya başlanmıştır. DNA'nın genetik bilgi taşıdığına dair ilk kanıtlar ise 1944 yılında O. Theodore Avery, Colin Macleod ve Marclyn McCarty tarafından bulunmuştur. 1952 yılında Alfred D.Hersey ve Martha Chase tarafından gerçekleştirilen deneyler, canlı hücrelerde genetik bilginin kromozomların önemli bir bileşeni olan DNA tarafından taşındığını ortaya koymuştur.

Bir organizmada meydana gelen biyolojik etkinlikler DNA'dan Ribo Nükleik Asite (RNA) ve RNA'dan da proteine aktarılan genetik bilgi akışıyla gerçekleştirilir. Hücreler arasında genetik bilgi genler aracılığıyla taşınır. Bireyin sahip olduğu özellikler *gen* olarak adlandırılan kalıtım elementleri ile ebeveynlerinden aktarılmaktadır.

Genlerin kalıtımda önemli rol oynadığını 1866'da Gregor Mendel tespit etmiştir. Ancak günümüzde DNA halen tam olarak çözümlenememiştir. Bunun yanı sıra moleküler biyolojide çok önemli gelişmeler yaşanmıştır. Bunlardan biri de mikro DNA çipleridir (microarray). Mikro DNA çipleri ile farklı organizmaların genom yapıları karşılaştırılabilmekte ve bu sayede de belirli hastalıklara neden olan genler ve bu genlerin ifadeleri elde edilebilmektedir.

Biyoçip ya da DNA çipleri olarak da adlandırılan mikrodiziler, gen ifade ölçümünde kullanılan en önemli teknolojilerden birisidir. Gen ifade verileri, sağlıklı ve hasta dokular arasındaki farklılıkların tespitinde önemli rol oynar. Günümüzde kanser gibi genetik hastalıkların temelinde yatan genetik faktörler yine bu teknoloji ile araştırılmaktadır.

Türk Kanser Araştırma ve Savaş Kurumu Derneği'nin tespitlerine göre dünyada her yıl yaklaşık 13 milyon insan kansere yakalanmakta ve bunların yaklaşık 8 milyonu bu nedenle hayatını kaybetmektedir. Meme kanseri de kadınlar arasında ölüme neden

olan kanser türlerinde ilk sırada yer almaktadır. GLOBOCAN 2008 verilerine göre meme kanserinden ölüm oranı Türkiye'de %17,6 Amerika'da %12,7 ve Çin'de %6,1 olarak açıklanmıştır [1]. Meme kanserinin erkeklerde görülme sıklığı, kadınlarda görülme oranının %1'i kadardır. Pek çok kanser türünde olduğu gibi meme kanserinde de hastalığın sebebi tam olarak anlaşılamamıştır. Ancak ailevi kalıtımın meme kanseri riskini arttırdığı bilinmektedir.

Hücresel düzeyde genetik bir hastalık olduğu bilinen meme kanserinde halen hangi genlerin bu hastalığa neden olduğu tam olarak bilinememektedir. Ancak bilinen şudur ki bu hastalığı açıklamak için tek geni ilgilendiren basit modeller yetersiz kalmaktadır. Birden fazla gen, kanser gelişim riskini arttırmakta, çevresel faktörler bu işi daha da karmaşık hale getirmektedir.

Son 10 yılda kanser genetiği ile ilgili yapılan çalışmalarda önemli gelişmeler kaydedilmiştir. Genetik çalışmalar ve yeni teknolojiler kanser için umut olsa da henüz bu çalışmalar kanser tedavisinde etkin olarak rol alamamaktadır. Günümüzde kanser tedavisi için halen 1960'lı yıllarda belirlenen cerrahi yöntemler, radyoterapi ve kemoterapi yöntemleri uygulanmaktadır. Genetik faktörler, hastalığın tanı ve tedavi sürecinde kullanılamamaktadır.

Meme kanserinin tarama ve tanısında kullanılan yöntemler, diğer kanser türlerinde olduğu gibi; fiziksel muayene, görüntüleme teknikleri ve kanda tümör belirleyici maddelerin (marker) taranması şeklindedir.

Erken ve doğru teşhis, hastalığın tedavisi açısından çok önemlidir. Bu sebeple hastalığın doğru evrelendirilmesi gerekir. Bu aşamada hem klinik hem de patolojik evrelendirme önem kazanmaktadır. Ayrıca bu konuda deneyimli uzmanlara da ihtiyaç duyulmaktadır.

TNM evrelemesi meme kanseri hastalarında tedaviye yön veren önemli bir araçtır. Hastalığın yayılımı ve ciddiyeti hakkında bilgi veren TNM evreleme sisteminde; tümör boyutu (T), aksiler lenf nodlarına yayılım (N) ve uzak bölgelere yayılım (M)

olmak üzere üç kriter kullanılır. TNM evreleme sistemi maalesef hastalığı doğru sınıflandırmakta yetersiz kalmakta ve dolayısıyla hastaya uygulanacak tedavi doğru olarak belirlenememektedir. Ayrıca kanserin yanlış evrelendirilmesi, hastanın çoğunlukla rahatsızlık veren gereksiz tedavileri görmesine ve hatta bazı cerrahi müdahalelere maruz kalmasına neden olmaktadır. Yeni teknikler geliştirildikçe hastalığın tespit edilmesi ve kanserin doğru evrelendirilmesinde başarı artar. Bunun sonucunda da daha etkin ve kurtarıcı tedaviler yapılabilir.

Genetik faktörlerin belirlenmesi; hastalığın tanı ve tedavi aşamasında, hastalığın daha erken teşhis edilmesinde, doğru evrelendirilerek daha etkin tedavinin uygulanmasında, yan etkisi az yeni ilaçların geliştirilmesi ve pahalı olan mikrodizi deney maliyetlerinin düşürülmesinde önemlidir.

Bugün genetik alanında yapılan çalışmalar ve teknolojik gelişmeler daha fazla verinin elde edilmesini sağlamıştır. Ancak bu durum beraberinde analiz problemlerini de getirmiştir. Ortaya çıkan bu büyük miktardaki verinin analizi için bilgisayar teknolojisi kaçınılmaz olmuştur. Veri içinde önceden bilinmeyen ve potansiyel olarak faydalı bilgilerin elde edilmesi olarak tanımlanan veri madenciliği, özellikle gen dizi analizleri, moleküler yapı analizleri ve moleküler fonksiyon analizleri için ciddi faydalar sağlar.

Bu çalışmada, meme kanseri hastalarına ait gen ifade verileri kullanılarak veri madenciliği teknikleri ile meme kanserinde etkin genlerin tespiti gerçekleştirilmiştir. Etkin genlerin bulunması için genetik algoritma ve veri füzyonu tabanlı yeni bir metot önerilmiştir. Genetik algoritma, nitelik seçme işlemlerinde başarılı sonuçlar vermesinden dolayı tercih edilmiştir.

Önerilen metot üç adımdan oluşmaktadır: İlk adımda, veri kümesinde yer alan hatalı ve eksik kayıtlar düzeltilerek veri madenciliği için ön hazırlık yapılmıştır. İkinci adım, iki aşamadan oluşmaktadır: İlk aşamada, genetik algoritma için başlangıç veri kümesi belirlenmiştir. Bunun için Fisher Korelasyon Skorlama, t-Skor ve WTS (Welch t-statistic) ile gen skorlaması yapılmıştır. Daha sonra her bir skorlama

yöntemi ile tespit edilen gen veri kümelerinde ayrı ayrı genetik algoritma ve destek vektör makinesi ile sınıflandırma başarısı en yüksek olan genler belirlenmiştir. Bu aşama, nihai gen seçimi için yüksek boyuta sahip gen ifade verisinin indirgenerek, başlangıç veri kümesinin belirlenmesi aşamasıdır. Elde edilen bu yeni veri kümesi, üç ayrı filtreleme yöntemiyle elde edilen sonuçların birleştirilmesi ve genetik algoritma ile belirlenmiştir. İkinci aşama meme kanserinde etkin genlerin belirlenmesi aşamasıdır. Bu amaçla Genetik Algoritma ile gen seçimi gerçekleştirilmiştir. Genetik algoritma için uygunluk fonksiyonu olarak Destek Vektör Makinesi kullanılmıştır. Üçüncü ve son adımda, belirlenen gen alt kümelerin sınıflandırıcı performansları, 10 kat çapraz doğrulama yöntemi ile test edilmiştir.

Filtreleme yöntemleri, biyoinformatik alanında gen seçme işlemlerinde sıklıkla kullanılmasına rağmen, uygun gen alt kümesini vermeyi garanti etmemekte ve gen seçme işlemlerinde tek başına yeterli olamamaktadır. Önerilen metot, genetik algoritma ve filtreleme yöntemlerinin birlikte kullanıldığı hibrit bir yaklaşım sunmaktadır. Genetik algoritmanın arama uzayını daraltmak için farklı filtreleme yöntemleri ile gen ifade verisi indirgenmiş ve bu işlemin ardından genetik algoritma ile etkin genlerin seçimi gerçekleştirilmiştir. Belirlenen genler ile meme kanseri hastalığı %96,15 gibi yüksek doğrulukta sınıflandırılabilmiştir.

Tezin ikinci bölümünde, biyoinformatik alanında veri madenciliği ile gerçekleştirilen çalışmalar hakkında literatür taraması yapılmıştır. Üçüncü bölümde, biyolojik verilerin bilgisayar destekli analiziyle ilgilenen biyoinformatik alanından, bu alanda başarılı bir şekilde kullanılan veri madenciliği yöntemlerinden, boyut indirgemeden ve genetik algoritmadan bahsedilmiştir. Dördüncü bölümde, meme kanseri, tanı ve evrelendirme sisteminden, gen ve genom yapısından bahsedilmiştir. Beşinci bölümde, meme kanserinde etkin genlerin belirlenmesi amacıyla geliştirilen metot anlatılmış, ayrıca elde edilen deneysel bulgulardan bahsedilmiştir. Altıncı bölümde sonuç ve değerlendirme yer almaktadır.

2. LİTERATÜR ARAŞTIRMASI

Günümüzde hücrelerin fonksiyonlarını anlamak ve hastalıkların temelinde yatan genetik faktörlerin belirlenmesi amacıyla pek çok çalışma yapılmaktadır. Ortaya çıkan verinin bilgisayar desteği olmaksızın analizi mümkün değildir. Bu verilerin analizi için veri madenciliği sıklıkla kullanılmaktadır. Bu bölümde biyoinformatik alanında veri madenciliği ile gerçekleştirilmiş çalışmalardan bahsedilmiştir.

Gloria Phillips ve arkadaşları [2] çalışmalarında, akciğer kanseri ile ilgili veriler üzerinde farklı veri madenciliği tekniklerini uygulamışlar, sonuçları karşılaştırmışlar, bu tekniklerin birbirlerine karşı güçlü taraflarını ortaya koymuşlardır. Ortaya konulan modeli büyük ve karmaşık veri kümelerinde uygulamışlar, ayrıca sağlık sigorta sistemi gibi genel uygulamalarda önceden tahminlerde bulunma amacıyla uygulanabilirliğini göstermişlerdir. Yapılan çalışmada, özellikle karar ağaçları ve yapay sinir ağları kullanıldığında çok başarılı sonuçlar elde edildiği gösterilmiştir. Yazarlar, medikal karar mekanizmalarında ve sosyal politikalar üretiminde büyük veri kümelerinin analizinin önemli bir yer tuttuğunu ifade etmektedirler. Klinik ve demografik bilgilerin tıbbi kararlar için önemini göstermişlerdir. Yazarlar bu çalışmada: Akciğer kanseri hastalarının demografik karakteristikleri, sosyoekonomik değerleri, etnik yapıları, tıbbi hikâyeleri ve sağlık hizmetlerine erişimi değerlendirmiştir. Akciğer kanseri ile ilgili elde edilen veriler veri madenciliği teknikleriyle analiz edilmiştir. Önerilen yeni yöntemin büyük sağlık sistemlerinde uygulanabilirliği gösterilmiştir.

Chun-Lang Chang ve Chih-Hao Chen [3] çalışmalarında, dermatoloji alanında deri hastalığı veri kümesinde yapay sinir ağları ve karar ağaçları ile sınıflandırma yapmışlardır. Çalışma sonucunda; önerilen yapay sinir ağı modeli ile %92,62, karar ağacı modeliyle ise %80,33 oranında doğru sınıflandırma yapılabildiğini göstermişlerdir.

Kuo-Sheng Lina ve Chen-Fu Chien [4] çalışmalarında, veri madenciliğinde önemli bir yeri olan nitelik seçme ve boyut indirgeme problemi çözümü için bir algoritma

önermektedirler. Bu amaçla korelasyon tabanlı bir yaklaşım önermişlerdir. Önerilen algoritma, Stanford Üniversitesinden alınan meme kanseri veri kümesinde denenmiş ve bu algoritma ile başarılı sonuçlar elde edilmiştir.

Walker ve arkadaşları [5], Alzheimer hastalığı üzerine veri madenciliği çalışması yapmışlardır. Bu çalışmada, Alzheimer görülen hastalar ile sağlıklı insanların yer aldığı iki ayrı mikrodizi veri kümesi kullanmışlardır. Alzheimer ile ilişkili genleri bulmak için üç ayrı teknik uygulamışlar ve çalışma sonucunda daha önce Alzheimer ve diğer nörolojik hastalıklarla ilişkili 17 genin yanı sıra toplamda 50 genin hastalık üzerinde etkili olduğunu tespit etmişlerdir. Tespit edilen bu 50 gen, daha önce bulunmuş ancak Alzheimer ile ilişkisi belirlenmemiş 20 gen ve karakterize edilmemiş EST (Expressed Sequence Tags) den oluşmaktadır.

Yuehui Chen ve Yaou Zhao [6] çalışmalarında, mikrodizi gibi büyük boyutlu verilerde sınıflandırma ve kümeleme performansını arttırmak için yeni bir metot önermişlerdir. Bu amaçla öncelikle ilgisiz niteliklerin atılması için korelasyon analizi ve Fisher oranı kullanmışlardır. Önerdikleri metodu kan kanseri ve kolon kanseri veri kümesinde uygulamışlardır. Sınıflandırma için EDA (Estimation of Distribution Algorithms) algoritmasını kullanmışlar ve diğer sınıflandırma algoritmalarıyla karşılaştırdıklarında önerdikleri metodun çok daha başarılı olduğunu tespit etmişlerdir.

Ming-Hseng Tseng ve Hung-Chang Liao [7], meme kanseri ile DNA virüsleri arasındaki ilişkiyi ortaya koymak amacıyla bir çalışma yapmışlardır. Meme kanserinde etkin faktörleri belirlemek için varyans analizi (ANOVA) ve bilgi ölçme metotlarını kullanmışlar, genetik algoritma tabanlı veri madenciliği uygulamışlardır. Çalışma sonucunda beş DNA virüsünün - HSV (Herpes Simplex Virus) EBV (Epstein–Barr Virus), CMV (Cyto Megalo Virus), HPV (Human Papillomavirus) ve HHV (Human Herpesv Virus) - meme kanseri oluşumunda etkili olduğunu ortaya koymuşlardır. HPV virüsünün ise meme kanserinin oluşumunda daha az etkili olduğunu belirlemişlerdir.

Jiyuan An ve Yi-Ping Phoebe Chen [8], gen ifade verilerinin geleneksel makine öğrenme metotları ile başarılı bir şekilde sınıflandırılamadığını ifade etmişler ve bu amaçla yeni bir metot önermişlerdir. Bu metotta, gen ifade veri kümesinin sınıflandırılmasında her bir gen belirli bir grup ile kısıtlanarak kullanılacak kural grubu oluşturulmuştur. Önerilen algoritma, olabilecek tüm kombinasyonları denemektedir. Kural grubu istenen sonucu veriyor ise gen doğru gruplandırılmış aksi halde yanlış gruplandırılmış anlamına gelir. Uygulanan bu metodun geleneksel makine öğrenme algoritmalarına göre daha başarılı sonuçlar verdiği görülmüştür.

Armañanzas ve arkadaşları [9], gen etkileşimlerini tanımlamak üzere yeni bir algoritma önermişler ve çalışmalarında Bayes ağlarını kullanmışlardır. Önerilen algoritma gen ifade verilerinden başarılı bir şekilde gen etkileşimlerini tespit edebilmektedir. Tespit edilen bu gen etkileşimleri, nedeni bilinmeyen pek çok hastalığın nedenini anlama noktasında bu alanda çalışan uzman kişilere yardımcı olabilecek niteliktedir. Önerilen algoritma fenotip değişkenler kullandığı için denetimli öğrenme modeli içinde sınıflandırılmıştır.

Resul Daş ve arkadaşları [10], kalp kapakçığı hastalığının teşhisi için veri madenciliği yöntemlerini kullanarak yeni bir metot önermişlerdir. Bu amaçla SAS 9.1.3 veri medenciliği yazılımını kullanmışlar, önerdikleri metot ile 215 denek üzerinde bir çalışma yapmışlar ve çalışma sonucunda %97,4 oranında doğru sınıflandırma yapabildiklerini ortaya koymuşlardır. Sınıflandırmalarında özellik seçimi için üç ayrı çok katmanlı ileri-beslemeli yapay sinir ağı kullanmışlardır ve bunun da daha doğru değerler elde etmede önemli rol oynadığını görmüşlerdir.

Nikolaos Giannakeasa ve Dimitrios Fotiadis [11], gen ifade ölçümü için kullanılan mikrodizi görüntü analizi üzerine çalışmışlardır. Mikrodizi görüntü analizi, çok büyük miktarda biyolojik verinin çözümlenmesi için kullanılan bir araçtır. Bu çalışmada otomatik mikrodizi görüntü analizi yapacak bir metot önerilmektedir. Önerilen metot grid ve bölümlere ayırma olmak üzere iki aşamadan oluşmaktadır. Mikrodizi görüntüsü, şablon eşleşmesi ile ön işlemden geçirilmektedir. Daha sonra ifade edilmeyen spotlar tespit edilmekte, Voronoi diyagramı ile grid

belirlenmektedir. Bölümleme için k-ortalama ve FCM (Fuzzy C means) kümeleme algoritmaları kullanılmıştır. Önerilen metot Stanford Microarray Database (SMD) veri kümesi üzerinde başarılı bir şekilde uygulanmıştır.

Xiaobai Zhang ve arkadaşları [12], gen ifade verilerinde yer alan hatalı kayıtların düzeltilmesi üzerine çalışma yapmışlardır. Mikrodizi analizlerinde hatalı değerler sonucu olumsuz etkilemekte bu nedenle doğru tahmin araçlarıyla hatalı değerlerin düzeltilmesi gerekmektedir. Bu problemin çözümü için SLLSimpute (Sequential Local Least Squares imputation) metodunu önermektedirler. Bu metotla hatalı değerler doğruya en yakın bir şekilde tahmin edebilmekte ve bu tahmin için otomatik parametre seçim algoritması kullanılmaktadır. Yapılan çalışma sonunda önerilen bu metodun daha önce var olan metotlara göre daha başarılı sonuçlar verdiği görülmüştür.

Horng ve arkadaşları [13], biyomarker genlerin belirlenmesi için üç aşamadan oluşan yeni bir yöntem önermişlerdir. Bu yöntemin veri girişi olarak adlandırılan ilk aşamasında, mikrodizi verilerinin bir matrise aktarılması gerçekleştirilmiştir. İkinci aşamasında, örnek sayısı Antonov yaklaşımı ile çoğaltılmış ve sistem C4.5 karar ağacı ile eğitilmiştir. Son aşamada WEKA yazılımı kullanılarak, Naive Bayes, Karar ağaçları ve Destek Vektör Makinesi sınıflandırma algoritmaları ile önerilen yöntemin başarısı test edilmiştir.

Li ve arkadaşları [14], Genetik Algoritma ve Destek Vektör Makinesi tabanlı gen seçimi gerçekleştirmişlerdir. Genetik algoritma bir arama motoru, destek vektör makinesi ise bir sınıflandırıcı olarak kullanılmıştır. Belirlenen genler ile %99 sınıflandırma başarısı elde edilmiştir.

Wang ve arkadaşları [15], filtreleme yöntemlerinden korelasyon tabanlı öznitelik seçme işlemi ile belirli mikrodizi verilerinde gen seçimini gerçekleştirmişlerdir. Elde ettikleri genlerin sınıflandırma başarısını farklı makine öğrenme yöntemleri ile test etmişler ve sonuçların başarılı olduğunu görmüşlerdir.

Uriarte ve Andres [16], Rasgele Orman (Random Forest) ile gen seçimi gerçekleştirmişler, elde ettikleri sonuçların sınıflandırma başarılarını k-En Yakın Komşu ve Destek Vektör Makinesi ile test etmişlerdir.

Alon ve arkadaşları [17], farklı hücre tiplerine sahip gen ifade verilerinden oluşan veri kümesinin analizi için iki yönlü kümeleme metodunu önermişlerdir. Önerilen bu metotla, 40 tümör ve 20 normal doku bilgisi ile 6 500 den fazla genetik bilgiyi içeren Affymetrix oligonükleotid dizisi başarılı bir şekilde sınıflandırılmıştır.

Dudoit ve arkadaşları [18], farklı kanser türlerine ait gen ifade verilerinden tümör sınıflandırması yapmışlar. Bu sınıflandırmada, k-En yakın komşu, doğrusal diskriminant analizi (LDA, Linear Discriminant Analysis) ve karar ağaçları gibi sınıflandırma algoritmalarının başarılı bir şekilde kullanılabileceğini göstermişlerdir.

Veer ve arkadaşları [19], meme kanseri hastalığının genetik faktörler aracılığıyla teşhis edilebilmesi amacıyla istatistiksel yöntemlere dayalı gen seçimi çalışması yapmışlardır. Çalışmada, 5 yıl içinde metastaz görülen hasta grubu ile görülmeyen hasta grubu olmak üzere iki sınıflı bir gen ifade verisi kullanmışlardır. Ve belirlenen 70 gen ile hastalığın %83 doğrulukta teşhis edilebileceğini ortaya koymuşlardır.

Literatür taramasında görülmüştür ki veri madenciliği, gen analizlerinde başarılı bir şekilde kullanılmaktadır. Ancak gen ifade verilerinin yüksek boyutu sınıflandırma performansını olumsuz etkilemektedir. Bu bağlamda nitelik indirgeme çok önemlidir. Gerçekleştirilen çalışmada gen seçme işlemi için filtreleme ve genetik algoritma tabanlı yeni bir metot önerilmektedir. Önerilen metotta, farklı filtreleme yöntemleri kullanılarak gen ifade verisi indirgenmiş daha sonra genetik algoritma ile etkin genlerin seçimi gerçekleştirilmiştir.

3. BİYOİNFORMATİK VE VERİ MADENCİLİĞİ

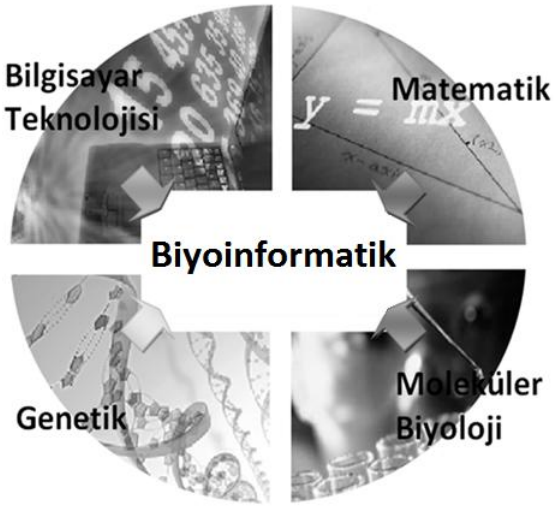
Yaşamın şifresi olarak adlandırılan DNA ve onun işleyişini anlamak için halen pek çok araştırma yapılmaktadır. Yaşanan teknolojik gelişmeler, DNA hakkında daha fazla veriyi bize sunmaktadır. Ancak ortaya çıkan verilerin bilgisayar desteği olmaksızın analizi mümkün değildir. Biyoinformatik ve veri madenciliği bu verinin analizinde en büyük yardımcımız olmuştur.

Bilgisayar desteği ile hastalıkların temelinde yatan genetik faktörlerin belirlenmesi biyoinformatiğin ilgi alanına girmektedir. Ayrıca veri madenciliği de gen analizlerinde başarılı bir şekilde kullanılmaktadır. Bu bölümde biyoinfomatik alanından, veri madenciliği tekniklerinden ve genetik algoritmadan bahsedilmiştir.

3.1. Biyoinformatik

Biyoloji ve informatiğin birleşimi olarak tanımlanan biyoinformatik, moleküler biyoloji alanında elde edilen verilerin bilgisayar destekli analizinin gerçekleştirildiği bir bilim dalıdır. Biyoinformatik, biyolojik bilgilerin elde edilmesi, veritabanlarında saklanması ile ilgilenir. Diğer bir ifade ile moleküler biyolojide elde edilen verilerin sayısal analizini yapan bilgisayar destekli bir bilim dalıdır. Biyoinformatik, DNA, RNA ve protein gibi biyolojik verilerin depolanması, kullanılması, gerektiğinde erişimi ve dağıtımı için bilgisayar teknolojisinden faydalanır.

Şekil 3.1'de görüldüğü gibi biyoinformatik, veritabanı teknolojilerini, veri madenciliğini, yapı ve süreç modellemesi gibi araçları kapsamaktadır.



Şekil 3.1. Biyoinformatik

Moleküler biyoloji alanında ortaya çıkan sorunlara hesaplamalı yaklaşımlarla çözüm arayan biyoinformatik, bu sorunlara çözüm bulmak için yine önceden elde ettiği büyük miktardaki moleküler veriyi kullanır. Moleküler biyoloji alanında yaşanan teknolojik gelişmeler ile elde edilen büyük miktardaki veri, analiz problemlerini de beraberinde getirmektedir. Ortaya çıkan büyük miktardaki bu verinin analizi için bilgisayar teknolojisi kaçınılmaz olmuştur.

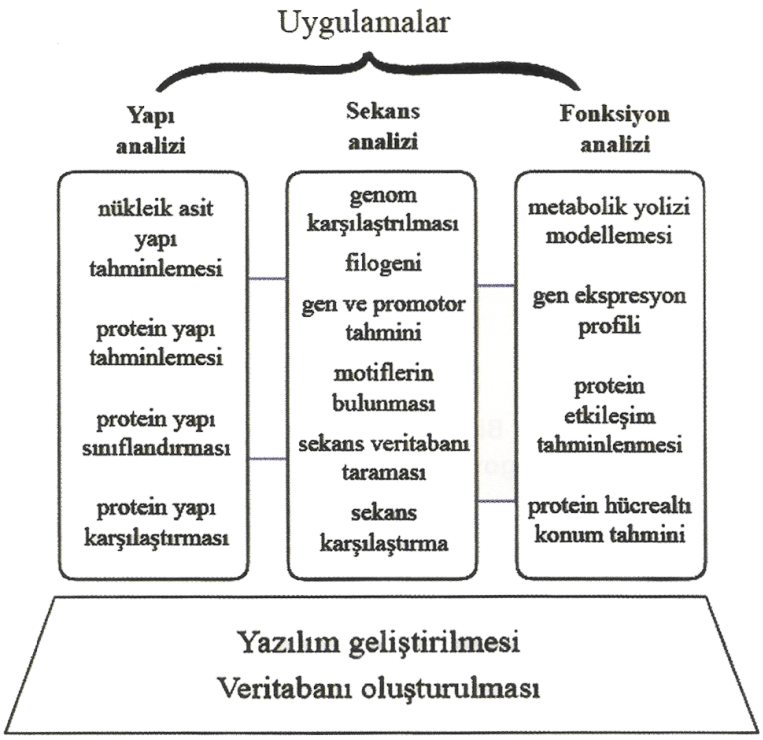
Biyoinformatiğin amacı şu iki başlıkta açıklanabilir.

1. Araştırmacıların kullanımına sunmak amacıyla elde edilen verileri düzenlemek ve depolamak,
2. Depolanan bu verileri anlamlandıracak araç geliştirerek, verilerin anlamlı hale getirilmesini sağlamak,

Geleneksel biyolojik çalışmalarla birkaç türün karşılaştırılması yapılabilmekte iken biyoinformatik veritabanlarında tutulan veriler ile aynı anda pek çok türün karşılaştırılması yapılabilmektedir. Bu amaçla kullanılacak verileri, ham DNA dizileri, protein dizileri, makromoleküler yapılar, genom dizileri ve tüm genom olarak gruplandırmak mümkündür. Ham DNA, dört bazın harflerinden oluşurken proteinler 20 aminoasidin harflerinden oluşmaktadır. Makromoleküler yapılar daha karmaşık yapıdadır. Protein bilgi bankasında tutulan verinin büyük çoğunluğu

protein yapısı hakkındadır. Ortalama protein bilgisi üç boyutlu yaklaşık 2 000 atom içerir. İnsan genom çalışmaları, 3 milyar baz uzunluğunda farklı boyutlarda dizileri ortaya koymuştur [20].

Biyoinformatikte tek sorun ortaya çıkan verinin depolanması değildir. Bu büyüklükte verinin analiz edilmesi ve bu amaçla yazılım geliştirilmesi ciddi bir problemdir. Bu yazılımlar özellikle moleküler dizi analizleri, moleküler yapı analizleri ve moleküler fonksiyon analizleri için önem taşımaktadır. Şekil 3.2'de biyoinformatiğin alt çalışma alanları görülmektedir.



Şekil 3.2. Biyoinformatiğin alt çalışma alanları [20]

Biyolojik veritabanları, biyolojik verilerin düzenli bir şekilde saklanması ve güncellenmesi amacıyla tasarlanmış bilgisayar yazılımlarıdır. Kullanılabilir nükleotid dizileri için üç temel veritabanı bulunmaktadır. Bunlar:

- NCBI tarafından oluşturulun GenBank
- Avrupa biyoinformatik enstitüsü tarafından oluşturulan EMBL
- Japonya tarafından oluşturulan DDBJ

Protein dizi verileri hizmeti sağlayan başlıca veritabanları ise şunlardır:

- GenBank
- EMBL
- Swiss-Prot

Biyoinformatikte kullanılan bazı terimler şunlardır:

Accession number (GenBank): Bir dizi GenBank'a kaydedildiği zaman bu kayda özel bir kimlik numarası verilir. Bir büyük harf, beş rakam veya iki büyük harf, altı rakamdan oluşur. Girilen dizi bilgisi değişse dahi kimlik bilgisi değişmez.

Accession number (RefSeq): Bütün RefSeq dizisine atanmış kimlik numaralarıdır. Sırayla iki büyük harf, bir alt çizgi "_" ve altı rakamdan oluşur.

- NT_123456 birleştirilmiş genomik kontigler
- NM_123456 mRNA'lar
- NP_123456 proteinler
- NC_123456 kromozomlar

Bit score: istatistiksel özelliklere bakılarak yapılan skorlamadır. Normalize edilmiş değerler, farklı karşılaştırmalar yapmak mümkün olabilir.

Blast (Basic Local Aligment search Tool): Aynı veya farklı organizmalar arasındaki nükleotid ya da protein dizisi karşılaştırmada kullanılan bir yazılımdır.

Blosum (Block substitution matrix): Proteinlerin karşılaştırılması sonucu elde edilmiş değişim frekansının gözlemi sonucu elde edilen değerler matrisi.

CDS: Bir nükleotid dizisinin kodonları oluşturan bölgesi veya kodlayan dizi.

Conserved sequence: Bir DNA molekülünde evrim sürecinde değişmeden kalan baz dizisi.

Contig: Bir kromozomun çakışma gösteren klonlanmış farklı DNA parçaları.

Domain: Bir proteinin bağımsız olarak çalışabildiği parçası.

E value (expection value): Beklenen değer. Bit skor değerine denk gelen ya da büyük skorlara sahip benzer dizilerin sayısı. Düşük E, büyük skora işaret eder.

EST (expressed sequence tag): Bir cDNA molekülünün, bir genin tanımlayıcısı olarak kullanılabilecek parçası. Genlerin konumlandırılmasında ve haritalanmasında kullanılır.

Homologue: Dizisi büyük oranda başak bir gene benzeyen gen. Bu genlerin benzer fonksiyon gösterdiği kabul edilir.

Motif: Protein dizisi içinde kısa, korunmuş bölge.

Orthologous: Benzer fonksiyon gösteren, farklı türlere ait homolog dizileri.

Paralogous: Aynı tür içinde gen duplikasyonu sonucu oluşmuş homolog diziler.

Query: Veritabanındaki tüm dizilerle karşılaştırılacak giriş dizisi.

3.2. Veri Madenciliği

Veri madenciliği veya bilgi keşfi (KDD- Knowledge Discovery from Data), veri içinde ilginç, önceden bilinmeyen ve potansiyel olarak faydalı bilgilerin elde edilmesi olarak tanımlanır [21]. Veri madenciliği, bilgi keşfinde önemli bir rol oynar ve biyoinformatik alanında başarıyla uygulanır [22-27].

Veri madenciliği şu yedi adımı içerir:

1. Veri kaynağından yeterli örnek veri seçilmesi,
2. Verilerin önişlemden geçirilmesi. Hatalı ve gereksiz verilerin temizlenmesi,
3. Veri madenciliği yapılacak kadar veri boyutunun azaltılması,
4. Veri madenciliği algoritmalarının uygulanması,
5. Elde edilen bilginin değerlendirilmesi,
6. Sonuç bilginin sunulması,
7. Yeni hipotezleri test etmek için 1. adıma dönülmesi.

Yukarıda bahsedilen işlemler, bilgi keşfi olarak da adlandırılan veri madenciliği adımlarıdır ve döngüsel bir yapıdadır. Her bir döngü belirli bir algoritma veya uzman görüşüne bağlı geri beslemeler içerir.

Veri madenciliği sadece büyük veri kümelerinde değil mikrodizi verisi gibi az örneğe sahip ancak büyük boyutlu verilerde de tercih edilir. Veri madenciliği algoritmaları biyoinformatik alanında ve diğer alanlarda çok hızlı analizler yapmamıza imkan sağlar.

Veri madenciliği ile bilgi keşfinde farklı algoritmalar kullanılabilir. Bunlardan bazıları denetimli (sınıflandırma), denetimsiz (kümeleme), regresyon analizi ve makine öğrenme algoritmalarıdır.

Kümeleme algoritmaları, birbirine benzer kayıtları bulmak için sıklıkla kullanılır. Sınıflandırma algoritmaları, önceden etiketlenmiş verilerden öğrendiği modeli kullanarak sınıfı bilinmeyen bir veri için öngörü uygulamalarında tercih edilir. Regresyon analizi ise bazı istatistiksel fonksiyonlar kullanarak girdi parametrelerine göre çıktıyı yaklaşık olarak kestirmeye çalışan bir model sunar. En basit yapısı doğrusal regresyondur ve $y{:}wx + b$ ile ifade edilebilir. x, giriş parametreleri w ise bu parametrenin ağırlık değeridir. Doğrusal olmayan problemler, daha karmaşık fonksiyonlar ile ifade edilebilir [28].

3.2.1. Verinin hazırlanması

Sınıflandırma ve tahmin doğruluğunu arttırmak için sınıflandırmadan önce veriler önişlemden geçirilebilir.

Veri temizleme: Bu aşama verinin gürültüden ve var olan hatalı verilerden temizlenmesi aşamasıdır. Her ne kadar sınıflandırma algoritmalarının pek çoğu gürültülü veriyi göz ardı etse de yine de gürültü ve hatalı veriler eğitim safhasında süreyi uzatır veya performansı düşürür.

Geçerlilik analizi: Veri kümesinde bazı nitelikler sınıflandırma için gereksiz olabilir. Var olan iki nitelik arasında ilişki olup olmadığını belirlemek için korelasyon analizleri sıklıkla kullanılmaktadır. Örneğin A_1 ve A_2 gibi iki nitelik arasında güçlü bir korelasyon var ise bu niteliklerden birinin atılması sınıflandırma işleminde faydalı olabilir. Veri kümesi aynı zamanda gereksiz nitelikler de içerebilir. İlgisiz niteliklerin bulunarak atılması önemli bir konudur. İlgisiz nitelikler, sınıflandırma performansını olumsuz etkiler.

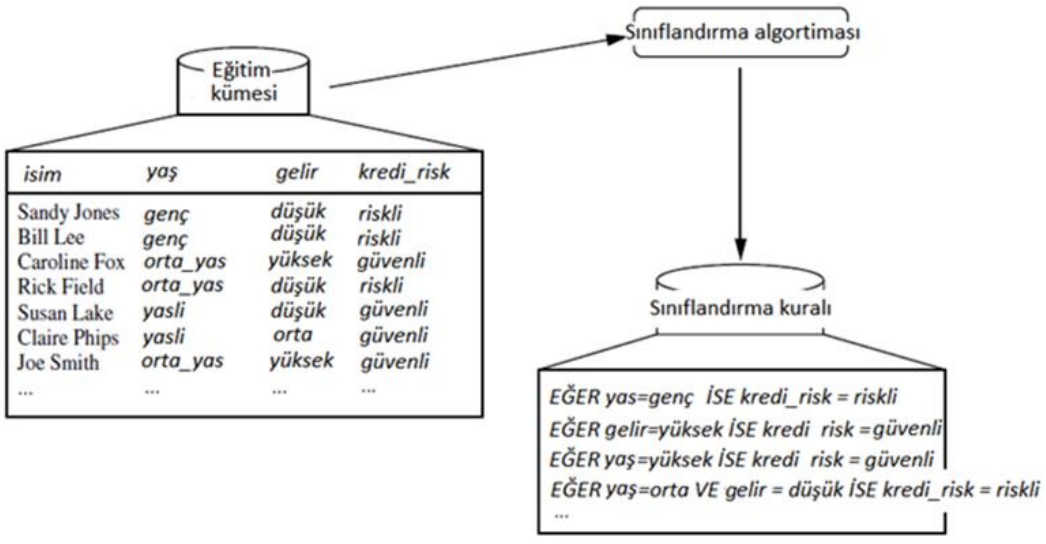
Veri dönüşümü ve indirgeme: Özellikle yapay sinir ağları gibi sınıflandırıcılar verilerin normalize edilmesini gerektirebilir. Normalizasyon, bütün değerlerin daha küçük bir aralıkta yeniden ölçeklendirilmesidir.

3.2.2. Sınıflandırma ve öngörü

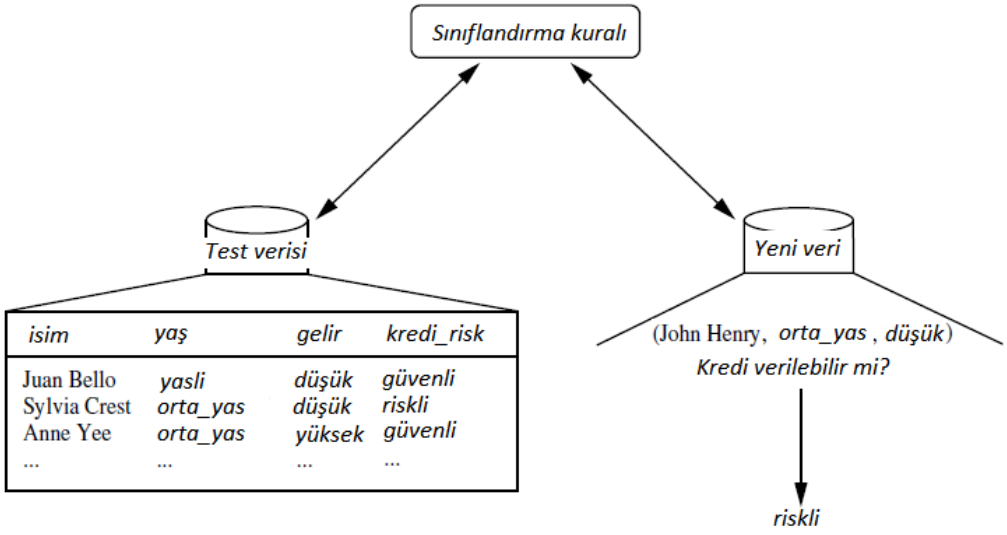
Sınıflandırma, veri madenciliği alanında sıklıkla kullanılmaktadır [29-31]. Veritabanları, öngörü sistemleri için pek çok gizli bilgi taşır. Sınıflandırma ve veri modelleme, geleceğe dair eğilimleri ortaya çıkarmada ve tahmin etmede kullanılan iki önemli yöntemdir. Bu tip analizler elimizde var olan geniş veri yığınlarını anlamamıza yardımcı olur. Sınıflandırma için verilerin etiketlenmiş yani kategorize edilmiş olması gerekir. Örneğin, bir bankanın verilen kredinin geri dönüp dönmeyeceğini öngörecek bir yazılım geliştirebilmesi için, müşterileri kredisi zamanında ödeyen ve ödemeyen şeklinde iki grup altında sınıflandırması gerekir.

Böyle bir sınıflandırma için makine öğrenmesi, istatistik ve örüntü tanımada pek çok metot önerilmiştir [32].

Sınıflandırma, öğrenme ve sınama olmak üzere iki adımlı bir işlemdir [33,34]. Bankacılıkta kredilendirme problemi için sınıflandırma Şekil 3.3'te görüldüğü gibi olacaktır. Birinci adımda, veri kümesinde bulunan kayıtlar düşük veya yüksek riskli olarak etiketlendirilmiştir. Öğrenme ya da eğitim safhası olarak adlandırılan adımda sınıflandırıcı veri ve etiket arasındaki ilişkiyi modeller. *n* boyutlu bir özellik vektörü *X* ile ifade edilse, özellik vektörü, $X = (x_1, x_2, \ldots, x_n)$ şeklinde gösterilebilir. Veritabanında bu nitelikler $A_1, A_2, \ldots A_n$ şeklinde gösterilebilir. Sınıflandırma, etiketlenmiş verileri kullanması sebebiyle, denetimli öğrenme olarak da bilinir. Verilerin etiketlenmiş olması, sınıflandırma ile kümeleme olarak bilinen denetimsiz öğrenme arasındaki temel farktır. Sınıflandırmanın ilk adımı model veya fonksiyon öğrenme olarak da adlandırılır. *X* öznitelikleri ile *y* sınıfını tahmin eden bir fonksiyon $y = f(X)$ şeklinde ifade edilebilir.



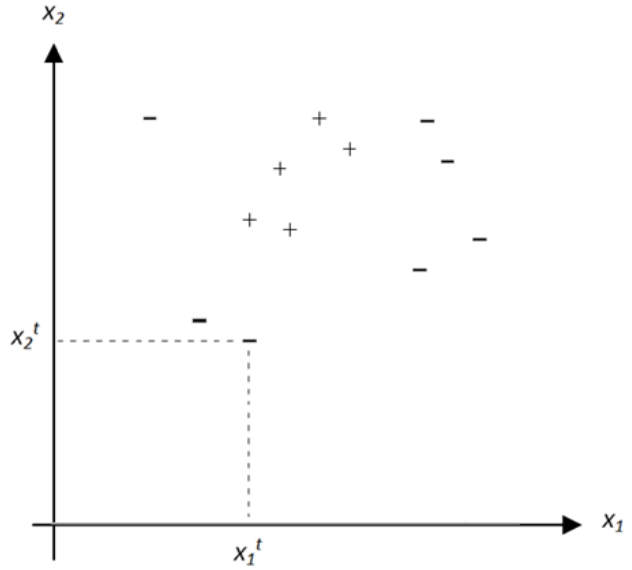
(a) Öğrenme ya da eğitim safhası



(b) Test safhası

Şekil 3.3. Sınıflandırma için bankacılık örneği

Böylece verileri sınıflara ayıran bir fonksiyon ya da model öğrenilmiş olacaktır. Tipik olarak bu eşleştirme kuralları karar ağaçları ya da matematiksel formül şeklinde olabilir [33].



Şekil 3.4. Nitelikleri x_1 *ve* x_2 olan veri kümesinin dağılımı

Şekil 3.4'te da görüldüğü gibi nitelikleri x_1 *ve* x_2 olan veri kümesinde her bir kayıt şu şekilde ifade edilebilir:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{3.1}$$

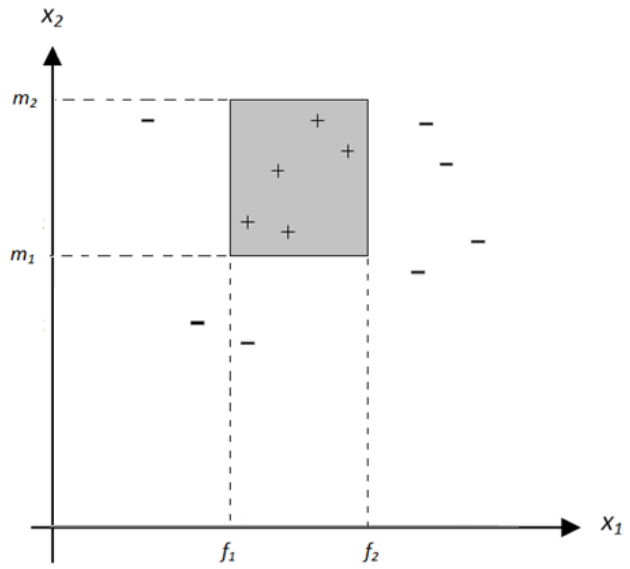
Etiket ise sınıf değerlerini belirler ve aşağıdaki gibi gösterilebilir:

$$y = \begin{cases} 1 \text{ eğer } x \text{ pozitif bir örnekse} \\ 0 \text{ eğer } x \text{ negatif bir örnekse} \end{cases} \tag{3.2}$$

Her bir kayıt sıralı (*x*,*y*) çiftiyle gösterilebilir ve öğrenme kümesi böyle *N* örnekten oluşur.

$$X = \{ x^t , y^t \}_{t=1}^{N} \tag{3.3}$$

Burada *t*, kümedeki farklı örnekleri gösteren indis değişkenidir. Pozitif ve negatif sınıflara sahip veri kümesi iki boyutlu x_1 *ve* x_2 uzayında Şekil 3.5'te görüldüğü gibi çizilebilir. Her *t* örneği (x_1^t , x_2^t) koordinatlarında bir veri noktasıdır. Sınıf, pozitif ve negatif olmak üzere, y^t *ile gösterilmiştir.*



Şekil 3.5. İki boyutlu x_1 *ve* x_2 uzayında veri kümesinin dağılımı ve pozitif sınıfın ayrımı

Buna göre pozitif sınıf şu şekilde ifade edilebilir:

$(f_1 \leq x_1 \leq f_2)$ *VE* $(m_1 \leq x_2 \leq m_2)$ (3.4)

Burada f_1, f_2, m_1 ve m_2 aralık sınır değerlerini göstermektedir.

Sınıflandırıcının doğruluğu, ezberleme riski sebebiyle eğitim veri kümesi ile test edilemez. Eğitim aşamasında kullanılmayan ayrı bir test grubu, sınıflandırıcının doğruluğunu test etmede gereklidir. Bu amaçla veri kümesi eğitim ve test veri kümesi olarak parçalara bölünebilir. Sınıflandırıcının, test verisinde doğru bulduğu kayıtlar, sınıflandırıcının doğruluk oranıdır. Sınıflandırıcının doğruluğu kabul edilebilir düzeyde olduğunda sınıflandırıcı gerçek veriler üzerinde kullanılabilir [33].

Destek Vektör Makinesi – DVM (Support Vector Machine - SVM)

Destek vektör makinesi, doğrusal ve doğrusal olamayan veriler için sınıflandırma yapabilen güçlü bir sınıflandırma algoritmasıdır [35-37]. Diskriminant tabanlı bir yöntem olan Destek Vektör Makinesi (DVM), 1995 yılında Vapnik tarafından ortaya atılmıştır. Vapnik, sınıflandırmada ayırt ediciyi öğrenmek için $p(x|C_i)$ sınıf olasılıklarını ya da $p(C_i|x) = p(C_j|x)$ koşulunu sağlayan x değerlerini arar. Burada x, veri kümesindeki örneği, C ise sınıfı göstermektedir. Öğrenme gerçekleştiğinde, doğrusal modelin parametresi olan ağırlık vektörü öğrenme kümesi için *destek vektör* olarak belirlenir. Sınıflandırmada bu değerler sınıra en yakın, muğlak değerlerdir. DVM'de çıktı destek vektörlerinin etkilerinin toplamı olarak gösterilir. Bu gösterim örnekler arasında uygulamaya özgü bir benzerlik tanımlayan *çekirdek fonksiyonu* tarafından belirlenir. Çoğu öğrenme algoritmasında girdiler vektör olarak gösterilmektedir. Benzerlik için Öklid uzaklığı kullanılır. Çekirdek fonksiyonları bunun üzerinde bir işlem yapar.

DVM, -1 ve +1 olarak etiketli iki sınıf için şöyle çalışır: $X = \{ x^t , y^t \}$ örnekleminde $x^t \in C_1$ *ise* $y^t = +1$ *ve* $x^t \in C_2$ *ise* $y^t = -1$ değerini aldığını kabul edelim. w ve b parametreleri ile koşul;

$$w^T x^t + b \geq +1 \qquad eğer \quad y^t = +1 \tag{3.5}$$
$$w^T x^t + b \leq -1 \qquad eğer \quad y^t = -1$$

olacaktır. Bu koşul şöyle de ifade edilebilir:

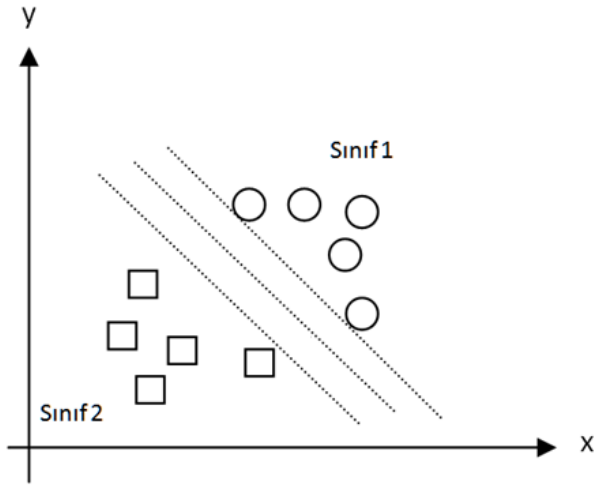
$$y^t(w^T x^t + b) \geq +1 \tag{3.6}$$

Örneklerin hiper düzlemin yakınında değil, daha iyi genelleme yapması yani ona belli bir uzaklıkta kalması için koşulun,

$$y^t(w^T x^t + b) \geq 0 \tag{3.7}$$

olmaması gerekir.

DVM nin amacı en büyük ayrımı yapan hiper düzlemi bulmaktır. Şekil 3.6'da iki sınıfı ayıran muhtemel hiper düzlemler görülmektedir.



Şekil 3.6. İki boyutlu, doğrusal ayrılabilen veri ve hiper düzlemler

Şekil 3.6'da görüldüğü gibi her hiper düzlem sınıfları doğru bir şekilde ayırabilmektedir. En uygun hiper düzlem, birbirine en uzak iki hiper düzlemin seçilmesi ile olur, bu da DVM nin genelleme başarısını arttırır.

$W.X + b$ şeklinde ifade edilen hiper düzlemde W ağırlık vektörü olmak üzere, $W=\{w_1, w_2, .. ., w_n\}$ şeklinde gösterilebilir. Burada n, nitelik sayısını göstermektedir. Bu durumda iki boyutlu yani iki niteliğe sahip veri kümesinde hiper düzlem şu şekilde yazılabilir:

$$b+w_1x_1+w_2x_2 \geq 0 \quad (3.8)$$

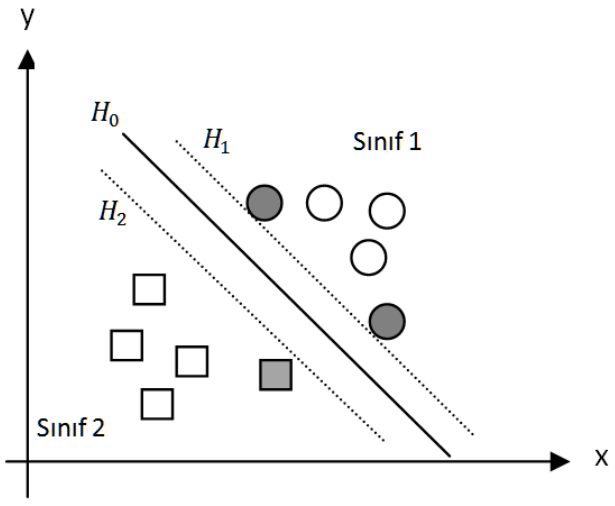
Böylece hiper düzlemin üzerinde kalan alan Eş. 3.9 ile ayrılabilirken

$$b+w_1x_1+w_2x_2 > 0 \quad (3.9)$$

hiper düzlemin altında kalan alan Eş. 3.10 ile ayrılabilir

$$b+w_1x_1+w_2x_2 < 0 \quad (3.10)$$

Maksimum marjinal hiper düzlemi bulmak için Öklid'den faydalanabiliriz. $\sqrt{W.W}$ ve eğer $W=\{w_1, w_2, .. ., w_n\}$ ise $\sqrt{W.W} = \sqrt{w_1^2 + w_2^2 + \ ... + w_n^2}$ şeklinde ifade elde edilebilir.



Şekil 3.7. İki sınıfı birbirinden ayıran en uygun hiper düzlem

Şekil 3.7'de görüldüğü gibi H_1 ve H_2 hiper düzlemleri göz önüne alındığında, bu düzlemler üzerindeki noktalar *destek vektör* adını alır. Bir destek vektör ile $H_0=WX+b$ arasındaki uzaklık d olmak üzere Eş. 3.11'de gösterildiği gibi mesafe şöyle ifade edilebilir:

$$d = \frac{|w_1\ x_1\ +\ w_2\ x_2\ +\ \ldots +\ w_n\ x_n\ +\ b|}{\sqrt{w_1^2 + w_2^2 +\ \ldots + w_n^2}} \tag{3.11}$$

Doğrusal olmayan optimizasyon problemlerinin çözümünde *Lagrange fonksiyonundan* faydalanılabilir. Bu amaçla şu yol izlenebilir. $f(x_1, x_2,\ldots, x_n)$ şeklinde tanımlı bir fonksiyonda,

$$\begin{aligned} &g_1(x_1, x_2,\ldots, x_n)= b_1 \\ &g_2(x_1, x_2,\ldots, x_n)= b_2 \\ &\ldots \\ &g_n(x_1, x_2,\ldots, x_n)= b_n \end{aligned} \tag{3.12}$$

koşulları altında optimize eden noktaları bulmak için, $\alpha_1, \alpha_2, \ldots \alpha_n$ olarak ifade edilen *Lagrange çarpanları* ile $L(x, \alpha)$ lagrange fonksiyonu şu şekilde ifade edilir.

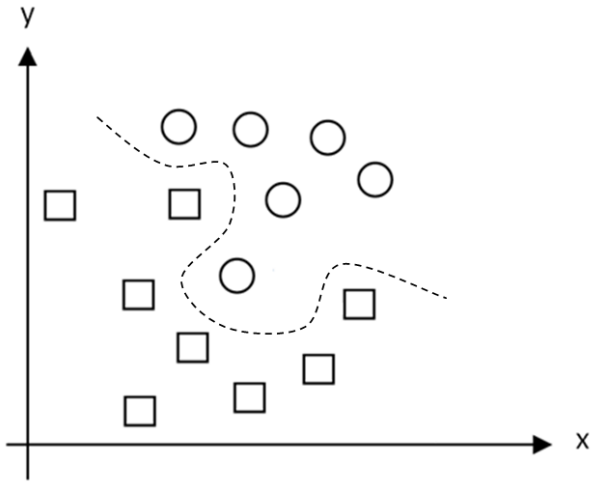
$$L(x_1, x_2,\ldots, x_n, \alpha_1, \alpha_2, \ldots \alpha_n) = f(x_1, x_2,\ldots, x_n) - \sum_{i=1}^{n} \alpha_i[g_i(x_1, x_2,\ldots, x_n)-b_i] \tag{3.13}$$

Bu durumda $f(x_1, x_2,\ldots, x_n)$ fonksiyonunu maksimize ve minimize eden noktaları bulmak için Eş. 3.14'ten faydalanılır.

$$\begin{aligned} &\frac{\partial L}{\partial x_i} = \frac{\partial f}{\partial x_i} - \sum_{i=1}^{n} \alpha_i \frac{\partial g_i}{\partial x_j} = 0 \qquad (j=1,2,\ldots,n) \\ &\frac{\partial L}{\partial \alpha_i} = g_i(x_1, x_2,\ldots, x_n)+ b_i = 0 \qquad (i=1,2,\ldots,n) \end{aligned} \tag{3.14}$$

Sıfırdan büyük α_i değerleri destek vektörü tanımlar.

Şekil 3.8'de görüldüğü gibi verilerin doğrusal ayrılamaması durumunda, negatif olmayan ve hataları ifade eden ξ gevşek değişkenin optimizasyon modeline eklenmesi ile problem çözülebilir.



Şekil 3.8. Doğrusal ayrılamama durumu

Gevşek değişken yardımı ile $y_i(w^T x + b) \geq 1 \; \forall$i yerine

$$y_i(w^T x_i + b) \geq 1 - \xi_i \qquad (i=1,2,\ldots,n),\ \xi_i \geq 0 \tag{3.15}$$

veya

$$(w^T x_i + b) \geq 1 - \xi_i \,,\ y_i = +1 \tag{3.16}$$
$$(w^T x_i + b) \leq 1 - \xi_i \,,\ y_i = -1$$

yazılabilir.

Burada $\xi_i > 1$ olan veriler, hiper düzlemin diğer tarafında kalan yani doğrusal ayrılmayan verilerdir. $0 < \xi_i < 1$ ise hiper düzlemin doğru bir şekilde ayırdığı ancak en büyük alan marjin bölgesi içinde kalan değerleri ifade eder.

Bu tip optimal hiper düzlem için maksimize edilecek fonksiyon, doğrusal ayrılmayı engelleyen durumlar için *C* olarak ifade edilen ceza parametresine ihtiyaç duyar. Bu fonksiyon Eş. 3.17'de görüldüğü gibi olacaktır.

$$L(w, \xi)=\frac{1}{2}\, w^T w + C(\sum_{i=1}^{n} \xi_i)^k \tag{3.17}$$

Burada *C* sıfırdan büyük bir sabittir ve kullanıcı tarafından seçilebilir. Eğer *C* küçük seçilirse istenmeyen bazı gözlemler girebilir. $L(w, b, \xi, \alpha, \beta)$ Langrange fonksiyonu $k=1$ için şu şekilde hesaplanabilir.

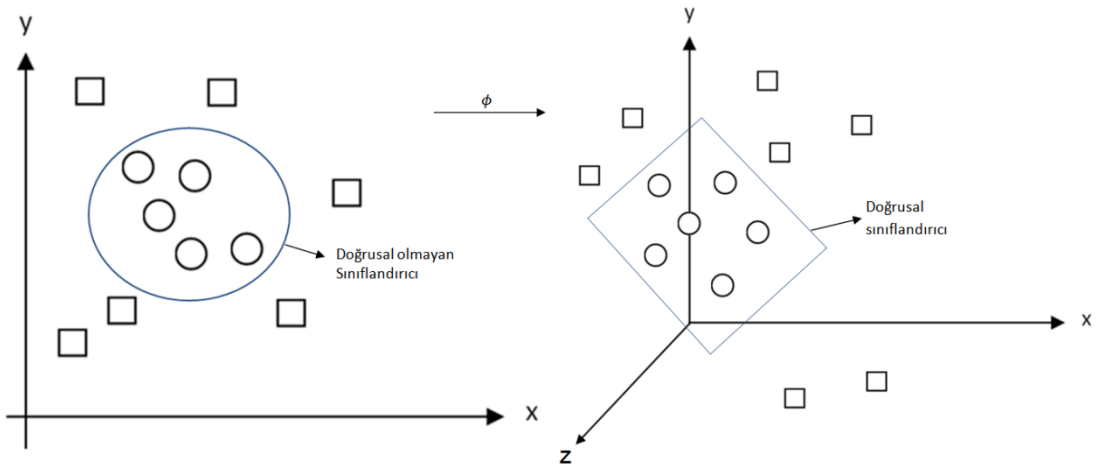
$$L(w, b, \xi, \alpha, \beta) \tag{3.18}$$
$$=\frac{1}{2}\, w^T \mathrm{w} + \mathrm{C}(\sum_{\mathrm{i}=1}^{\mathrm{n}} \xi_i) - \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_i \{y_i[w^T x_i + b] - 1 + \xi_i\} - \sum_{\mathrm{i}=1}^{\mathrm{n}} \beta \xi_i$$

Burada α_i *ve* β_i Lagrange çarpanlarıdır.

$$\frac{\partial L}{\partial w} = 0,\ w=\sum_{i=1}^{n} \alpha_i y_i x_i \tag{3.19}$$
$$\frac{\partial L}{\partial b} = 0,\ \sum_{i=1}^{n} \alpha_{\mathrm{i}} y_i = 0$$
$$\frac{\partial L}{\partial \xi} = 0,\ \alpha_i + \beta_i = 0$$

Doğrusal ayrılamayan veri kümeleri ξ gevşek değişkenler kullanılarak doğrusal hiper düzlem ile ayrılabilmektedir. Bunun dışında *doğrusal olmayan sınıflandırıcılar* kullanılabilir. Şekil 3.9'da görüldüğü gibi doğrusal ayrılmayan veri kümesi daha yüksek dereceli bir uzayda *z* vektörlerine dönüştürülürse, bu veriyi doğrusal sınıflandırıcılar ile ayırmak mümkün olabilir. *z* vektörünün yer aldığı uzay *F* ile gösterildiğinde ϕ ifadesi $R^n \rightarrow R^F$ eşleşmesini yapmak üzere $z= \phi(x)$ şeklinde gösterilebilir.

$$x \in R^n \rightarrow z(x)=[\, \alpha_1, \phi_1(x), \ldots, \alpha_n, \phi_n(x)]^{\mathrm{T}} \in R^F \tag{3.20}$$



Şekil 3.9. İki boyutlu uzaydaki verilerin üç boyutlu uzaydaki verilere dönüştürülmesi ve doğrusal hiper düzlem ile sınıflandırılması

Veri kümesinin doğrusal ayırt edilemediği durumlarda veriyi daha büyük boyutlu uzaya taşıyarak doğrusal hiper düzlem ile ayırmak için $\phi_1(x) \ldots \phi_n(x)$ fonksiyonları kullanılabilmektedir. Ancak bu fonksiyonlar yerine söz konusu dönüşümler için *çekirdek fonksiyonlar* kullanılabilir. Çekirdek fonksiyon için *Mercer* teoremine göre $K(x_i,x_j)=\phi(x_i)^T\phi(x_j)$ yazılabilmesini sağlayan ϕ eşlemesi varsa pozitif kesin ve simetri $K(x_i,x_j)$ bir çekirdek fonksiyondur denir. Bir çekirdek fonksiyonu şu şartları sağlamalıdır:

- Sürekli fonksiyon olmalıdır.
- Simetrik olmalıdır. $K(x_i,x_j)= K(x_j,x_i)$
- Herhangi $x_1,x_2, \ldots ,x_n$ değerleri için pozitif kesin olmalıdır.

Sıklıkla kullanılan çekirdek fonksiyonlar şunlardır [37]:

- Doğrusal: $K(x_i,x_j) = x_i^T x_j$
- Polinom: $K(x_i,x_j,c,d) = (c+x_i^T x_j)^d$
- Radyal: $K(x_i,x_j,\sigma) = e^{\frac{|x_i-x_j|^2}{2\sigma^2}}$

3.2.3. Sınıflandırma algoritmalarının karşılaştırılması

Sınıflandırma algoritmaları şu kriterlere göre karşılaştırılabilir:
Doğruluk: Bir sınıflandırıcının daha önce görmediği bir veriyi doğru sınıflandırma başarısıdır.

Hız: Bir sınıflandırıcının hesaplama maliyeti olarak da ifade edilir.

Dayanıklılık: Bir sınıflandırıcının gürültü veya hatalı veriler olması durumunda bile başarısını sürdürebilmesidir.

Ölçeklenebilirlik: Bir sınıflandırıcının büyük veri kümelerinde bile etkinliğini sürdürebilmesidir [33].

3.2.4. Geçerlik ve güvenirlik

Geçerlik, herhangi bir sınıflandırıcının, ölçülmek istenen özelliği doğru bir şekilde ölçebilmesidir. Güvenirlik ise aynı sınıflandırıcının her uygulamada tutarlı davranmasıdır. Güvenirlik, bağımsız ölçümler arasındaki kararlılık olarak da ifade edilebilir.

Bir öğrenme algoritmasının başarısı ölçülürken, farklı bir geçerleme kümesine ihtiyaç duyulur. Veri üzerinde öğrenme bir kez yapılırsa elimizde bir model ve bir de hata değeri olur. Eğer rastsallık üzerinden öğrenme birden fazla uygulanır ise elimizde birden çok model olacaktır. Bu modellerin hepsi denendiğinde geçerleme hatalarından oluşan bir örneklem elde ederiz. Ancak bunun için öğrenme ve test kümelerinin aynı veriden olması gerekir. Öğrenme algoritmasının doğruluğunu ölçmede bu geçerleme değerlerinin dağılımına bakılır. Böylece algoritmanın beklenen hatası ölçülebilir.

Bu amaçla verinin öğrenme ve geçerleme kümeleri olarak ayrılması, yalnızca ölçme ve değerlendirme içindir. Sonuç modelin başarısını ölçmek için farklı ve daha önce

öğrenme ya da geçerlemede kullanılmayan yeterince büyük bir deneme kümesi gereklidir.

Çapraz geçerleme

Mikrodizi gibi az örnekleme sahip veri kümeleri, yeteri kadar öğrenme ve geçerleme kümesine sahip olmayabilir. Bu durumda veri kümesinden birden fazla öğrenme ve geçerleme kümesi çiftleri oluşturulmalıdır. Bunun için veri kümesi *k* parçaya bölünür. Elde edilen her bir *k* parçası hem eğitim hem de test için kullanılır. Buna *çapraz geçerleme* adı verilir [38].

k Kat çapraz geçerleme

Bu geçerleme yönteminde veri kümesi rasgele *k* tane eşit parçaya bölünür. Bu parçalardan her biri geçerleme için ayrılırken, geri kalan *k-1* parça öğrenme kümesini oluşturur. Böylece elimizde *k* tane çift veri kümesi olur. Literatürde yapılmış çeşitli geçerlilik çalışmalarında verinin 10 alt-kümeye bölünmesinin (k=10) en iyi sonucu verdiği saptanmıştır [39]. *k* arttıkça öğrenme kümelerinin yüzdesi artar ancak geçerleme kümesi küçülür [29].

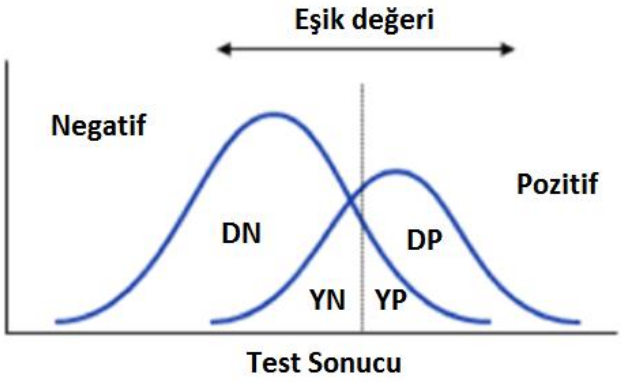
3.2.5. Sınıflandırma performansının ölçülmesi

Pozitif ve negatif etiketli iki sınıf içeren bir veri kümesinde yapılan sınıflandırma sonucunda, muhtemel dört sonuç vardır: Pozitif bir örnek için sınıflandırıcı pozitif karar verirse bu durum *doğru pozitif*, sınıflandırıcı negatif karar verirse, *yanlış negatif* olarak adlandırılır. Negatif bir örnek için sınıflandırıcı pozitif karar verirse *yanlış pozitif,* negatif karar verirse *doğru negatif* olarak adlandırılır. Bu durum Şekil 3.10'da görüldüğü gibi iki sınıf için hata tanımı olarak ifade edilebilir.

Tahmin edilen sınıf	Gerçek sınıf: Pozitif	Gerçek sınıf: Negatif
Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Şekil 3.10. İki sınıf için hata tanımı

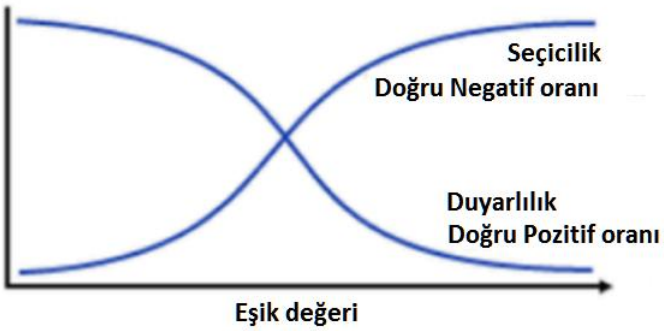
Alıcı İşletim Karakteristiği (Receiver Operating Characteristics - ROC), sınıflandırıcı performansını test etmek veya iki farklı sınıflandırıcının karşılaştırılması amacıyla biyoinformatikte sıklıkla kullanılan bir yöntemdir [31,40,41]. İki farklı grup için dağılım Şekil 3.11'de görüldüğü gibi gösterilebilir. Nadiren bu iki grup mükemmel bir şekilde birbirinden ayrılabilir.



Şekil 3.11. İki farklı sınıfın normal dağılımı

Bu iki sınıfı ayırmak için bir eşik değeri ya da kesim noktasına ihtiyaç vardır. Bu durumda Şekil 3.10'da ifade edildiği gibi dört muhtemel sonuç ortaya çıkar.

Bir sınıflandırıcıda duyarlılık (sensitivity) pozitif örneği doğru bulma, seçicilik (specificity) negatif örneği doğru bulma gücü olarak tanımlanmaktadır. Şekil 3.12'de gösterildiği gibi eşik değeri yüksek seçildiğinde DP oranı ve duyarlılık düşerken, seçicilik artar. Eşik değeri küçük seçildiğinde DP oranı ve duyarlılık artar. Ancak bunun yanı sıra YP de artacak, DN oranı ve seçicilik düşecektir.



Şekil 3.12. Eşik değerine bağlı olarak seçicilik ve duyarlılık eğrisi

Seçicilik ve duyarlılık şu şekilde hesaplanır:

$$\text{Doğru Pozitif Oranı} = \frac{\text{DP}}{\text{DP} + \text{YN}} \tag{3.21}$$

$$\text{Yanlış Pozitif Oranı} = \frac{\text{YP}}{\text{YP} + \text{DN}} \tag{3.22}$$

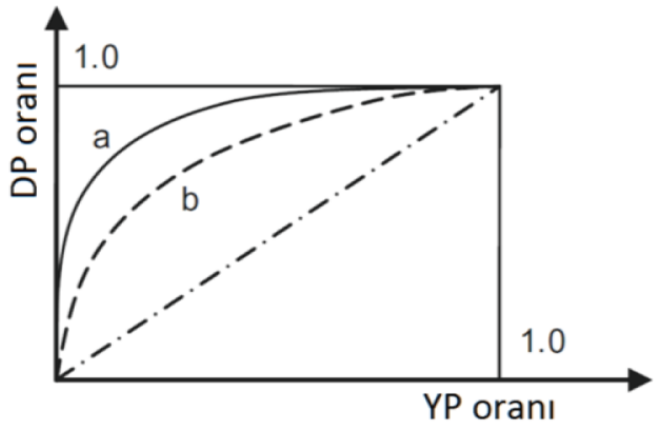
$$\text{Sınıflandırma doğruluğu} = \frac{\text{DP} + \text{DN}}{\text{DP} + \text{YP} + \text{DN} + \text{YN}} \tag{3.23}$$

$$\text{Duyarlılık (Sensitivity)} = \text{DP oranı} \tag{3.24}$$

$$\text{Seçicilik (Specificity)} = 1 - \text{YP oranı} \tag{3.25}$$

Alıcı İşletim Karakteristiği eğrisinin çizilmesi

Eş. 3.23 ile sınıflandırma doğruluğu hesaplanırken, Eş. 3.24 ve Eş. 3.25 kullanılarak, ROC eğrisi Şekil 3.13'te görüldüğü elde edilebilir.



Şekil 3.13. ROC eğrisi

x ekseni $YP_{oranı}$, y ekseni $DP_{oranı}$ olarak çizildiğinde, diyagonal eğrinin üstünde ve sol üst köşeye yaklaşan ROC eğrisine sahip sınıflandırıcı performansı iyi kabul edilir.

Başarılı bir sınıflandırmada doğru pozitif (DP) oranının daha yüksek, yanlış pozitif (YP) oranının da daha düşük olması beklenir. Ayrıca eğri altında kalan alanın (Area Under the Curve-AUC) 1'e yakın olması sınıflandırma performansının yüksek olduğu anlamına gelmektedir. Bu durumda Şekil 3.13'te a sınıflandırıcısının b sınıflandırıcısına göre daha başarılı olduğu söylenebilir.

3.3. Boyut İndirgeme

Yüksek boyutlu gen ifade verilerinden gen seçimi işleminde boyut indirgeme yöntemleri çok önemli bir yer tutar [42]. Gerek sınıflandırma, gerekse kümeleme amaçlı olsun, bilgi kabul edilen tüm gözlem verileri girdi olarak kullanılmaktadır. Ancak bu gözlemlerin hangilerinin bizim için önemli, hangilerinin önemsiz olduğunu bilmemiz gerekir. Veri kümelerinde boyut indirgemenin bize sağlayacağı avantajlar şöyle sıralanabilir:

- Pek çok öğrenme algoritmasında karmaşıklık, veri kümesi örneklem sayısına (N) ve girdi boyutuna (d) bağlıdır.
- Gereksiz bir girdinin tespit edilmesi ile bu girdinin elde edilme maliyeti düşürülür, tahmin modeline dahil edilmeyerek performans artırılabilir.
- Veri kümesinin küçük olması daha basit ve güvenilir modeller geliştirilmesine imkan sağlar. Küçük veri kümelerinde varyans düşüktür yani gürültü, aykırı gözlem vb. durumlardan daha az etkilenir.
- Veri kümesi daha az değişkenle ifade edildiğinde bu verilerin elde edilme süreci daha iyi anlaşılabilir, bu da bilginin elde edilmesini kolaylaştırır.
- Bilgi kaybı olmaksızın veri boyutunun azaltılması, verinin yapısı ve var olan aykırı değerler hakkında sonuç çıkarmamıza yardımcı olur.

Boyut azaltmak için iki yöntem vardır: *öznitelik seçimi* ve *öznitelik çıkarımı*. *Öznitelik seçimi*, *d* değişkenden en çok bilgi içeren *k* tanesinin bulunarak geri kalanının atılması işlemidir. *Öznitelik çıkarımında*, var olan *d* değişkeni birleştirilerek *d*'den daha az *k* tane yeni değişken elde edilir. Bu yöntemler gözetimli ya da gözetimsiz olabilir. Öznitelik çıkarımı için en sık kullanılan yöntem, *temel bileşenler analizi* ve *doğrusal ayırtaç analizi*'dir. Bunların her ikisi de doğrusaldır ve temel bileşenler analizi gözetimsiz, doğrusal ayırtaç analizi ise gözetimlidir.

3.3.1. Öznitelik seçimi

Öznitelik seçimi için *altküme seçimi* yöntemi sıklıkla kullanılmaktadır [43].

Altküme seçimi

Altküme seçiminde amaç öznitelikler içinden en iyi altkümeyi bulmaktır. En iyi alt küme, modelin başarısını en yüksek yapan ve en az sayıda boyuttan oluşan kümedir. Geri kalan nitelikler atılır. *d* değişkeninin olduğu bir veri kümesinde 2^d farklı altküme vardır, *d* küçük değilse bütün altkümelerin sınanması mümkün değildir. Bu sebeple en iyi altkümeyi bulamasak da yeterince hızlı ve daha iyi performans gösterecek altkümeler bulmak için *sezgisel yöntemler* kullanılabilir.

Altküme seçiminde, *İleri Doğru Seçim* ve *Geriye Doğru Seçim* olmak üzere iki yaklaşım vardır. *İleri Doğru Seçim* yaparken boş bir altküme ile başlanır, model hatasını en çok azaltan nitelikler seçilerek birer birer eklenir. *Geriye Doğru Seçim* yönteminde nitelik kümesinin tamamıyla başlanır. Bir niteliğin çıkarılması için hatayı önemli derecede arttırıncaya dek her yinelemede hatayı en çok azaltan (ya da çok az arttıran) nitelikler birer birer çıkarılır.

x_i, $i = 1,2,...d$ ile gösterilen girdi niteliklerinin bir kümesi *F* ile, *F* kümesi kullanılarak eğitilmiş modelin geçerleme hatası (uygulamaya göre bu hata ortalama kare hatası ya da sınıflandırma hatası olabilir). *E(F)* ile ifade edildiğinde, ileri doğru sırayla seçim yaparken boş bir öznitelik kümesi ile başlanır. $F=\phi$. Her adımda olası tüm x_i modele eklenip, öğrenme kümesinde eğitilir. Test kümesi üstünde $E(F \cup x_i)$ hesaplanır. En küçük hata gösteren x_j seçilir ve modeli eklenir.

$$j = \arg\min E(F \cup x_i) \tag{3.26}$$

Bu şöyle de ifade edilebilir:

$$x_i \text{ , } F \text{ kümesine eklenir EĞER } E(F \cup x_i) < E(F) \tag{3.27}$$

Eklenen yeni nitelik, E değerini azaltmıyor veya fark çok az çıkıyorsa işlem tamamlanır. Bu şekilde öznitelik seçimi karmaşıklığı oldukça fazladır. Veri kümesini d boyuttan k boyuta indirmek için $d + (d\text{-}1) + (d\text{-}2) + \ldots + (d\text{-}k)$ kez eğitmek ve geçerleme yapmak gerekir. Bu da $O(d^2)$ anlamına gelmektedir. Ayrıca daha önce tek tek denenmiş ve seçilmemiş, ancak diğer bir nitelikle yan yana geldiğinde daha küçük hata gösteren birleşimlerin de denenmesi gerekebilir. Bu durum arama uzayını daha da genişletecek ve karmaşıklığı artıracaktır. Bütün bunlara rağmen seçilen öznitelik altkümesinin en iyi olduğu garantisi verilemez.

Geriye doğru seçim, tüm nitelikleri içeren F kümesiyle başlar. İleri doğru seçim yöntemine benzer biçimde, ancak ekleme yerine çıkararak her adımda en az hataya neden olan öznitelikleri tespit eder.

$$j = \arg\min E(F - x_i) \tag{3.28}$$

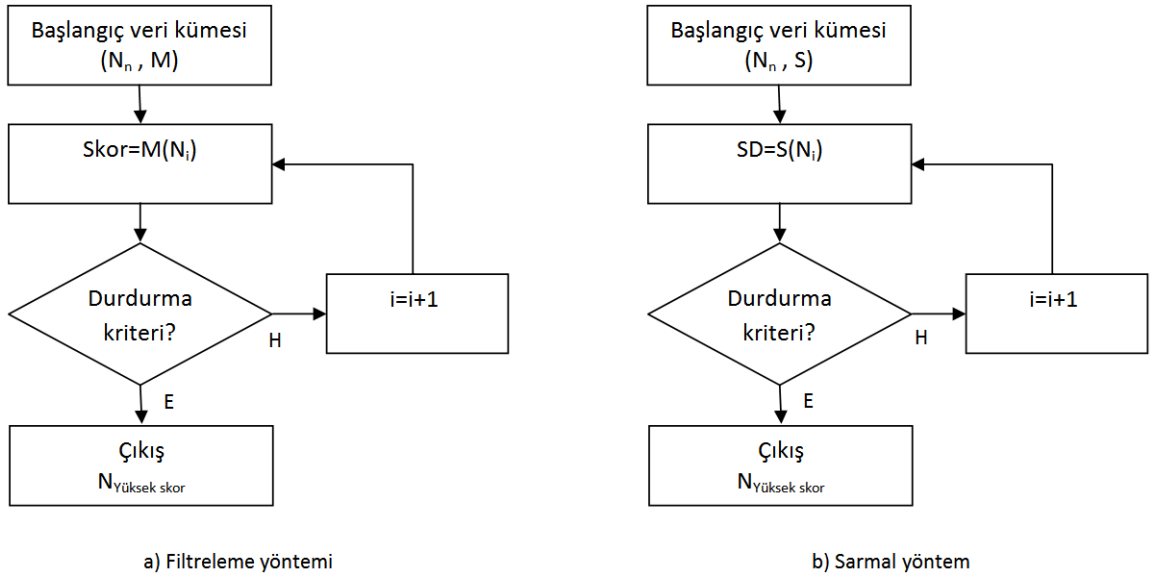
Bu şöyle de ifade edilebilir:

$$x_i \text{ , } F \text{ kümesinden çıkarılır EĞER } E(F - x_i) < E(F) \tag{3.29}$$

Herhangi bir nitelik çıkarma işlemi hatayı azaltmıyor ise geriye doğru seçim tamamlanır. Geriye doğru seçim ile ileri doğru seçim modelleri birbirine oldukça benzer. Her ikisinin de dezavantaj ve avantajları aynıdır. Hatta değişkenlerin pek çoğunun gereksiz olduğu beklenen durumlarda ileri doğru seçim avantajlı olabilir. Ayrıca görüntü tanıma uygulamalarında öznitelik seçimi iyi bir yöntem değildir. Çünkü görüntü noktaları (piksel) tek başına ayırt edici bilgi taşımaz. Örneğin yüz

tanıma için bilgi tekil noktalardan değil birçok noktanın değerinin birleşimi ile elde edilir.

Öznitelik seçme, filtreleme ve sarmal olmak üzere iki ayrı grupta da incelenebilir [44,45]. Şekil 3.14'te nitelik seçme yöntemleri görülmektedir. Şekil 3.14(a) da görüldüğü gibi filtreleme yöntemi, herhangi bir öğrenme algoritmasından bağımsız, zayıf bilgi içeren nitelikleri süzmek için istatistiksel özellikleri kullanır [46,47]. Çoğu uygulamada özellik ilişki skoru hesaplanır. Bunun sonucunda az skora sahip nitelikler atılır. Daha sonra elde edilen nitelik altkümesi sınıflandırma için kullanılır [48]. Burada N_n, n adet niteliği, M ise bağımsız testleri ifade eder. Durdurma kriteri şunlardan birine göre belirlenebilir: Herhangi bir niteliğin eklenmesi ya da çıkarılması daha iyi bir nitelik altkümesi vermiyor ise veya istenilen nitelik sayısına ulaşılmış ise durdurma gerçekleşir.



a) Filtreleme yöntemi b) Sarmal yöntem

Şekil 3.14. Nitelik seçme yöntemleri

Biyoinformatik alanında sıklıkla kullanılan Fisher Korelasyon Skorlama, t-Skor ve Welch t-İstatistik aşağıda gösterildiği gibi hesaplanmaktadır [43]:

Fisher korelasyon skorlama:

$$FKS(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^+ + \sigma_i^-} \tag{3.30}$$

t-Skor:

$$t(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{(n^+(\sigma_i^+)^2 + n^-(\sigma_i^-)^2)/(n^+ + n^-)}} \tag{3.31}$$

Welch t-istatistik:

$$WTS(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \tag{3.32}$$

Burada $\mu_i^+ ve\, \mu_i^-$ sınıfların aritmetik ortalaması, $\sigma_i^+ ve\, \sigma_i^-$ sınıfların varyansı ve $n_i^+ ve\, n_i^-$ sınıf örnek sayılarıdır. Bu çalışmada Eş. 3.30, Eş. 3.31 ve Eş. 3.32 kullanılarak nitelik seçimi gerçekleştirilmiştir.

Şekil 3.14(b)'de görüldüğü gibi sarmal yöntemde, bağımsız testler yerine özel makine öğrenme metotları (Destek vektör makinesi, Karar ağaçları vb) kullanır. Nitelik seçme ölçüsü, sınıflandırıcının doğruluk oranıdır. Her bir iterasyonda belirli nitelik altkümesi için sınıflandırma sonucu elde edilir. Burada, *S*, sınıflandırıcıyı, *SD*, sınıflandırma doğruluk oranını ifade etmektedir. Durdurma kriteri filtreleme yönteminde olduğu gibi gerçekleştirilir [48,49]. Sarmal yöntemde, nitelik alt küme uzayı üstsel büyüdükçe sezgisel arama yöntemleri tercih edilir. Sarmal yapıda, model seçimi ile nitelik alt küme araması etkileşimlidir. Ancak filtre yöntemine göre aşırı eğitim riski bulunması ve sınıflandırma maliyetinin fazla olması en büyük dezavantajıdır.

3.4. Genetik Algoritma

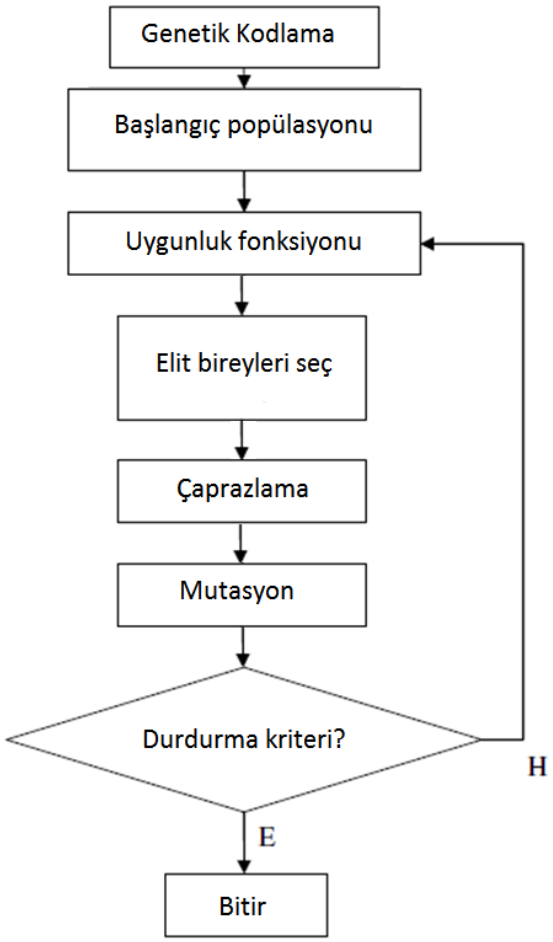
Genetik algoritma, gen seçme işlemlerinde başarılı bir şekilde kullanılan sezgisel bir arama algoritmasıdır. Holland 1975 yılında canlıların bu sürecini sanal ortamda gerçekleştirmek amacıyla ortaya attığı Genetik Algoritma (GA) fikri, yine Holland'ın öğrencisi olan Goldberg'in 1989 yılında yaptığı doktora çalışması ile ilgi odağı haline gelmiştir [50]. Genetik algoritma, biyoinformatikte nitelik seçme işlemlerinde sıklıkla tercih edilmektedir [51-54].

Genetik algoritma (GA), en iyiyi arama aracıdır. En uygun sonucu bulmak için doğal evrim ve seçim teoremini taklit eder. Genetik algoritma, geleneksel optimizasyon yöntemlerine göre parametre kümesini değil bunların kodlanmış biçimlerini kullanırlar [55]. Genetik algoritmanın diğer özellikleri şöyle sıralanabilir:

- GA'da bir nokta yerine aynı anda noktalar topluğunda hareket edilir. Evrim sonucunda en iyi sonuç elde edilmeye çalışılır. Bu evrim sırasında GA yerel en iyiye takılmaz.
- GA, karar vermek için hedef fonksiyonu kullanır. Türev ve integral işlemlerine gerek olmadığından başlangıç ve sınır değerleri gibi kabullere gerek yoktur.
- GA, deterministik değildir. Yani belirlilik değil belirsizlik (rasgelelik, ihtimal) kurallarına dayanır.

Genetik algoritmada amaç, karar uzayında mevcut olan en iyi çözümü bulmaktır. Evrim sayesinde adım adım en iyiye ulaşmaya çalışır. Genetik algoritma en iyiyi bulduğunda son bulmak zorundadır. En iyi çözüm %100'e varmakla olur; ancak bu teorik bir beklentiden ibarettir. Bu nedenle genetik algoritmanın sonlandırılması için belirli bir kriterin olması gerekir. Bu kriter, iterasyon sınırlaması, uzman görüşü, bulanık karar verme veya elde edilen hedef değerler arasındaki farkın önceden belirlenen bir miktardan daha düşük olması ile sağlanabilir.

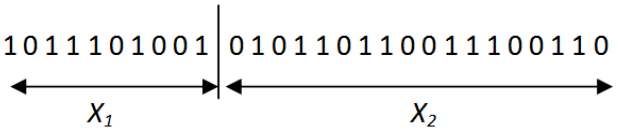
GA şu adımlarla gerçekleştirilir: kodlama, başlangıç popülasyonu, uygunluk fonksiyonu, genetik operatörler (seçim, çaprazlama ve mutasyon). Şekil 3.15'te genetik algoritma için akış diyagramı görülmektedir.



Şekil 3.15. Genetik algoritma akış diyagramı

Stokastik bir yöntem olan genetik algoritmada her bir çözüm adayı *kromozom* (birey) olarak adlandırılır. Kromozomlar bir araya gelerek çözüm uzayını yani *popülasyonu* (topluluğu) meydana getirir. Popülasyon belirlenen fonksiyona bağlı olarak daha iyi olacak şekilde biçimlendirilir. Bu aşamada bazı kromozomlar, topluluktan çıkarılıp yenileri eklenebilir. Kromozomların en küçük parçası (özelliği) *gen* olarak adlandırılır.

Kromozom, kodlanmış bir genetik sayı dizidir. GA'ların kodlanmasında hesaplama kolaylığı sebebiyle ikili sayı sistemi sıklıkla kullanılmasına rağmen, bazı problemler için farklı kodlama sistemi de kullanılabilir. Eğer bir problemin çözümünde x_1 ve x_2 gibi iki değişken varsa bu iki değişken ikili sayı siteminde Şekil 3.16'da görüldüğü gibi kodlanabilir.



Şekil 3.16. Bir kromozomun yapısı

Bu kromozomda x_1 ve x_2 sırayla 10 ve 18 hane (gen) ile temsil edilmektedir. Kromozomların en küçük yapısı *gen* olarak adlandırılır. İkili sayı sistemi ile kodlanmış bir kromozomun genleri *0* ya da *1* olacaktır.

Başlangıç popülasyonu, her biri rasgele belirlenmiş genlerden meydana gelen kromozomlardan oluşur. Probleme göre değişmekle birlikte en az 20 civarında kromozom olması gerekir. Popülasyonun büyüklüğü GA'nın çalışma zamanını artırırken, küçük popülasyon GA'nın verimliliğini düşürür.

Başlangıç popülasyonu çözüm olmaya aday kromozomlardan oluşmaktadır ve bunlardan sadece en iyiler yeni popülasyona dahil edilir. GA'nın yerel en iyiye takılma gibi problemi yoktur. Yeni oluşturulacak popülasyona seçilecek kromozomlar, uygunluk fonksiyonuna göre belirlenir.

Karar uzayında yer alan her bir kromozomun çözüme ne kadar yakın olduğunu belirlemek için *uygunluk fonksiyonuna* (Amaç fonksiyonu) ihtiyaç vardır. Uygunluk fonksiyonu, dinçlik fonksiyonu olarak da bilinir ve kromozomun yeni popülasyonda yer almasına bu fonksiyonun sonucu karar verilir.

GA, en iyi çözümü bulmak için her yeni nesilde mevcut popülasyonda yer alan en iyileri seçme işlemini de yapar. Böylece sonraki popülasyonda GA işlemlerine tabi tutulacak yeni kromozomlar belirlenir. Yeni popülasyonu, önceki popülasyon içinde hayatta kalmayı başaran en iyi kromozomlar oluşturur. Böylece yeni oluşturulacak nesil daha güçlü olabilir. Sıklıkla tercih edilen yöntemlerden birisi uygunluk değeri yüksek olan kromozomlar seçilerek uygunluk değeri küçük olan kromozomlar üzerine kopyalanır. Daha sonra çaprazlama ve mutasyon gerçekleştirilir. Böylece uygunluk değerleri düşük olan kromozomlar yerine uygunluk değeri yüksek olan kromozomlar baskın yeni bir popülasyon oluşturulmuş olacaktır. Popülasyon içinde kromozomların seçim işlemi için farklı yöntemler mevcuttur. Rastsal seçim, rulet tekerleği, turnuva seçimi sıklıkla kullanılan yöntemlerdir.

Rastsal seçim, popülasyon içinden eşleştirmenin rasgele yapıldığı yöntemdir. Herhangi bir kromozom herhangi diğer bir kromozomla eşleştirilebilir.

Rulet tekerleği, kromozomların uygunluk değerleri hesaba katılarak oluşturulan rulet çarkında kromozomların rasgele seçilmesi temeline dayanır. Çark çevresi uzunluğu U, seçim sayıları s_1 , s_2 , … , s_n kabul edilirse, i_1 , i_2 , … , i_n seçilme ihtimali şöyle hesaplanabilir:

$$s_1 + s_2 + \ldots + s_n = U \tag{3.33}$$

$$\frac{s_1}{U} + \frac{s_2}{U} + \ldots + \frac{s_n}{U} = \frac{U}{U} \tag{3.34}$$

$$i_1 + i_2 + \ldots + i_n = 1 \tag{3.35}$$

olacaktır.

Her kromozomun seçilme ihtimali, Eş. 3.35 ile elde edilmiş olur. Böylece uygunluk değeri yüksek olan kromozomların seçilme ihtimali yüksek olacaktır.

Turnuva seçimi, rasgele belirlenen birkaç kromozom içinde uygunluk değeri en yüksek olan birinci gelir. Benzer şekilde diğer eş de seçilerek eşleştirme gerçekleştirilir.

Bir önceki nesildeki en güçlü adayların yeni oluşturulan popülasyona dahil edilmesi *elitizm* olarak adlandırılır. Böylece bir şekilde seçilememiş önceki neslin güçlü adayları yeni popülasyona dahil edilmiş olacaktır. Bu amaçla popülasyon içindeki en zayıf kromozomlar yok edilerek yerine bir önceki neslin en güçlü kromozomları yerleştirilir.

GA'da iki kromozom eşleştirildiğinde yeni iki tane kromozom elde edilir. Eşleşme, toplumdan atılan daha kötü kromozomların yerine daha güçlü adaylar geçinceye kadar devam eder. GA'da iki kromozomun eşleştirilmesi ile iki kromozomun elde edilmesi işlemi *çaprazlama* olarak adlandırılır. Çaprazlama, eşleştirilen kromozomların genlerinin birbirleri ile karşılıklı yer değiştirmesi işlemidir. Hangi genlerden itibaren değiştirileceği rasgele seçilebilir. Burada amaç önceki nesilden farklı toplumlar elde etmektir. Çaprazlama sonucunda yeni iki birey elde edilir. Eşleştirmeden sonra kromozom sayısı böylece sabit kalır.

Farklı çaprazlama yöntemleri vardır. Tek kesimli çaprazlamada, öncelikle rasgele bir kesim noktası bulunarak, genler karşılıklı çapraz değiştirilir. Çift kesimli çaprazlamada ise tek kesimli çaprazlamaya benzer, kesim kromozom boyunca iki noktada gerçekleştirilir. Bu yöntemde kromozom üç parçaya bölünmüş olur. Bu parçalardan karşılıklı her ikisi çaprazlama yer değiştirmesi sonucu altı yeni kromozom elde edilir. Yeni kromozomların hepsi kullanılmayabilir. İstenilen sayıda yeni kromozom yeni topluma katılabilir. Çok kesimli çaprazlamada, kromozom ikiden fazla noktadan kesilir ve genler yer değiştirilir. Böylece çok sayıda kromozom elde edilmiş olacaktır. Bire bir çaprazlamada ise kromozomlarda karşılıklı rasgele belirlenen genler arası çaprazlama gerçekleştirilir.

Mutasyon (değişim), tek kromozom üzerinde yapılan GA işlemidir. Mutasyon, kromozomun genlerinde yapılan değişikliktir. Bir veya birden fazla gen

değiştirilebilir. Böylece başlangıçta bulunmayan yeni kromozom türleri elde edilir. Genelde her iterasyonda hanelerin %1 ile %0.1 miktarında değişim yapılır. Genelde en iyi kromozomlarda mutasyon önerilmez [50].

4. MEME KANSERİ VE GENETİK

Günümüzde birçok ülkede kalp damar hastalıklarından sonra en çok ölüme neden olan kanser, hücrelerin kontrol dışı çoğalması sonucu ortaya çıkan genetik bir hastalıktır [56,57]. Ancak bu hastalığı tek bir gene bağlayarak açıklamak mümkün değildir. Birden fazla gen kanser gelişim riskini arttırmakta, çevresel faktörler bu işi daha da karmaşık hale getirmektedir.

Günümüzde mikrobiyoloji alanında çok önemli gelişmeler yaşanmıştır. Bu sayede hastalıkların temelinde yatan genetik faktörler hakkında daha fazla veri elde edilebilmekte ve bu verilerin analizi üzerine yapılan çalışmalar artarak devam etmektedir. Genetik faktörlerin belirlenmesi, hastalıkların erken teşhis edilmesi ve yeni tedavi yöntemlerinin geliştirilmesinde çok önemlidir [58].

Kanser gelişimine birden fazla gen neden olmaktadır. Onkogenler, kansere neden olan genlerdendir. Bazı kanser türlerinde birden fazla onkogen rol oynayabilir. Proto-onkogenler ise büyümeyi artırıcı genlerdir ve normal hücre gelişimi için gereklidir. Ancak bu genler, aktif olduklarında onkogenlere dönüşür. Günümüzde kırk civarında proto-onkogen tespit edilmiştir. Onkogen aktivasyon mekanizmaları, sonradan olan somatik mutasyonlardır. Dolayısıyla onkogenler ailevi kanserlerin oluşumunda rol almazlar. Ailevi kansere yatkınlık, ikinci bir tür gen grubundaki mutasyon ve bunların kalıtımı ile gerçekleşir. Bu genlere *tümör baskılayıcı genler* ya da *anti-onkogen* denilmektedir. Çizelge 4.1 ‘de Ailesel kanser sendromları ve bunlarla ilişkili tümör baskılayıcı genler listelenmiştir [59].

Çizelge 4.1. Ailesel kanser sendromları ve bunlarla ilişkili tümör baskılayıcı genler.

Ailesel Kanser sendromu	Tümör Baskılayıcı Gen	İşlevi	Kromozom Lokasyonu	Gözlenen Tümör Tipi
Ailesel Retinoblastoma	RB1	Hücre siklusu düzeni	13q 14	Retinoblastoma, osteogenik sarkoma
Li-Frumeni sendromu	P53	H. siklusu düzeni ve apoptozis	17p 13	Beyin tümörleri, sarkomalar, lösemi, Meme kanseri
Ailesel adenomatoz polipozis	APC	Hücre yüzeyinde nükleusa adezyon	5q 21	Kolon kanseri
Von Hippel-Lindau sendromu	VHL	Transkripsiyonal uzama düzenlemesi	3p 25	Renal kanserler, hemangioblastoma,
Vilms tümörü	WT1	Transkripsiyonal düzen	11p 13	Pediatrik böbrek kanseri
Ailesel melanoma	CDKN2A	Hücre siklusu düzenlemesi	9p 21	Melanoma, pankreatik kanser ve diğerleri
Kalıtsal nonpolipozis	MLH2	DNA mismatch tamiri	2p 16	Kolon kanseri
Kolon kanseri	MLH1	DNA mismatch tamiri	3p 21	Kolon kanseri
Ailesel meme kanseri	BRCA1	Bilinmiyor	17q 21	Meme ve ovaryum kanseri
Ailesel meme kanseri	BRCA2	Bilinmiyor	13q 12	Meme ve ovaryum kanseri
Gorlin sendromu	PTCH	Erken gelişim ve hücre farklılaşması	9q 22	Bazal hücre deri kanseri
Multipi endokrin neoplazi tip 1	MEN1	Bilinmiyor	11q 13	Paratiroid ve ptüiter adenomlar, islet
Neurofibramatozis Tip1	NF1	RAS inaktivasyonu katalizi	17q 11.2	Nörofibromalar, sarkomalar, glimolar

Son 10 yılda kanser genetiği ile ilgili ciddi gelişmeler yaşanmıştır. Genetik çalışmalar ve yeni teknolojiler kanser için umut olsa da gerçekte henüz kanser tedavisi için somut adım atılamamaktadır. Günümüzde kanser tedavileri için halen 1960'lı yıllarda belirlenen cerrahi, radyoterapi ve kemoterapi yöntemleri uygulanmaktadır.

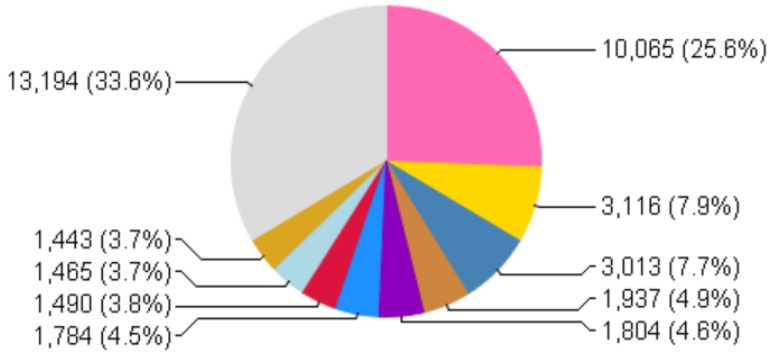
DNA testleri ile kansere yatkınlık genlerinin belirlenmesinde karşılaşılan problemler şunlardır:

- Tümör baskılayıcı genlerinin karmaşık yapısı
- Popülasyonda birden fazla farklı mutasyon olması
- Etkin olan mutasyon ile etkin olmayan mutasyonların ayırt edilememesi
- Yüksek maliyetli testler
- Test sonuçlarının yorumlanmasında karşılaşılan zorluklar [59].

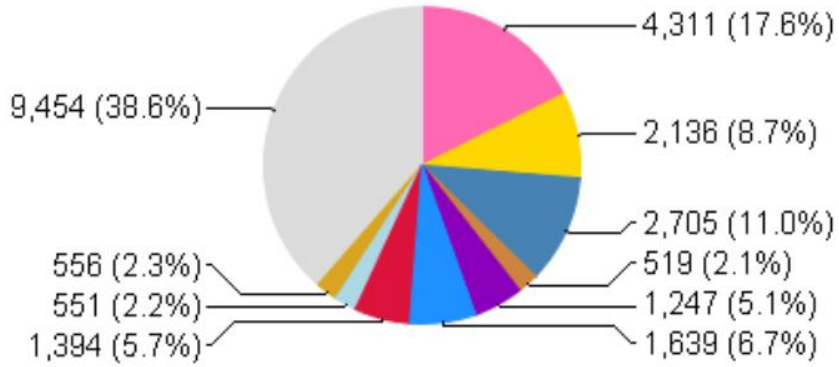
Yapılan tez çalışmasında, meme kanserine neden olan genetik faktörlerin belirlenmesi amacıyla gen ifade verilerinden faydalanılmıştır. Gen ifadesi (gene expression), sağlıklı ve hasta dokular arasındaki farklılığın tespitinde kullanılmaktadır. Bu bölümde meme kanseri ve genetik alanından bahsedilmiştir.

4.1. Meme Kanseri

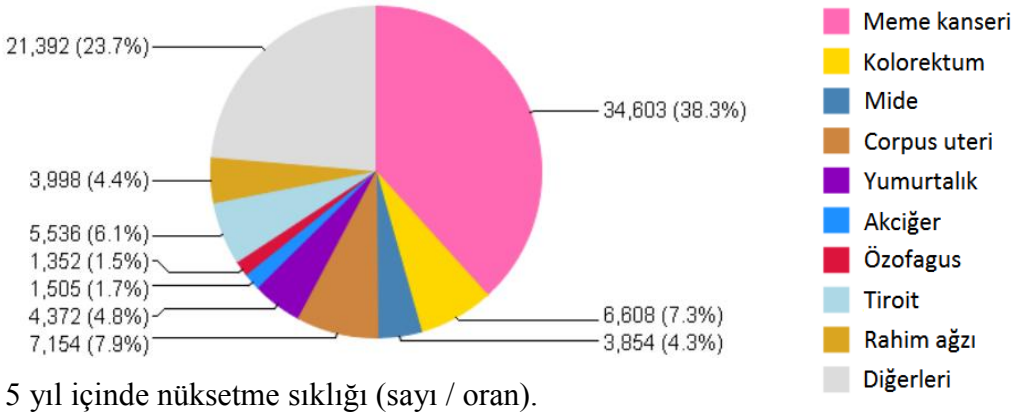
Meme kanseri kadınlar arasında ölüme neden olan kanser türlerinde ilk sırada yer almaktadır. GLOBOCAN 2008 verilerine göre meme kanserinden ölüm oranı Türkiye'de %17,6 Amerika'da %12,7 ve Çin'de %6,1 olarak açıklanmıştır [1]. Şekil 4.1'de Türkiye'de meme kanseri ve diğer kanser türlerinin görülme sıklığı ve oranları verilmektedir. Şekil 3.2 (a)'da meme kanseri görülme sıklığı, Şekil 3.2 (b)'de ölümle sonuçlanma ve Şekil 3.2 (c)'de 5 yıl içinde nüksetme değerleri görülmektedir. Meme kanserinin erkeklerde görülme sıklığı, kadınlarda görülme oranının %1'i kadardır [60].



a) Görülme sıklığı (sayı / oran).



b) Ölümle sonuçlanma (sayı / oran).



c) 5 yıl içinde nüksetme sıklığı (sayı / oran).

Şekil 4.1. Dünya sağlık örgütü (WHO) kanser araştırma enstitüsü 2008 verilerine göre a) Çeşitli kanser türlerinin görülme sıklığı, b) Ölüm oranı, c) 5 yıl içinde nüksetme oranı [1]

Pek çok kanser türünde olduğu gibi meme kanserinde de hastalığa sebep olan faktörler tam olarak anlaşılamamıştır. Ancak ailevi kalıtımın meme kanseri riskini arttırdığı bilinmektedir. Menopoz öncesi bilateral meme kanseri görülen kadınların yakın akrabalarında risk belirgin bir derecede artmaktadır. Bu kadınların %50'den fazlasında meme kanseri geliştiği görülmektedir. Meme kanseri olmuş kadınlarda, iyileşmeden sonra ikinci memede kanser gelişme riski her sene %1 artar [61,62].

Meme kanseri gelişiminde virüslerin rolü olabileceği bazı deneysel çalışmalarda gösterilmiştir. Ancak bunun etkisi değişik türlerde genetik yatkınlığa ve endokrin faktörlerine bağlanmıştır.

İnvazif meme kanserlerinin çoğu lobül epitelinde ve süt kanallarında gelişir. Lobüler karsinoma mikroskopik bir lezyon olduğundan klinik olarak saptanması zordur. Patalog tarafından biyopsi ile teşhis edilebilir. Vakaların %38'inde invazif tümör gelişir. Biyopsi yerinde, biyopsi yapılan memenin başka yerinde veya diğer memede gelişebilir. İntrduktal karsinoma olarak da bilinen duktal karsinoma kolaylıkla teşhis edilebilir. Vakaların %70'inden biyopsi yerinde invazif hale gelir. Çoğunlukla çok merkezli (multicentric) ve vakaların %1-3'ünde nodal metastaz gelişir.

İnvaziv karsinomlar, meme kanserinin %60'ını oluştururlar ve büyük bir çoğunluğu başka bir şekilde tanımlanamayan karsinomlardır. Medüller karsinomlar, infiltratif kanserlerin %5'ini oluşturur ve daha geç metastaz yaparlar. Jelatinöz, tübüler ve papiller karsinomlar daha nadir görülür. İnflamatuvar karsinom, lenf yoluyla çabuk ve erken yayılır.

Meme kanseri gelişimi çok uzun olup, çoğu vakada ilk malign hücrenin ortaya çıkmasından 1 cm çapında bir hacme ulaşması 7-8 senelik sürede gerçekleşir. Meme kanseri lenfatiklerle ve kan yoluyla yayılır. Lenf bezleri primer tümörden kaçan hücreler için bir filtre görevi görür. Bir kez aksilerden geçtiğinde tümör hücreleri supraklavikular lenf bezlerini tutar ve sonra damar (venöz) dolaşımına karışır. Kan yoluyla yayılım sıktır ve iskelet sistemi, akciğer, karaciğer ve beyin metastazları görülebilir [57].

4.1.1. Meme kanseri tarama ve tanı

Meme kanseri tarama ve tanıda genellikle kullanılan yöntem fiziksel muayene ve mamografidir [63,64]. Mamografi tekniği, fiziksel muayenede ele gelen lezyonları olmayan asemptomatik kadınlarda sıklıkla kullanılmaktadır. Ayrıca yeni geliştirilen cihazlarla radyasyona maruz kalma riski 0,25 rad (cGy)'a kadar düşürülmüştür.

Küçük tümörler, meme kanseri tanısını zorlaştırmaktadır. Bu tür hastalarda uzmanların deneyimi önem kazanmaktadır. Meme kanseri çoğunlukla fiziksel muayenede fark edilebilir. Ancak nadiren ilk semptomu aksiller, kemik ve akciğer metastazı olabilir.

Mamografi dışında ultrasonografi de meme kanseri taramalarında kullanılmaktadır. Mamografi, yoğun meme dokusu olan kadınlarda ultrasonografinin mamografiye ek olarak meme kanserini saptayabildiğini gösteren araştırmalar vardır [65,66].

Fiziksel muayene sıklıkla kullanılmasına rağmen, mamografi tümörün büyüklüğü hakkında kesin bilgi vermekle beraber, klinik olarak saptanmayan diğer neoplastik lezyonların varlığını da bildirir. Deneyimli uzmanlar eşliğinde gerçekleştirilen aspirasyon biyopsi, hastalığa dair önemli bilgiler verebilir [61].

4.1.2. Meme kanseri evrelendirmesi

Doğru ve erken tanı hastalığın tedavisi açısından çok önemlidir. Bu sebeple hastalığın doğru evrelendirilmesi gerekir. Bu aşamada hem klinik hem de patolojik evrelendirme önem kazanır. Günümüzde kullanılan sistem TNM evrelemedir [67,68]. Çizelge 4.2'de TNM tedavi öncesi klinik evrelendirme çizelgesi görülmektedir.

Çizelge 4.2. TNM tedavi öncesi klinik evrelendirme çizelgesi

Primer Tümör (T)	
T_x	Primer tümör saptanamamaktadır
T_0	Primer tümör yok
T_{is}	Karsinoma in situ
$T_{is(DCIS)}$	Duktal karsinoma in situ
$T_{is(LCIS)}$	Lobuler karsinoma in situ
$T_{is(LCIS)}$	Meme başının kitlesiz Paget hastalığı (Tümörlü Paget hastalığında sınıflama tümörün boyutuna göre yapılır)
T_1	Tümörün en büyük boyutu 2 cm veya daha az
T_{1mic}	En büyük boyutu 0.1 cm veya daha az olan mikroinvazyon
T_{1a}	En büyük boyutu 0.1 cm.den büyük olan ancak 0.5 cm.yi geçmeyen tümör
T_{1b}	En büyük boyutu 0.5 cm.den büyük olan ancak 1 cm.yi geçmeyen tümör
T_{1c}	En büyük boyutu 1 cm.den büyük olan ancak 2 cm.yi geçmeyen tümör
T_2	En büyük boyutu 2 cm.den büyük olan ancak 5 cm.yi geçmeyen tümör
T_3	En büyük boyutu 5 cm.den büyük olan tümör
T_4	Herhangi bir boyutta ancak (a) göğüs duvarına veya (b) cilde direkt yayılım
T_{4a}	Pektoral kasa ulaşmamış göğüs duvarı yayılımı
T_{4b}	Meme cildinde ödem veya ülserasyon, veya aynı memede satellit deri nodülleri
T_{4c}	T_{4a} ve T_{4b} birlikte
T_{4d}	Enflamatuar karsinom

Bölgesel Lenf Nodları (N)	
N_x	Bölgesel lenf nodları saptanamamaktadır (daha önce çıkarılmış)
N_0	Bölgesel lenf nodu metastazı yok
N_1	Ipsilateral lenf nodu veya nodlarında metastaz
N_2	Fikse veya gruplaşmış ipsilateral aksiller lenf nodlarında metastaz veya klinik olarak belirgin aksiller lenf nodu metastazı olmadığı durumlarda klinik olarak belirgin ipsilateral internal mammaryal nodlarında metastaz
N_{2a}	Birbirlerine veya çevre dokulara fikse ipsilateral aksiller lenf nodlarında metastaz
N_{2b}	Sadece klinik olarak aksiller lenf nodu metastazı olmadığında klinik olarak belirgin ipsilateral internal mammaryal nodlarda metastaz olduğunda

Çizelge 4.2. (Devam) TNM tedavi öncesi klinik evrelendirme çizelgesi

N_3	Aksiller lenf nodu tutulumu olsun ya da olmasın ipsilateral infraklavikular lenf nodları metastazı veya klinik olarak belirgin ipsilateral internal mammaryal lenf nodları metastazı ile birlikte klinik olarak belirgin aksiller lenf nodu metastazı; veya aksiller ya da internal mammaryal lenf nodu metastazı olsun ya da olmasın ipsilateral supraklavikular lenf nodları metastazı
N_{3a}	Ipsilateral infraklavikular lenf nodlarında metastaz
N_{3b}	Ipsilateral internal mammaryal lenf nodlarında veya aksiller lenf nodlarında metastaz
N_{3c}	Ipsilateral supraklaviküler lenf nodlarında metastaz

Uzak Metastazlar (M)	
M_x	Uzak metastazların varlığı değerlendirilemeyebilir
M_0	Uzak metastaz yok
M_1	Uzak metastazlar mevcut

Hastalığın yayılımı ve ciddiyeti hakkında bilgi edinilmesini sağlayan TNM evreleme sisteminde tümör boyutu (T), aksiler lenf nodlarına yayılım (N) ve uzak bölgelere yayılım (M) olmak üzere üç kritere göre yapılır. TNM evrelemesi meme kanseri hastalarında tedaviye yön veren önemli bir araçtır. Yeni teknikler geliştirildikçe kanser evrelemesinde doğruluk artar ve hastalığın tespiti ve bunun sonucunda da tedavisi daha doğru yapılabilir.

AJCC (American Joint Committee on Cancer), periyodik olarak evreleme standartlarını günceller [69]. 2003 yılında meme kanseri evrelemesinde de yeni düzenlemeler yapmıştır. Lenf nodları, bölgesel metastazların boyutu, sayısı ve saptanma metotları 1997 yılındakinden farklılıklar gösterir. Çizelge 4.3'te 2003 yılı TNM sınıflamasına göre evrelerin gruplandırılması görülmektedir [70].

Çizelge 4.3. Meme kanserinde TNM sınıflamasına göre evrelerin gruplandırılması.

Evre 0	T_{is}	N_0	M_0
Evre I	T_1*	N_0	M_0
Evre IIA	T_0	N_1	M_0
	T_1*	N_1	M_0
	T_2	N_0	M_0
Evre IIB	T_2	N_1	M_0
	T_3	N_0	M_0
Evre IIIA	T_0	N_2	M_0
	T_1*	N_2	M_0
	T_2	N_2	M_0
	T_2	N_2	M_0
	T_3	N_1	M_0
	T_3	N_2	M_0
Evre IIIB	T_4	N_0	M_0
	T_4	N_1	M_0
	T_4	N_2	M_0
Evre IIIC	Herhangi T	N_3	M_0
Evre IV	Herhangi T	Herhangi N	M_1

T_1*, T_{1mic} de içerir.

4.1.3. Meme kanserinde genlerin rolü

Meme kanseri, genetik bir hastalıktır. Sporadik kanser için iki mutasyon gerekli iken, ailesel kanserlerin oluşumu için tek mutasyon yeterlidir. Günümüzde yapılan çalışmalar tüm kanser türlerinin oluşumunda üç gen sınıfının etkili olduğunu ortaya çıkarmıştır. Bu genler proto-onkogenlerin mutasyonu ortaya çıkan onkogenler, tümör supresör genleri ve DNA tamir genleridir [71].

Tümör belirleyiciler, bazı kanser tiplerinde tarama ve tanı koymada kullanılmaktadır [72]. Tümör belirleyicileri, kan veya vücut sıvılarında tümör tarafından üretilen veya ilişkili olarak ortaya çıkan maddeler olarak tanımlanır. Tümör belirleyiciler, tümör tarafından üretilebildiği gibi, vücudun tümör dokusuna karşı ürettiği maddeler de olabilir. Ancak çoğu kez tek başına tanı koymada kullanılmazlar.

Ucuz, kolay ve doğruluğu yüksek tümör belirleyicileri hastalığın tedavisinde önemlidir. Tümör belirleyiciler klinikte şu amaçlar için kullanılır:

- Neoplazi taranması
- Kanser tanısı
- Kanser sınıflandırma
- Prognozun belirlenmesi
- Tedavi izleme
- Rekürrens takibi
- Metastaz takibi

Duyarlılığı ve seçiciliği yüksek bir tümör belirleyicisi erken tanı sağlar, tedavi başarısını arttırır. Tümör belirleyicisinin duyarlılığı, meme kanseri olan hastaların büyük bir yüzdesini doğru tespit etmesidir. Seçiciliği ise meme kanseri olmayan hastaların büyük bir yüzdesinin doğru tespit etmesidir [71].

4.2. Genetik

Yaşamın şifresi olarak adlandırılan DNA, hücrelerde meydana gelen biyolojik faaliyetlerden sorumludur. Gen ise protein veya RNA sentezi için gerekli bilgiyi içeren DNA parçasıdır. DNA'nın kalıtsal materyal olduğu ilk kez 1944 yılında Avery, Macleod ve McCarty tarafından kesin olarak kanıtlanmıştır. Friedrich Miescher'in 1869'da yapmış olduğu deneyler sonucunda tespit ettiği zayıf asit günümüzdeki modern moleküler genetiğin çalışmalarının merkezine yerleşen DNA olmuştur [20,59].

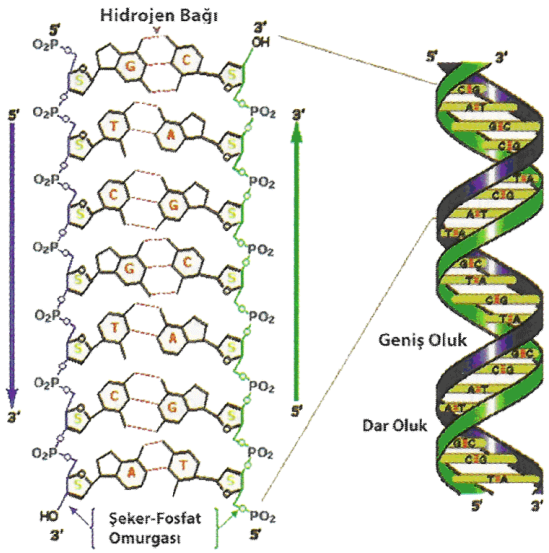
4.2.1. Genel DNA yapısı

Bir organizmada meydana gelen biyolojik etkinlikler DNA'dan RNA'ya (Ribonükleik Asit) ve RNA'dan da proteine aktarılan genetik bilgi akışıyla gerçekleştirilir. Hücreler arasında genetik bilgi genler arcılığıyla taşınır [73]. Bir hücrede, her bir proteinin aminoasit ve RNA molekülünün nükleotit dizilimi, hücrenin kendi DNA'sı tarafından belirlenir. Bir protein ve RNA için gerekli olan nükleoitit dizileri DNA'da bulunmaktadır. Gen, protein veya RNA sentezi için

gerekli bilgiyi içeren DNA parçasıdır. Her bir hücre binlerce gene sahiptir. DNA'daki nükleotit dizilimi ve onun temsil edildiği protein dizilimi arasındaki ilişki *genetik kodlama* olarak ifade edilir [20].

Bir organizmanın genetik yapısına *genotip* adı verilirken, kalıtsal bir özelliğin fiziksel görünümüne *fenotip* adı verilmektedir. Genler, fenotip karakterin gelişiminde sihirli güçtür. Ancak fenotip, yalnızca genler ve onların etkileşimlerinden değil aynı zamanda çevresel faktörlerden de etkilenir. Örneğin bir kişinin boyu, kilosu pek çok genle denetlenir. Ancak bu genlerin ifadesi iç (ergenlik çağındaki hormonlar vs.) ve dış (beslenme vs.) çevresel etmenlerden oldukça etkilenir. *Mutasyon* ise genetik mesajı değiştiren kalıtsal bir olaydır [59].

DNA'nın biyokimyasal olarak araştırılması Fried Miescher tarafından ilk kez 1868 yılında yapılmıştır. DNA'nın genetik bilgi taşıdığına dair ilk kanıtlar ise 1944 yılında Osward T.Avery, Colin Macleod ve Marclyn McCarty tarafından bulunmuştur. 1952 yılında Alfred D.Hersey ve Martha Chase tarafından gerçekleştirilen deneyler, canlı hücrelerde genetik bilginin kromozomların önemli bir bileşeni olan DNA tarafından taşındığını ortaya koymuştur [20]. Şekil 4.2'de anti-paralel DNA zincirlerinin hidrojen bağı ile bağlanarak DNA çift sarmalı oluşturması görülmektedir.



Şekil 4.2. Anti-paralel DNA zincirlerinin hidrojen bağı ile bağlanarak DNA çift sarmalı oluşturması [20]

Genetik kodun doğru olarak aktarılması *replikasyon* olarak adlandırılır. Replikasyon ile iki DNA ipliğinin ayrılması ve her bir tamamlayıcı ipliğin baz eşleşmesi kuralları içinde nükleotitlerin birbirlerine eklenmesi yoluyla sentezlenmesi aşamalarını içerir.

Bir organizmanın sahip olduğu tüm DNA bilgisi *genom* olarak adlandırılır. Genomdaki toplam DNA miktarı, her bir organizma için karakteristik olup, *c-değeri* olarak adlandırılmaktadır.

Genetik kod, mRNA boyunca *kodon* olarak adlandırılan üçlü gruplar halinde bulunan ve protein sentezleme sırasında üretilen aminoasit dizilerinin düzenini belirleyen nükleotit dizilerdir. Kodon üzerindeki her bir pozisyon için dört nükleotitten biri bulunabileceğinden 20 amino asit için 4^3=64 farklı muhtemel üçlü nükleotit dizilimi vardır [20].

4.2.2. Genel gen yapısı

Bireyin sahip olduğu özellikler (fenotip) *gen* olarak adlandırılan kalıtım elementleri ile ebeveynlerinden aktarılmaktadır. Kromozomlar üzerinde yer alan genler, RNA moleküllerinin şifrelenmesinden sorumlu genetik bilgiyi taşır. Daha geniş anlamda gen, organizmanın karakterlerinin ortaya çıkmasında sorumlu olan proteinlerin birincil yapıları olan *polipeptitlere,* tRNA, rRNA ve diğer RNA moleküllerine ait bilgileri taşıyan genetik bilgi üniteleridir [73].

Genlerin kalıtımı nesilden nesile nasıl aktardığını, 1866 yılında Gregor Mendel ispatlamıştır. Friedrich Miescher'in 1869'da yapmış olduğu deneyler sonucunda tespit ettiği zayıf asit günümüzdeki modern moleküler genetiğin çalışmalarının merkezine yerleşen DNA olmuştur. Ancak 20.yy ortalarına kadar gen ve DNA tam olarak bir araya getirilememiştir.

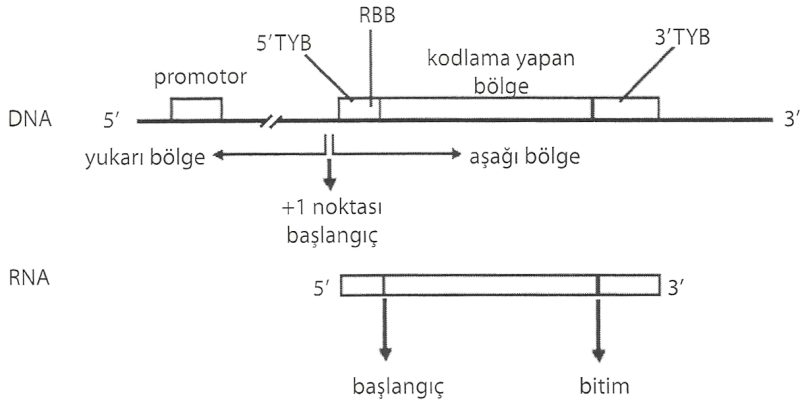
Alman botanik bilimci Hans Winkler 1920 yılında GEN ve kromozOM terimlerini bir araya getirerek GENOM kelimesini terminolojiye katmıştır. Genom, bir

organizmanın yapısı, işlevleri, işlevlerinin devamlılığı ve kalıtım için gerekli biyolojik bilgi olarak tanımlanabilir [59].

Genler, kromozomlar üzerinde rasgele dizilmiş nükleik asit zinciri değildir. Genler, genetik bilgiyi RNA moleküllerine aktarma görevleri dışında birçok işlevsel birimin bir araya gelmesinden oluşmaktadır. DNA molekülü üzerinde gen ifadesinin gerçekleştirilmesinde işlevsel bölgeler vardır. Bunlar arasında promotor, gen ifadesinin ilk aşaması olan transkripsiyonun yapılabilmesi için RNA polimeraz enziminin geni tanıyıp bağlandığı bölgedir. Transkripsiyonun bitim noktaları ve transkripsiyon ile translasyonun kontrol edildiği bölgeler gibi daha pek çok bölge gen yapısını oluşturmaktadır. Genlerin yapı ve işlevlerini inceleyen çalışmalar moleküler genetik içinde gerçekleştirilmektedir.

Genin, kromozom üzerinde kapladığı alana lokus adı verilir. Ökaryotlarda, genin farklı formları *allel* olarak adlandırılmaktadır. Genler işlevlerine göre, yapısal genler, RNA genleri, düzenleyici genler olmak üzere 3 sınıf altında toplanabilir [20].

DNA molekülü üzerinde gen ifadesinin gerçekleştirilmesinde aktif olan bölgeler Şekil 4.3.'te görülmektedir.



Şekil 4.3. Temel gen yapısı

Şekil 4.3'te de görüldüğü gibi promotor, RNA polimeraz enziminin geni tanıyıp bağlandığı bölgedir. Translasyonda ribozomun, primer transkripsiyon ürünü olan mRNA (mesaj taşıyan RNA-messenger RNA) molekülüne bağlandığı ve RBB olarak

adlandırılan ribozama bağlanma bölgesi de burada yer alır. Ayrıca, bir gen zincirinde polipeptit zincirinde karşılığı olmayan 5'translasyonu yapılmayan bölge (5'TYB) ve 3'translasyonu yapılmayan bölge (3'TYB) nükleotit dizileri de burada yer almaktadır [20].

Canlı bir organizmanın genomunun ifade edilmesinde kullanılan ilk ölçüt genom boyutudur. Genom boyutu hücrenin haploitindeki DNA miktarıdır ve *c değeri* olarak gösterilir. Biyolojik anlamda *c değeri* karşılaştırmalı çalışmalarda sıklıkla kullanılmaktadır. Genom boyutu, gelişmiş organizma ile basit organizma arasında ciddi farklılık gösterir. Çizelge 4.4'te farklı organizmaların c değeri, DNA moleküler yapısı ve haploit kromozom sayısı görülmektedir.

Çizelge 4.4. Farklı organizmalara ait c değeri, DNA molekülünün formu ve haploit kromozom sayısı.

Genom	**c değeri**	**DNA molekülünün şekli**
Virüs		
SV40	5 kb	Çift zincirli halkasal DNA
Herpes simpleks	152 kb	Çift zincirli halkasal DNA
Prokaryot		
Escherichia coli	4600 kb	Çift zincirli halkasal DNA
Borrelia burgdorferi	910 kb	Çift zincirli halkasal DNA
Ökaryot		**Haploit kromozom sayısı**
Saccharomyces cervisiae	13 Mb	16
Caenorhabditis elegans	97 Mb	6
Arabidposis thaliana	100 Mb	5
Drosophile melanogaster	180 Mb	4
Homo sapiens	3 000 Mb	23
Zea mays	4 500 Mb	10

Günümüzde, genom bilgisinin farklı fenotipler ortaya koyduğu yani genlerde yazılı olan şifrelerin aminoasit dizilerine nasıl aktarıldığı, aminoasit dizilerinin de proteinlerin üç boyutlu yapısında nasıl etkin olduğu bilinmektedir.

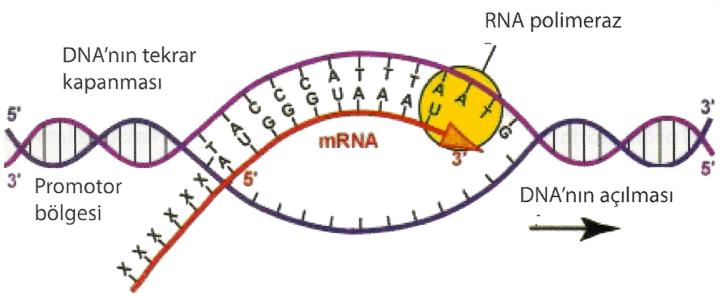
Genom projeleri sayesinde pek çok organizmanın nükleotit dizileri tamamen açıklanmıştır. Bu alanda yapılan çalışmalar, *Genomik,* yani genomu oluşturan genleri ve onların işlevlerini aydınlatmak üzere yapılan çalışmalar, *transkriptomik,* belirli bir dokuda belirli bir zamanda aktif olmuş genlerin ürünleri olan transkriptleri araştıran çalışmalar, *proteomik*, genom tarafından kodlanan yeni terimler üzerine yapılan çalışmalar bunlardan bazılarıdır.

Günümüzde mikro DNA çipleri (microarray) kullanılarak farklı organizmaların genom yapıları karşılaştırılabilmektedir. Bu sayede belirli hastalıklara neden olan genler ve bu genlerin ifadeleri elde edilebilmektedir [74].

4.2.3. Genetik bilginin ifade edilmesi: Transkripsiyon

Genetik bilgi ve genetik bilginin çevresel şartlara vermiş olduğu cevap, bir organizmanın sahip olduğu özellikleri ifade eder. Genetik bilginin ifadesi, transkripsiyon ve translasyon olarak ifade edilen iki basamaklı protein sentezi ile gerçekleştirilir.

Şekil 4.4'te görüldüğü gibi transkripsiyon, bilginin, haberci RNA (mRNA) olarak adlandırılan RNA molekülü formuna dönüştürülmesidir. Transkripsiyon ile DNA replikasyonu birbirine çok benzer. Ancak transkripsiyonda replikasyona göre daha küçük bir molekül üretilir ve DNA'nın sadece bir ipliği kopyalanır. Bu sebeple transkripsiyon, replikasyona göre daha basittir.



Şekil 4.4. Bir gene komplementer mRNA'nın transkripsiyonu

4.2.4. Mikrodizi teknolojisi

Gen ifade verisi mikrodizi olarak adlandırılan bir teknoloji ile elde edilmektedir. Mikrodizi teknolojisi, *Southern hibritleme* tekniğine dayanmaktadır. 1975 yılında Profesör E.M. Southern tarafından keşfedilen ve kendi adıyla anılan *Southern hibritleme* tekniği belirli büyüklükteki DNA parçasını saptamaya yarar. Bu işlemde, çift iplikli DNA parçaları, jel elektroforezi yardımıyla ayrıldıktan sonra denatüre edilir. *Denatürasyon* ya da erime, DNA çift sarmalının iki ipliği arasındaki tüm hidrojen bağlarının kırılarak, sarmalların birbirinden ayrılması işlemidir. Jelde ayrılan DNA bantları, jeldeki pozisyonlarında membrana aktarılmış olur. Daha sonra membrana, radyoaktif işaretli ve ilgilenilen genle homoloji gösteren bir probla hibritlenir. Böylece prob DNA, sadece kendisinin komplementeri olan DNA parçası ile hibritleşir. Kullanılan radyoaktif işaretli prob, X-ışınına maruz bırakılarak fotoğraflanır böylece probun bağlandığı yer tespit edilir. Bu işleme *otoradyografi* ve elde edilen görüntüye de *otoradyogram* denir. Mikrodizi analizinin temeli Southern ve Northern hibritleme tekniğine dayanır.

Bir dizi, bilinen DNA sekanslarının katı bir substrat üzerine düzenli bir şekilde yerleştirilmesi olarak tanımlanır. Bir mikrodizi, mikroskopik lamın üzerine yazılmış 200 mikrondan daha küçük çapta binlerce noktadan oluşur. Mikrodizi, görüntüleme için özel yüksek çözünürlüklü tarayıcılara ihtiyaç duyar. Mikrodizi, Southern ve Northern hibritleme tekniğine dayanmasına rağmen, eş zamanlı olarak binlerce gen sekansı aynı anda ölçülebilir. Mikrodiziler, mikroskopik spotlar içerir ve her bir spot

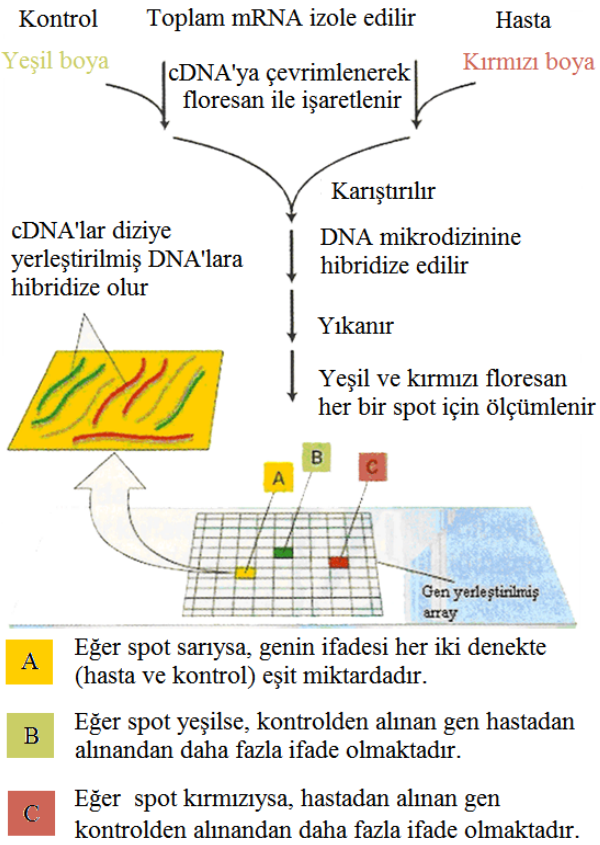
ayrı bir test için dizayn edilmiştir. Mikrodizi yüzeyine tutturulan nükleotidlere *prob* adı verilir.

Mikrodiziler, DNA sekanslama, genetik hastalıkların teşhisi ve gen ifade ölçümlerinde kullanılır. Biyoçip ya da DNA çipleri olarak da adlandırılan mikrodiziler, gen ifade ölçümünde kullanılan en önemli teknolojilerden biridir. Gen ifadeleri sağlıklı ve hasta dokular arasındaki farklılıkların tespitinde sıklıkla kullanılır.

Gen çipindeki her bir spot yalnızca tek bir zincir DNA molekülünü içerir. Test edilecek mRNA örnekleri florasan boya ile işaretlenmiş olarak ya da işaretlenmeden, cDNA (komplementer DNA, mRNA'dan DNA'ya kopyalanmış) olarak kullanılmaktadır.

Gen ifade ölçümünde, tamamlayıcı DNA (cDNA) dizileri ve Oligonükleotid dizileri olmak üzere iki temel teknik vardır. cDNA dizileri daha ucuzdur, Oligonükleotid teknolojisi daha pahalı ve kapsamlıdır. Bu sebeple mikrodizi çipi ve tarayıcı üreten birkaç endüstriyel firma vardır. Affymetrix sistemler bunlardan biridir. cDNA dizilerinde, kontrol ve örnek aynı çip üzerine hibridize edilerek gerçekleştirilirken, oligonükleotid dizide kontrol ve örnek karşılaştırması için iki ayrı çipe ihtiyaç vardır.

cDNA dizilerde, 100 - 5 000 baz uzunluğundaki DNA zincirleri sentezlenerek cam veya plastik plak üzerine yerleştirilir. Problar, mürekkep püskürtmeli yazıcı gibi çalışır. DNA problarının sekansı DNA veritabanlarından elde edilir. Deneysel örneklerdeki RNA materyali izole edilerek DNA'ya çevrilir. Elde edilen cDNA zincirleri florasan boyalarla işaretlenerek DNA mikrodizisine sunulur. Çipe sabitlenmiş problar ile eşleşen cDNA molekülleri (hedef) hibridize olur yani tek zincir DNA tamamlayıcısı ile bağlanarak çift zincirli yapı oluşturur. Tamamlayıcısı olmayan hedef moleküller ise bağlanamazlar. Hibridizasyon olan spotlar optik olarak florasan görüntüleyicisi ile okunur. Florasan konumu ve şiddeti, hücrede hangi genlerin ne derecede ifade edildiğini gösterir. Şekil 4.5'te tipik bir cDNA mikrodizi deneyi görülmektedir.



Şekil 4.5. cDNA mikrodizi deneyi

cDNA deneyinde karşılaştırma işlemi için sağlıklı ve hasta dokudan mRNA izole edilerek, cDNA'ya kopyalanır. Her biri farklı florasan boyalarla (kırmızı, yeşil) etiketlenir. DNA çipine sokularak, hibridizasyon olan spotlar görüntülenir. Dört farklı sonuç elde edilebilir:

1. Kırmızı etiketli moleküller hibridize olabilir.
2. Yeşil etiketli moleküller hibridize olabilir.
3. Her iki molekül hibridize (Sarı renk) olabilir.
4. Hibridize olmayabilir (Renksiz).

Her bir probdan gelen renk yoğunluğu, hücredeki genin sağlıklı ve hasta dokuda ne derece ifade olduğunu gösterir. Mikrodizi teknolojisi tek bir çip ile binlerce genin ifadesini aynı anda ölçebilmekte, böylece hastalıkla ilişkili genleri tespit edilebilmektedir.

Tek nükleotid polimorfizmler (SNP) insan genomunda tanımlanmış bir noktada tek bir nükleotid farklılığı gösterirler (örnek AGGC, AGGT ye değişir). Her 1 000 baz çiftinde bir görülen SNP'ler hastalık genlerini tanımlamak için marker olarak kullanılabilir [75].

5. MEME KANSERİNİN SINIFLANDIRILMASINDA ETKİN GENLERİN TESPİTİ

Bu tez çalışmasında meme kanserinin sınıflandırılmasında etkin olan genlerin tespiti için literatürdeki mevcut metotlardan farklı bir metot önerilmiştir. Bunun için meme kanseri gen ifade verilerinden faydalanılmıştır. Meme kanseri gen ifade verileri çok az örnekten (hastadan) oluşmakta buna karşın on binlerce nitelik (gen) bilgisini içermektedir. Ancak bu genlerin pek çoğu ilgisiz ya da gürültü olarak adlandırabileceğimiz niteliklerdir. İlgisiz niteliklerin atılması, diğer taraftan özniteliklerin bulunması bu çalışmanın amacını oluşturmaktadır.

5.1. Meme Kanseri Gen İfade Veri Kümesi

Bu çalışmada meme kanserinde rol alan etkin genlerin belirlenmesi amacıyla, DNA mikrodizi tekniği ile elde edilmiş 97 meme kanseri hastasına ait gen ifade verileri kullanılmıştır. Bu veri kümesinde yer alan 78 hastanın 44'ünde 5 yıl içinde uzak metastaz görülmezken (iyi prognoz), 34'ünde 5 yıl içinde uzak metastaz görülmüştür (kötü prognoz). Ayrıca 19 lenf nodu negatif meme kanseri hastasının 7 sinde 5 yıl içinde uzak metastaz görülmezken, 12 sinde 5 yıl içinde uzak metastaz görüldüğü rapor edilmiştir. Mikrodizi veri kümesi Rosetta Inpharmatics'den alınmıştır [76].

Meme kanseri hastalarına ait gen ifade verileri şu niteliklere sahiptir:

Systematic name : Her bir gen veya sekans için verilmiş olan sistematik isim.

Gene name : Araştırmacılar tarafından atanmış gen isimleri.

Log10(Intensity) : Çip üzerinde kırmızı ve yeşil kanallar için geometrik ortalama yoğunluğu. Yüksek değer elde edilmesi, yüksek sinyal alınması anlamına gelir.

Log10(ratio) : Kırmızı yeşil kanalların ortalama yoğunluk oranı.

P-value : Güven düzeyi.

Veri kümesinde gen ifade verilerinin yanı sıra hastalara ait aşağıda gösterilen klinik bulgular da yer almaktadır.

SampleID : Hasta kimlik bilgisi.

Posnodes : Patoloji raporu sonucunda lenf nodu durumu.

EVENTmeta : Uzak metastaz durumu, 0 = hayır, 1 = evet.

EVENTdeath : Hastanın hayatta kalma durumu, 0 = hayır, 1 = evet.

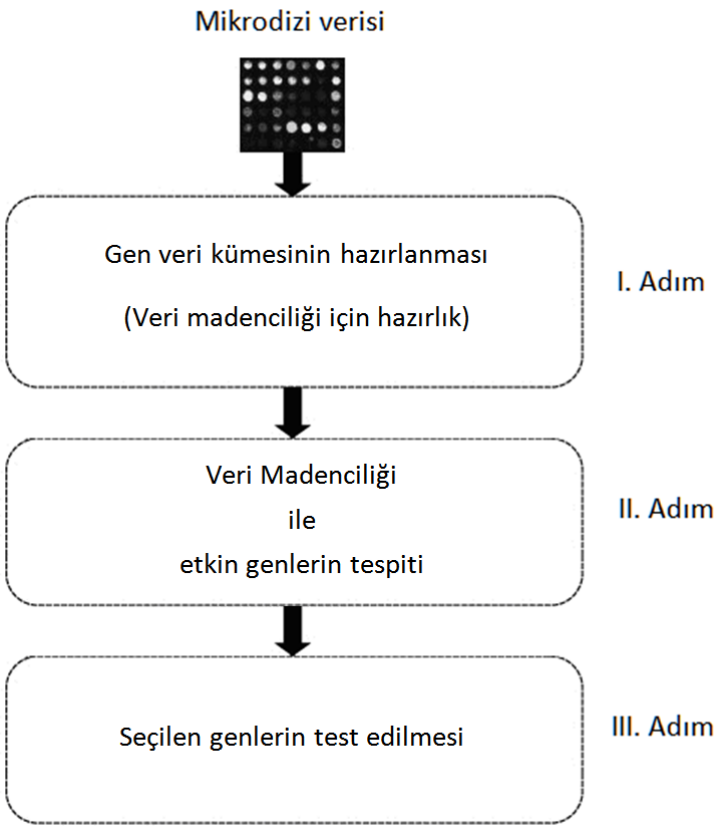
Meme kanserinde etkin genlerin tespiti için kullanılan veri kümesine, hastalara ait klinik bulgular dahil edilmemiştir. Bu amaçla sadece gen ifade verileri kullanılmıştır. Şekil 5.1'de meme kanseri hastalarına ait gen ifade verileri görülmektedir.

	A	B	C	D	E	F	G	H	I
1		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Samp
2	Accession	Log10(ratio)	Log10(ratio)	Log10(ratio)	Log10(ratio)	Log10(ratio)	Log10(ratio)	Log10(ratio)	Log10(
3	J00129	-0,4480	-0,4800	-0,5680	-0,8190	-0,1120	-0,3910	-0,6240	-0,52
4	Contig29982_RC	-0,2960	-0,5120	-0,4110	-0,2670	-0,6700	-0,3100	-0,1200	-0,44
5	Contig42854	-0,1000	-0,0310	-0,3980	0,0230	0,4210	-0,0600	-0,2360	-0,25
6	Contig42014_RC	-0,1770	-0,0750	0,1160	-0,2300	-0,1900	-0,1640	-0,1750	0,01
7	Contig27915_RC	-0,1070	-0,1040	-0,0920	0,1980	0,0320	-0,1730	0,2530	0,65
8	Contig20156_RC	-0,1100	-0,2340	-0,1660	-0,5100	0,2810	-0,0340	-0,1250	0,36
9	Contig50634_RC	-0,0950	-0,2250	0,0360	0,5290	0,3100	-0,0910	-0,1270	0,06
10	Contig42615_RC	-0,0760	-0,0940	0,3970	0,3540	0,0560	0,0360	-0,0200	0,18
11	Contig56678_RC	-0,1340	0,1150	-0,1940	-0,2610	0,1160	0,3460	0,0470	-1,14
12	Contig48659_RC	-0,1400	0,0190	-0,1280	0,0120	0,0740	0,0070	-0,1500	-0,11
13	Contig49388_RC	0,0060	0,1500	0,1390	-0,2600	0,0410	0,2510	0,2660	-0,15
14	Contig1970_RC	0,1110	0,0380	-0,0330	-0,0690	0,0670	0,2290	0,2460	-0,41
15	Contig26343_RC	-0,2360	0,0920	0,0390	-0,1150	0,2790	0,2970	0,1420	0,11
16	Contig53047_RC	-0,8660	-1,0350	-1,1140	-1,0210	-1,0060	-1,0590	-0,6950	-1,15
17	Contig43945_RC	0,1260	-0,0620	0,0110	-0,9990	0,2110	-0,1000	-0,1940	-0,05
18	Contig19551	-0,6920	-0,2100	-0,4620	0,2730	0,2420	-0,8830	0,2060	0,17
19	Contig10437_RC	0,1320	-0,1390	-0,1850	0,1590	0,2760	-0,1460	-0,3010	-0,07
20	Contig47230_RC	0,0950	0,0680	-0,1680	-0,3980	-0,6040	0,3820	-0,5490	-0,63
21	Contig20749_RC	0,2520	0,2680	-0,2890	-0,7340	0,0800	0,4030	-0,0120	-0,58
22	AL157502	0,1390	-0,1790	-0,3780	-0,4270	0,3720	-0,0140	-0,0220	-0,82
23	Contig36647_RC	-0,0970	0,1810	-0,4940	0,8480	-0,0100	0,6000	-0,9840	0,07
24	D31887	0,1130	0,0600	-0,2110	-0,3380	0,0760	-0,0250	0,0750	-0,03
25	AB033006	-0,2090	-0,1980	-0,3310	-0,2390	-0,1180	-0,3170	-0,2500	-0,08
26	AB033007	0,1070	0,0400	0,1140	0,0810	0,0720	0,1340	0,1310	0,06

Şekil 5.1. Meme kanseri hastalarına ait gen ifade verileri

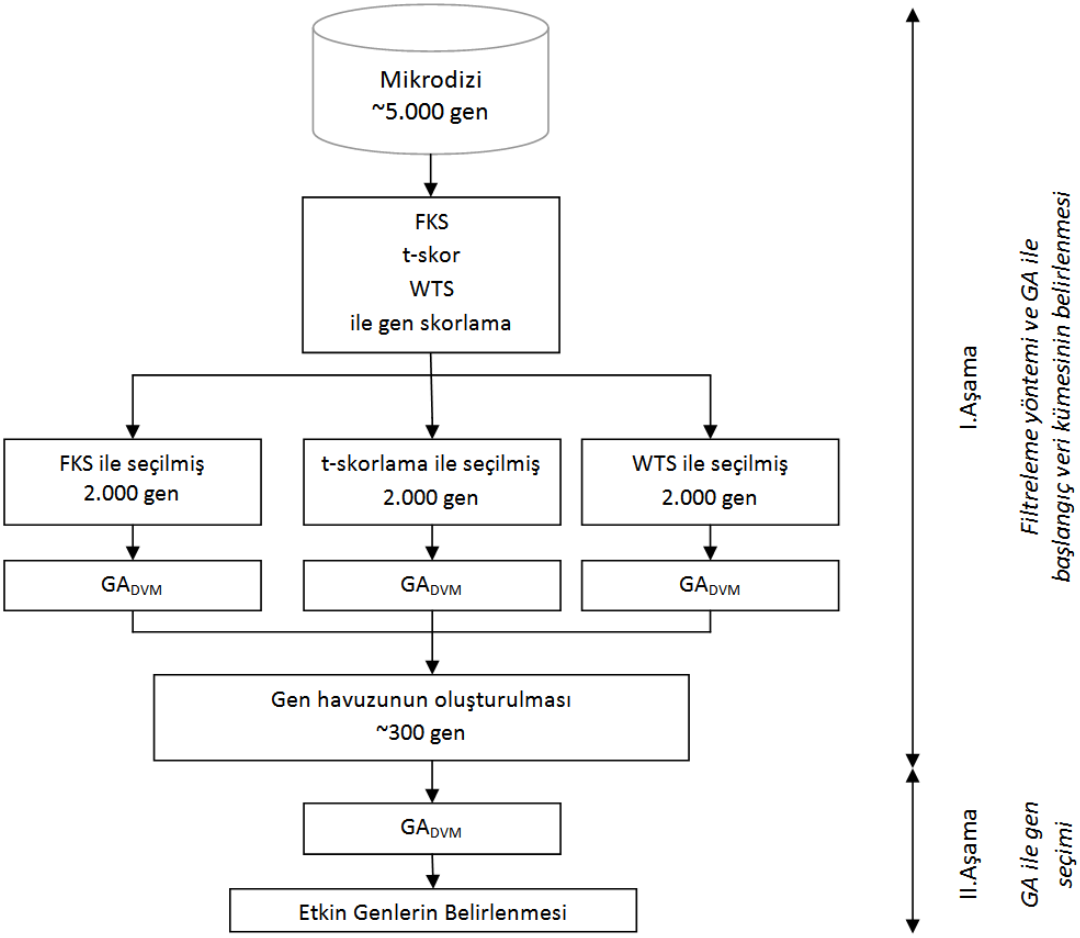
5.2. Önerilen Metodun Yapısı

Önerilen metot Şekil 5.2'de görüldüğü gibi üç adımda gerçekleştirilmektedir.



Şekil 5.2. Önerilen metodun yapısı

Birinci adımda, hatalı ve boş kayıtlar algoritma performansını olumsuz etkilediği için, gen ifade verisinde yer alan bu kayıtlar düzeltilmiştir. İkinci adımda, veri madenciliği teknikleri ile gen seçimi gerçekleştirilmiştir. İkinci adım Şekil 5.3'te görüldüğü gibi kendi içinde iki aşamadan oluşmaktadır. İlk aşamada, başlangıç veri kümesi belirlenmiş, ikinci aşamada genetik algoritma ve destek vektör makinesi ile gen seçimi gerçekleştirilmiştir. Genetik algoritma için destek vektör makinesi, uygunluk fonksiyonu olarak kullanılmıştır. Destek vektör makinesi için radyal tabanlı çekirdek fonksiyon tercih edilmiştir. Çalışmanın üçüncü adımında, belirlenen genlerin sınıflandırma başarısı ile geçerlik ve güvenirlik testleri yapılmıştır. Önerilen metot, etkin genleri belirlemesindeki başarısı ve farklı veri kümelerinde de uygulanabilir olması sebebiyle literatürdeki metotlardan daha üstündür.



Şekil 5.3. Mikrodizi verisinden gen seçimi

5.3. Veri Kümesinin Hazırlanması

Veri kümesinde yer alan hatalı ve boş kayıtlar, algoritmanın başarısını doğrudan etkilemektedir. Bu sebeple gen seçimi gerçekleştirilmeden önce gen ifade veri kümesi içinde var olan hatalı ve boş kayıtlar düzeltilmiştir. Literatürde hatalı kayıtların düzeltilmesinde iki yöntem sıkça kullanılmaktadır. İlki hatalı verinin silinmesi, ikincisi boş veya hatalı veriler yerine en yakın tahmine dayalı verilerin yerleştirilmesidir.

Mevcut gen ifade verisi 97 meme kanseri hastasına ait kaydı içermektedir. Bu veri kümesinde yer alan 78 hastanın 44'ünde iyi prognoz, 34'ünde ise kötü prognoz gözlemlenirken, 19 lenf nodu negatif meme kanseri hastasının 7'sinde iyi prognoz,

12'sinde kötü prognoz gözlemlenmiştir. Ayrıca veri kümesinde her bir hastaya ait yaklaşık 25 000 gen bilgisi yer almaktadır. Gen ifade verisi incelendiğinde NaN olarak adlandırılmış hatalı kayıtlar olduğu tespit edilmiştir. Bu kayıtlar ölçüm alınamayan veya hatalı ölçüm alınmış gen ifade verileridir. Veri kümesindeki hatalı veriler temizlenmeden başarılı bir şekilde etkin genlerin belirlenmesi mümkün değildir. Boş veya geçersiz kayıtlar sınıflandırma başarısını da etkilemektedir.

Veri kümesinin az örnek içermesi sebebiyle hatalı kayıtlar silinmemiş, olabilecek değerler tahmin edilerek yeniden düzenlenmiştir. Hatalı kayıtlar, benzer davranış gösteren diğer örneklerden Öklid mesafe ölçümü ile tahmin edilerek elde edilmiştir. Böylece hatalı kayıtlar diğer bir deyişle NaN ile ifade edilen kayıtlar yeniden düzenlenmiştir. Bu sayede az örnekleme sahip gen ifade veri kümesinde örnek sayısı sabit tutulmuştur.

Hatalı kayıtların düzeltilmesinden sonra gen ifade veri kümesinde istatistiksel olarak anlamlı olmayan nitelikler elenmiştir. Gen ifade verilerinden en az beş örnekte iki kat daha fazla ifade olmuş ve p-değeri 0,001'in altında kalan anlamlı yaklaşık 5 000 gen bilgisi seçilmiştir.

5.4. Veri Madenciliği ile Gen Seçimi

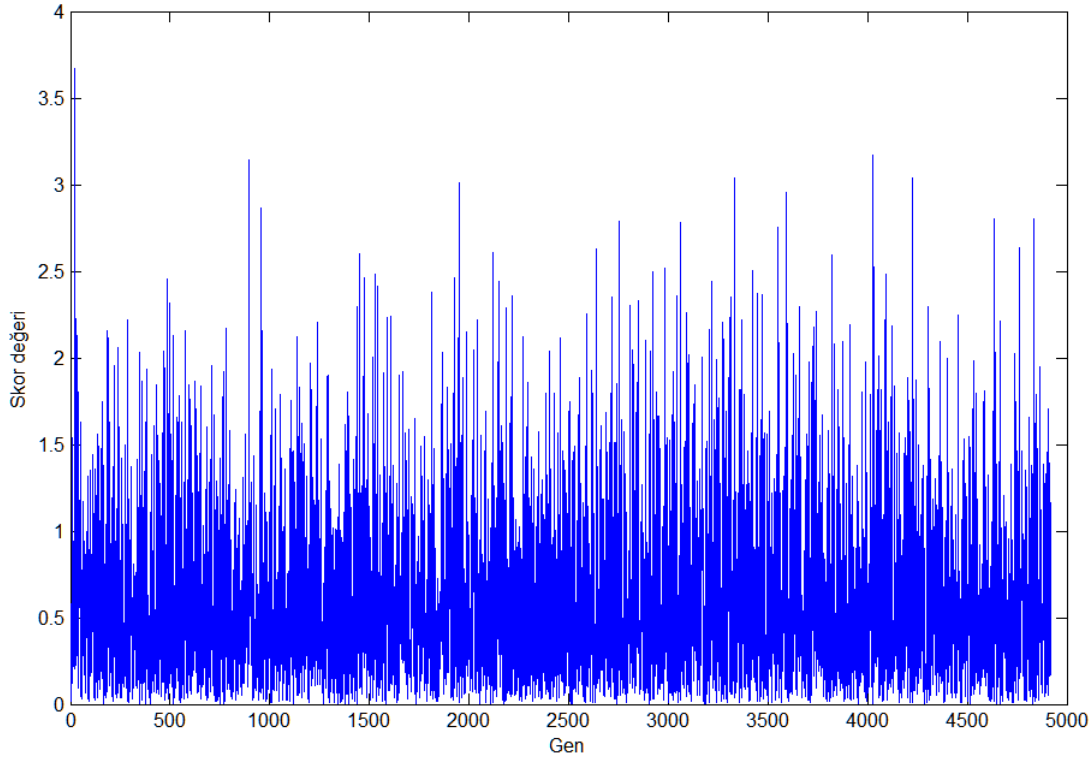
Veri madenciliği ile gen seçimi adımı kendi içinde iki aşamadan oluşmaktadır: İlk aşama genetik algoritma için başlangıç veri kümesinin belirlenmesi aşamasıdır. Gen ifade veri kümesinin yüksek boyutu hesap karmaşıklığını arttırmakta ve uygun gen alt kümesini bulma başarısını düşürmektedir. Arama uzayını daraltmak ve böylece genetik algoritmanın performansını artırmak için Şekil 5.3'te görüldüğü gibi I. aşamada filtreleme yöntemleri kullanılarak veri kümesi boyutu azaltılmıştır. Böylece 2 000 gen bilgisi içeren yeni veri kümesi elde edilmiştir. Her bir yöntemin arkasından genetik algoritma ile ayrı ayrı en yüksek başarıyı gösteren genler belirlenerek yaklaşık 300 gen bilgisine sahip yeni veri kümesi oluşturulmuştur. Bu veri kümesi içinden nihai gen seçimi gerçekleştirilmiştir.

Veri kümesinin başlangıç veri boyutunun indirgenmesinde Fisher Korelasyon Skorlama, t-Skor ve WTS yöntemleri kullanılmıştır. Her bir skorlama yöntemine göre en yüksek skora sahip ilk 2 000 gen belirlenmiştir. Daha sonra her bir yöntemin arkasından ayrı ayrı genetik algoritma ile en yüksek sınıflandırma başarısı gösteren genler belirlenerek yaklaşık 300 gen bilgisi içeren başlangıç veri kümesi (Gen havuzu) oluşturulmuştur. Başlangıç veri kümesi Eş. 5.1 ile ifade edilebilir.

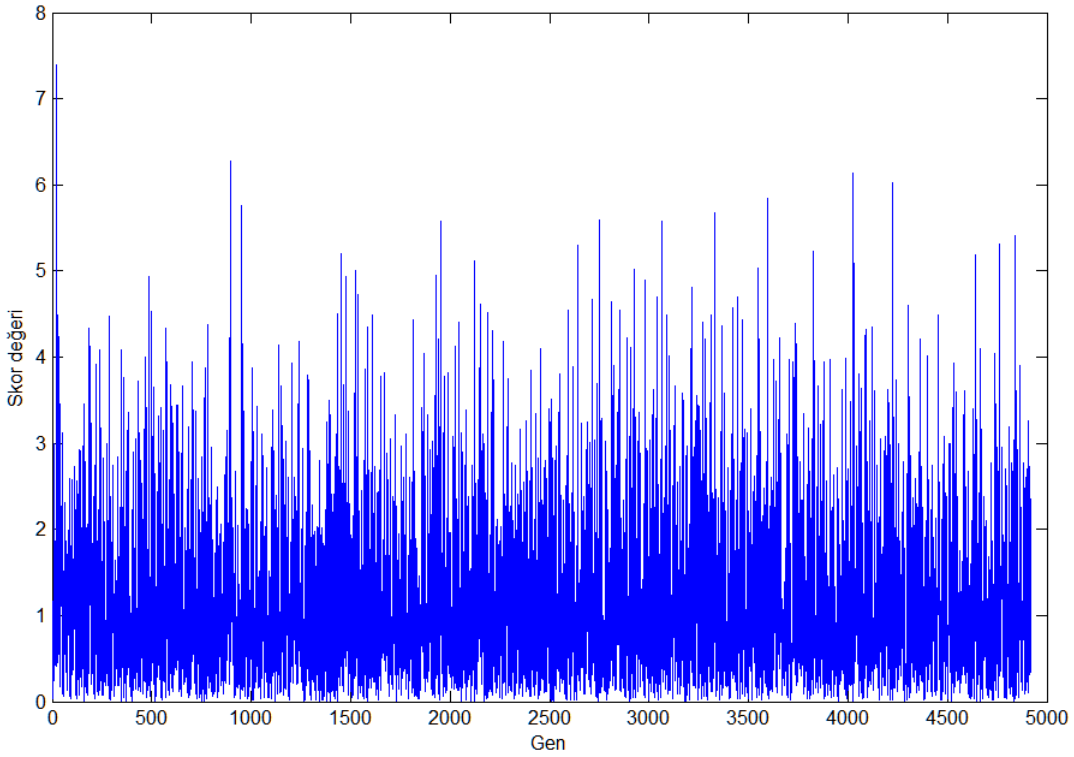
$$\text{Gen Havuzu}_{\sim 300\text{gen}} = \{\text{FKS}_{2000}\}_{\text{GA}} \cup \{\text{t-Skor}_{2000}\}_{\text{GA}} \cup \{\text{WTS}_{2000}\}_{\text{GA}} \quad (5.1)$$

5.4.1. Başlangıç veri kümesinin belirlenmesi

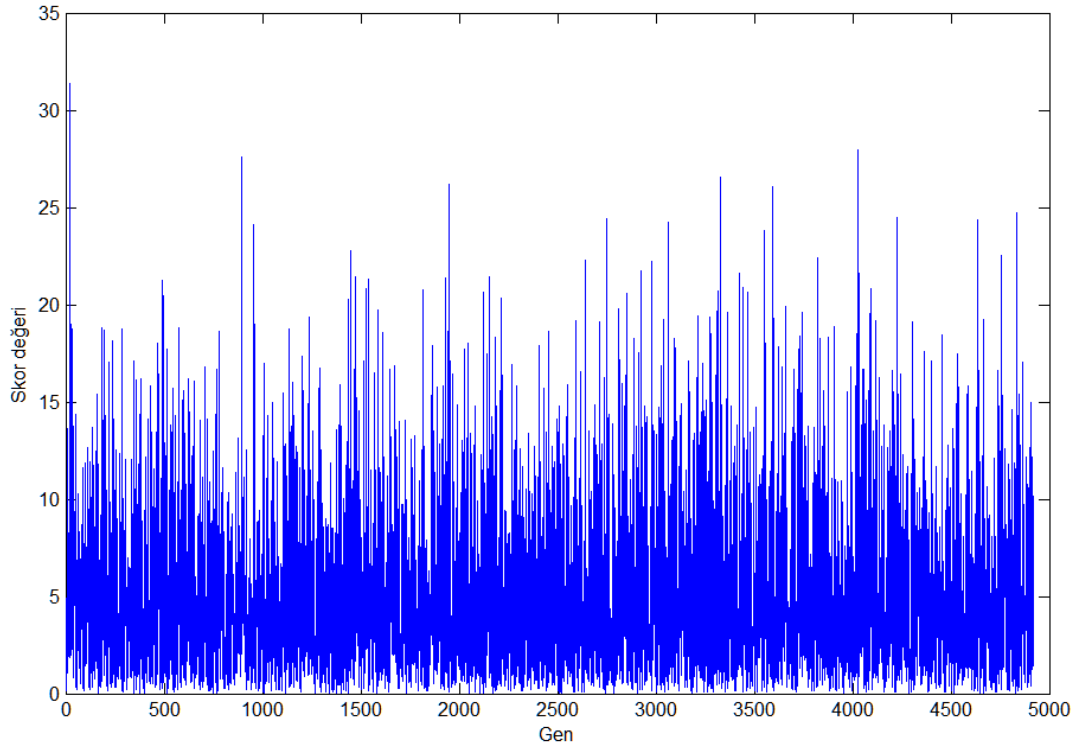
Gen ifade veri kümesinin boyutu Bölüm 3'te bahsedildiği gibi FKS, t-skorlama ve WTS skorlama yöntemleri ile Eş. 3.30, Eş. 3.31 ve Eş. 3.32 kullanılarak indirgenmiştir. Her üç skorlama yöntemi sonucunda ayrı ayrı en yüksek skora sahip ilk 2 000 gen belirlenmiştir. Şekil 5.4'te Fisher Korelasyon Skorlama, Şekil 5.5'te t-Skorlama Şekil 5.6'da WTS skor grafikleri görülmektedir.



Şekil 5.4. Fisher korelasyona skorlama sonuçları



Şekil 5.5. t-skor sonuçları



Şekil 5.6. WTS sonuçları

Skorların hesaplanması sonucu FKS ile 0,572 – 3,669 skor değerleri aralığındaki, t-skor ile 1,13 - 7,399 skor değerleri aralığındaki ve WTS ile 4,849 - 31,4 skor değerleri aralığındaki ilk 2 000 gen seçilmiştir. Çizelge 5.1'de 5 000 ve ilk 2 000 gene ait skor ortalama ve standart sapma değerleri görülmektedir.

Çizelge 5.1. FKS, t-Skor ve WTS ile elde edilen skor ortalama ve standart sapma değerleri

		5 000 gen	İlk 2 000 gen
FKS	μ	0,5906	1,0773
	σ	0,5088	0,4470
t-Skor	μ	1,1584	2,1144
	σ	0,9974	0,8723
WTS	μ	5,0168	9,2007
	σ	4,3780	3,8672

FKS, t-Skor ve WTS ile ayrı ayrı her bir niteliğin skoru hesaplanmıştır. Daha sonra en yüksek skora sahip 2 000 gen seçilmiştir. Elde edilen grafikler birbirine çok benzer görünmesine rağmen detayda farklılıklar olduğu gözlemlenmiştir. Böylece FKS, t-Skor ve WTS'nin tek başına ayırt edemediği nitelikler de başlangıç veri kümesine dahil edilmiştir.

Eş. 5.1'de ifade edildiği gibi filtreleme yöntemlerinden sonra ayrı ayrı genetik algoritma ile en yüksek sınıflandırma doğruluğu gösteren ortak genler belirlenmiştir. Böylece yaklaşık 300 gen bilgisi içeren başlangıç veri kümesi elde edilmiştir. Filtreleme ve genetik algoritma ile belirlenen başlangıç veri kümesini oluşturan genlerin skor ortalama ve standart sapma değerleri Çizelge 5.2'de görüldüğü gibi elde edilmiştir.

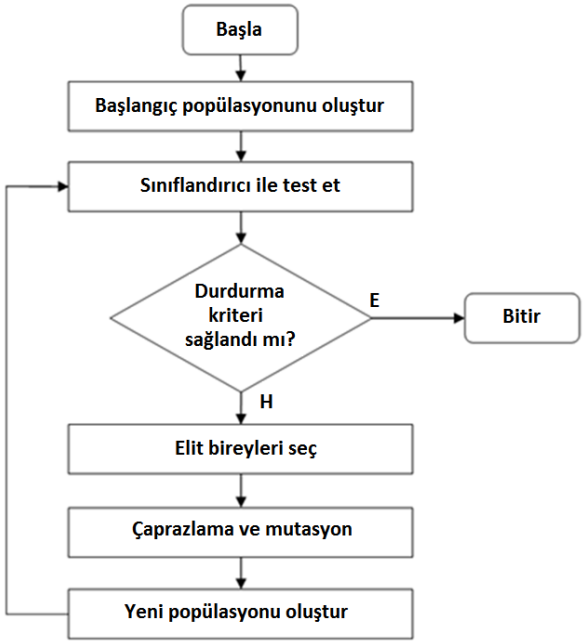
Çizelge 5.2. Başlangıç veri kümesinde yer alan yaklaşık 300 gene ait skor ortalama ve standart sapma değerleri.

	FKS	t-Skor	WTS
μ	1,2754	2,4585	10,8080
σ	0,6164	1,2045	5,3923

Elde edilen sonuçlar Çizelge 5.1 ile karşılaştırıldığında skor ortalama ve standart sapma değerlerinin yükseldiği görülmüştür. Bu durum yüksek skorlu niteliklerin yanı sıra nispeten düşük skorlu niteliklerin de başlangıç veri kümesine dahil edilmesinden kaynaklanmaktadır.

5.4.2. Genetik algoritma ile etkin genlerin belirlenmesi

Bölüm 3'te bahsedildiği gibi Genetik algoritma, en iyiyi arama aracıdır. Bizim için en iyi, sınıflandırma başarısı en yüksek olan genlerdir. I. aşamada belirlenen yeni gen veri kümesi, meme kanseri hastalığında etkin genlerin bulunmasında genetik algoritma için arama uzayı olacaktır. Şekil 5.7'de kullanılan genetik algoritmanın akış diyagramı görülmektedir.



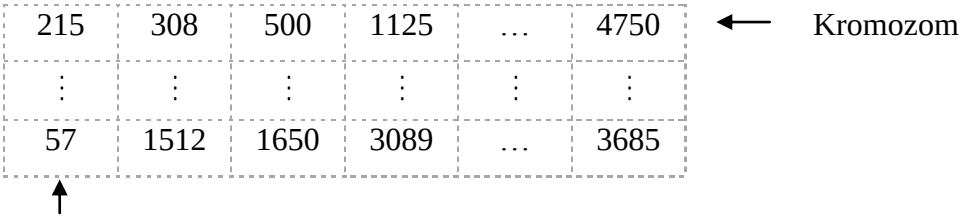
Şekil 5.7. Genetik algoritma akış diyagramı

Genetik algoritmanın kodlanması

Genetik algoritmanın dört önemli bileşeni vardır: genetik kodlama, başlangıç popülasyonu, uygunluk fonksiyonu, genetik operatörler (seçme, çaprazlama ve mutasyon). Bunlar aşağıda gösterildiği gibi kodlanmıştır.

Genetik kodlama

Genetik algoritmada sıklıkla tercih edilen ikili kodlama, gen seçimi için uygun görülmüştür. Bir (1) olarak kodlanmış gen uygunluk fonksiyonuna gönderilirken, sıfır (0) olarak kodlanmış gen uygunluk fonksiyonunda hesaba katılmayacak anlamına gelmektedir. Şekil 5.8'de kromozom yapısı görülmektedir.



Şekil 5.8. Kromozom yapısı

Başlangıç popülasyonu

Başlangıç popülasyonu, rasgele gen değerlerine sahip kromozomlardan meydana gelmektedir. Kromozom sayısı 100 olarak belirlenmiştir. Ancak kromozom sayısının (popülasyonun) artması çalışma zamanını olumsuz etkilemektedir.

Uygunluk fonksiyonu

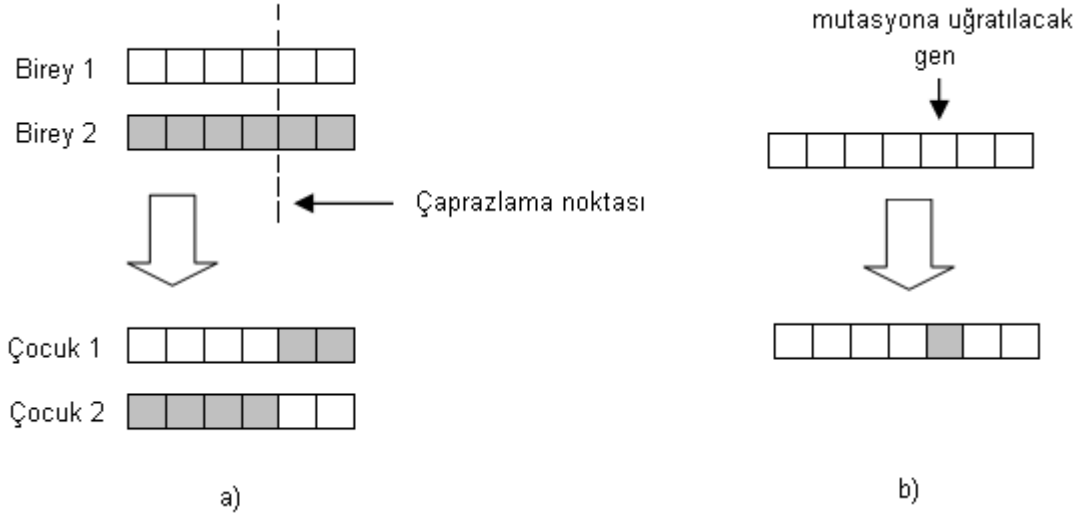
Meme kanserinde etkin genlerin seçiminde, seçilen genlerin sınıflandırma başarı oranı çok önemlidir. Genetik algoritmanın, sınıflandırmada en yüksek başarıyı gösteren gen alt kümesini seçmesi gerekmektedir. Etkin genlerin sınıflandırma başarısı da yüksek olacaktır. Destek vektör makinesi sınıflandırma doğruluk oranı, uygunluk fonksiyonu olarak belirlenmiştir. Destek vektör makinesi için radyal tabanlı çekirdek fonksiyon tercih edilmiştir.

Seçme

Rasgele belirlenen başlangıç popülasyonunda, her kromozomun sınıflandırma doğruluk oranı belirlenir. En yüksek başarı oranına sahip kromozomlar, elit kromozomlar olarak yeni popülasyona eklenir. Elit kromozomlar başlangıç popülasyonunu oluşturan bireylerin %10'u olarak kabul edilmiştir. Geri kalan bireyler rulet tekerleği ile belirlenmiştir.

Çaprazlama ve mutasyon

Eşleştirme işleminde tek noktalı çaprazlama metodu kullanılmıştır. Şekil 5.9'da görüldüğü gibi rasgele belirlenen iki kromozom için yine rasgele belirlenen iki noktadan kromozomun genleri çaprazlanır. Çaprazlama sonucunda yeni iki birey elde edilecektir. Geliştirilen uygulamada bireylerin mutasyon oranı sabit değildir. Başarı oranına göre her birey farklı oranda mutasyona tabii tutulmaktadır. Nispeten yüksek başarı gösteren birey küçük bir mutasyona uğratılırken, daha düşük başarı oranı gösteren bireyler yüksek mutasyona uğratılmıştır. Böylece genetik algoritmanın daha etkin çalışması sağlanmıştır.



Şekil 5.9. a) Çaprazlama b) Mutasyon işlemi

5.5. Deneysel Bulgular

FKS ile elde edilen skor sıralamasına göre ilk 100, ilk 50 ve ilk 20 genin sınıflandırma başarısı hesaplanmıştır. Destek vektör makinesi ile elde edilen sınıflandırma başarı oranları Çizelge 5.3'te görülmektedir.

Çizelge 5.3. FKS ile elde edilen genlerin sınıflandırma sonuçları

	İlk 100 gen	İlk 50 gen	İlk 20 gen
Sınıflandırma başarısı	% 67,9	% 58,97	% 70,51

t-skorlama ile elde edilen skor sıralamasına göre ilk 100, ilk 50 ve ilk 20 genin sınıflandırma başarısı hesaplanmıştır. Destek vektör makinesi ile elde edilen sınıflandırma başarı oranları Çizelge 5.4'te görülmektedir.

Çizelge 5.4. t-skorlaması ile elde edilen sınıflandırma sonuçları

	İlk 100 gen	İlk 50 gen	İlk 20 gen
Sınıflandırma başarısı	% 77,56	% 69,62	% 69,10

WTS ile elde edilen skor sıralamasına göre ilk 100, ilk 50 ve ilk 20 genin sınıflandırma başarısı hesaplanmıştır. Destek vektör makinesi ile elde edilen sınıflandırma başarı oranları Çizelge 5.5'te görülmektedir.

Çizelge 5.5. WTS ile elde edilen sınıflandırma sonuçları

	İlk 100 gen	İlk 50 gen	İlk 20 gen
Sınıflandırma başarısı	% 71,74	% 67,18	% 72,69

I. aşamada, filtreleme yöntemine göre yapılan nitelik indirgeme sonucunda her bir yöntemin sınıflandırma doğruluk oranlarının farklı farklı olduğu görülmüştür. En yüksek skorlu ilk 100 gene ait sınıflandırma doğruluk oranları en yüksek %77,56 ve ilk 50 genin sınıflandırma doğruluk oranları en yüksek %69,62 ile t-skor ile elde edilmiştir. İlk 20 genin sınıflandırma doğruluk oranı en yüksek %72,69 ile WTS ile elde edilmiştir.

Her bir filtreleme yöntemi sonrasında genetik algoritma ile en yüksek sınıflandırma başarısı gösteren genler seçilmiştir. Çizelge 5.6, Çizelge 5.7 Çizelge 5.8'de sırayla FKS, t-skor ve WTS sonrasında yüksek sınıflandırma başarısı gösteren gen alt kümelerinin sınıflandırma başarıları görülmektedir.

Çizelge 5.6. FKS sonrasında GA ile belirlenen gen alt kümelerinin sınıflandırma başarıları

Belirlenen gen sayısı	10 gen	11 gen	12 gen	13 gen	14 gen
Sınıflandırma başarısı	% 91,03	% 90,24	% 89,74	% 89,07	% 87,18

Çizelge 5.7. t-skor sonrasında GA ile belirlenen gen alt kümelerinin sınıflandırma başarıları

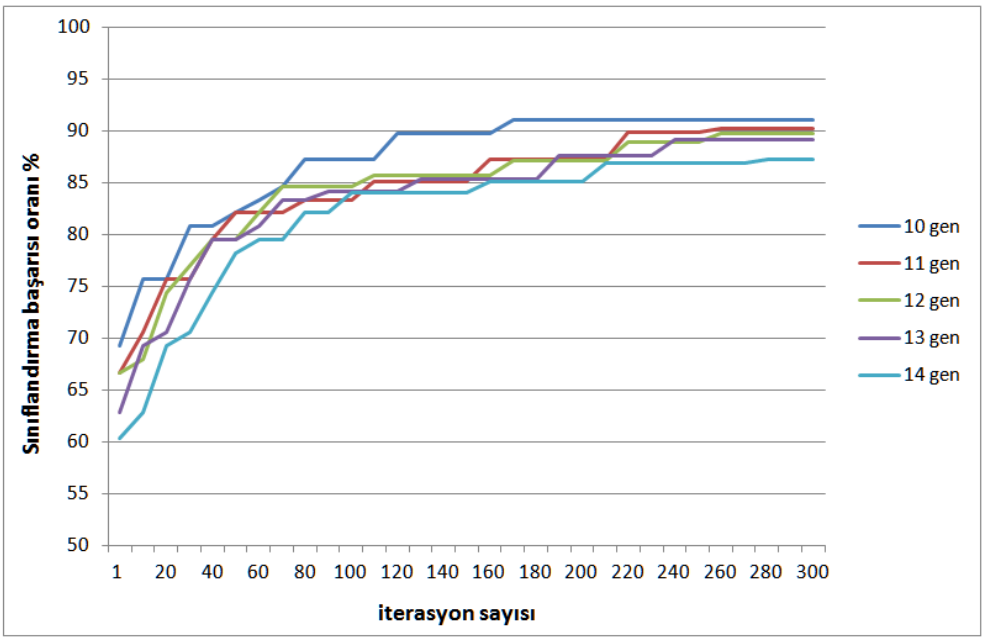
Belirlenen gen sayısı	9 gen	10 gen	11 gen	12 gen	13 gen
Sınıflandırma başarısı	% 89,91	% 87,97	% 89,74	% 88,07	% 86,77

Çizelge 5.8. WTS sonrasında GA ile belirlenen gen alt kümelerinin sınıflandırma başarıları

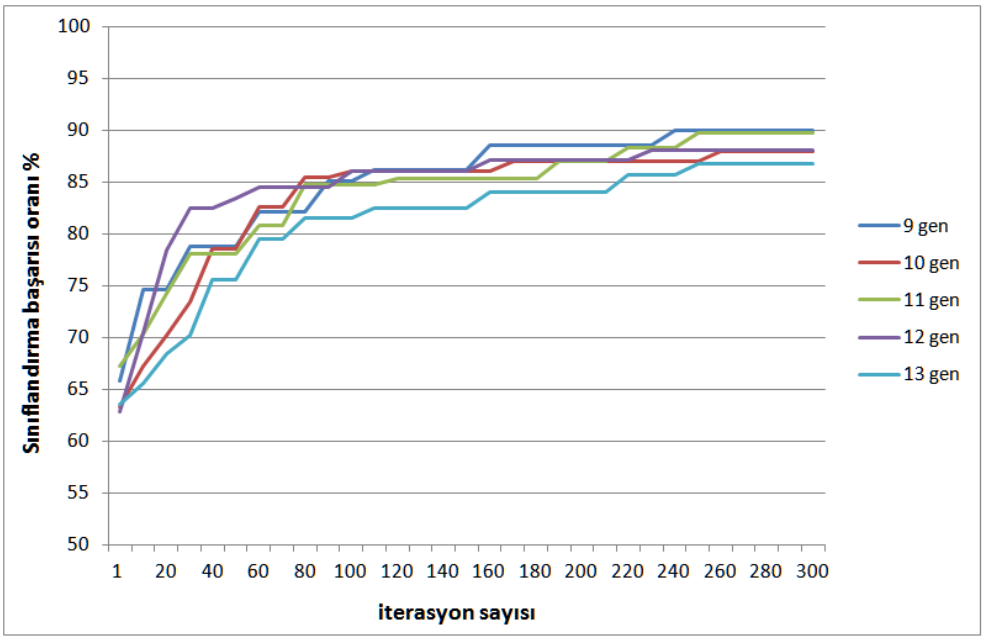
Belirlenen gen sayısı	10 gen	11 gen	12 gen	13 gen	14 gen
Sınıflandırma başarısı	% 90,46	% 88,46	% 87,66	% 86,65	% 84,75

FKS sonrasında GA ile belirlenen gen alt kümesi ile en yüksek sınıflandırma başarısı 10 gen ile %91.03 elde edilirken, t-skor sonrasında GA ile elde edilen 9 gen alt kümesi ile %89,91 sınıflandırma başarısı elde edilmiştir. WTS sonrasında GA ile elde edilen 10 gen alt kümesi ile %90,46 sınıflandırma başarısı elde edilmiştir.

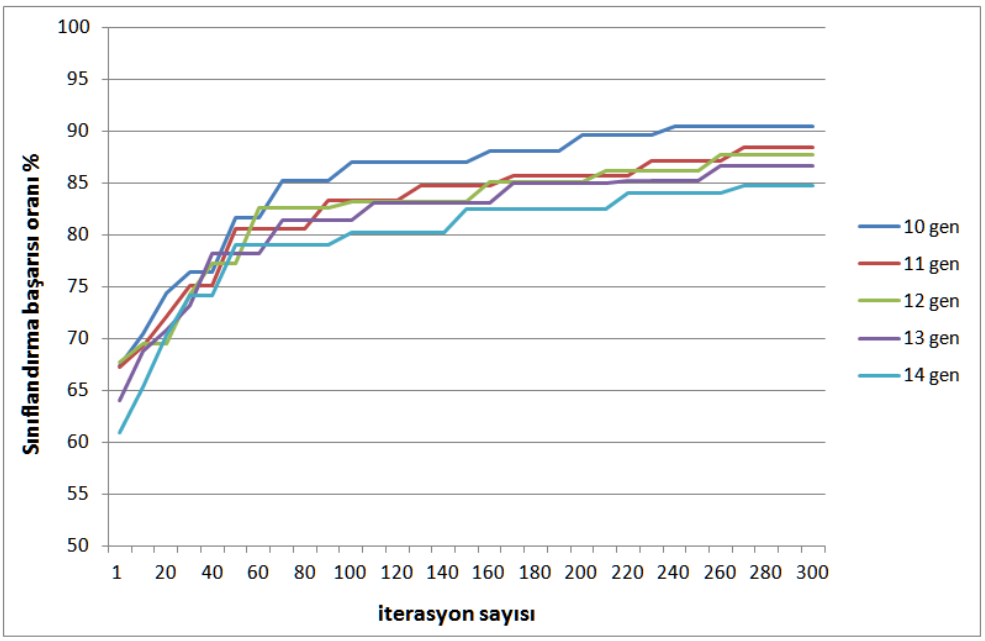
Şekil 5.10'da her bir filtreleme yönteminin arkasından gerçekleştirilen genetik algoritma ile belirlenen genlerin sınıflandırma başarısı-iterasyon eğrileri görülmektedir.



a) FKS sonrasında yüksek başarı gösteren genlerin sınıflandırma başarısı-iterasyon eğrileri



b) t-skor sonrasında yüksek başarı gösteren genlerin sınıflandırma başarısı-iterasyon eğrileri

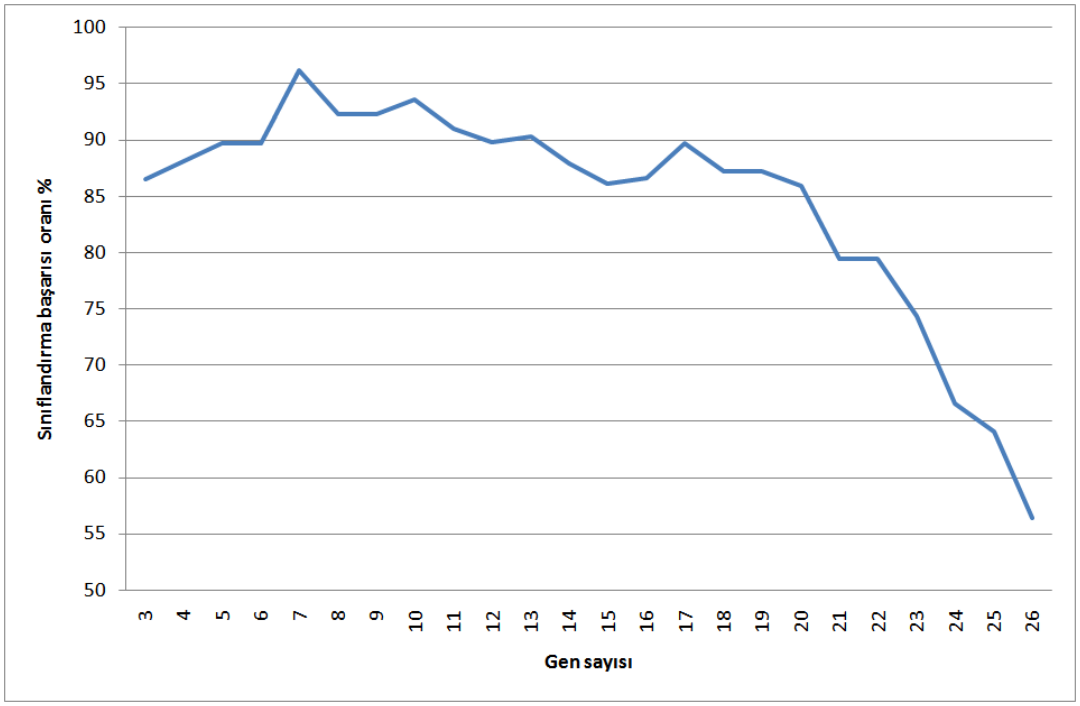


c) WTS sonrasında yüksek başarı gösteren genlerin sınıflandırma başarısı-iterasyon eğrileri

Şekil 5.10. Filtreleme yöntemleri sonrasında genetik algoritma ile belirlenen genlerin sınıflandırma başarısı-iterasyon eğrileri

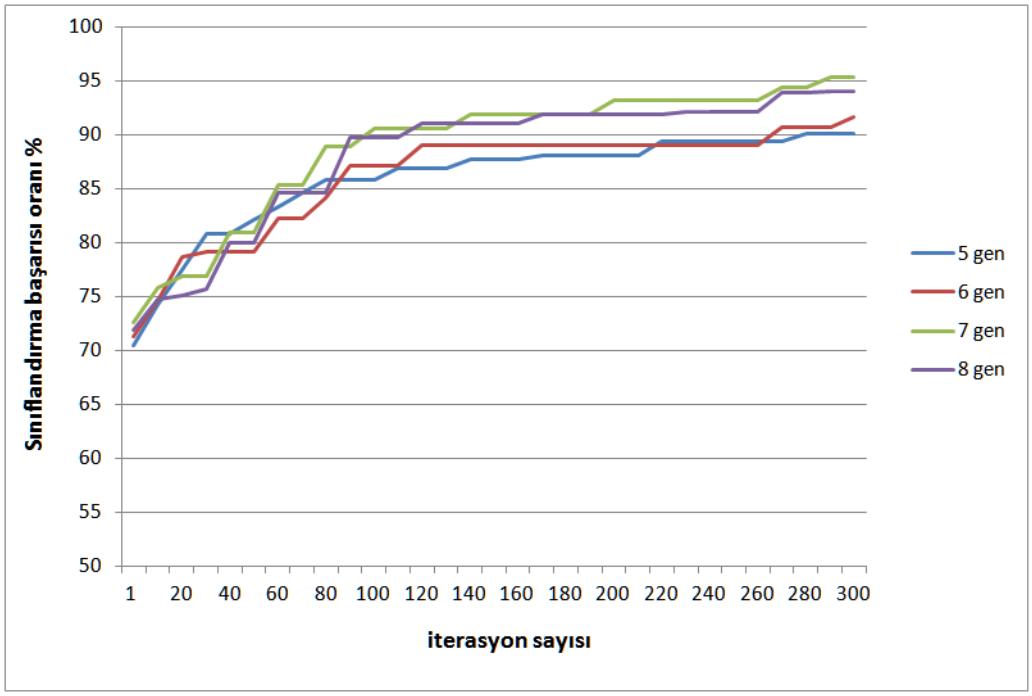
İlk aşamada başlangıç veri kümesi belirlendikten sonra nihai gen seçimini yapmak üzere II. aşamaya geçilmiştir. II. aşamada, ilk aşamada oluşturulan başlangıç veri kümesinden genetik algoritma ile gen seçimi gerçekleştirilmiştir. Başlangıç popülasyonu 100 ve iterasyon sayısı 300 belirlendiğinde en uygun çalışma performansı elde edilmiştir. 300 ve üstü iterasyonda çalışma zamanının yüksek olduğu ve ayrıca iterasyon sayısının daha fazla artmasının seçilen nitelik alt kümelerini değiştirmediği gözlemlenmiştir.

Şekil 5.11'de GA_{DVM} ile belirlenen genlere ait sınıflandırma doğruluk oranları görülmektedir. Tek nitelik bilgisi ile sınıflandırma doğruluğunun çok düşük olduğu görülürken 3 ve üstü nitelik bilgisine sahip alt kümelerin bazılarında sınıflandırma doğruluk oranının %80'nin üzerinde olduğu gözlemlenmiştir. Nitelik sayısı 20 ve üzerinde olduğunda sınıflandırma doğruluk oranının düştüğü gözlemlenmiştir. En yüksek sınıflandırma doğruluk oranı 7 nitelik (gen) ile tespit edilmiştir. %96,15 sınıflandırma başarısı ile belirlenen 7 genin en yüksek doğrulukta sınıflandırma yaptığı belirlenmiştir. Şekil 5.11'de yüksek sınıflandırma başarısı gösteren gen alt kümelerinin sınıflandırma başarısı ve seçilmiş gen sayısı eğrisi görülmektedir.

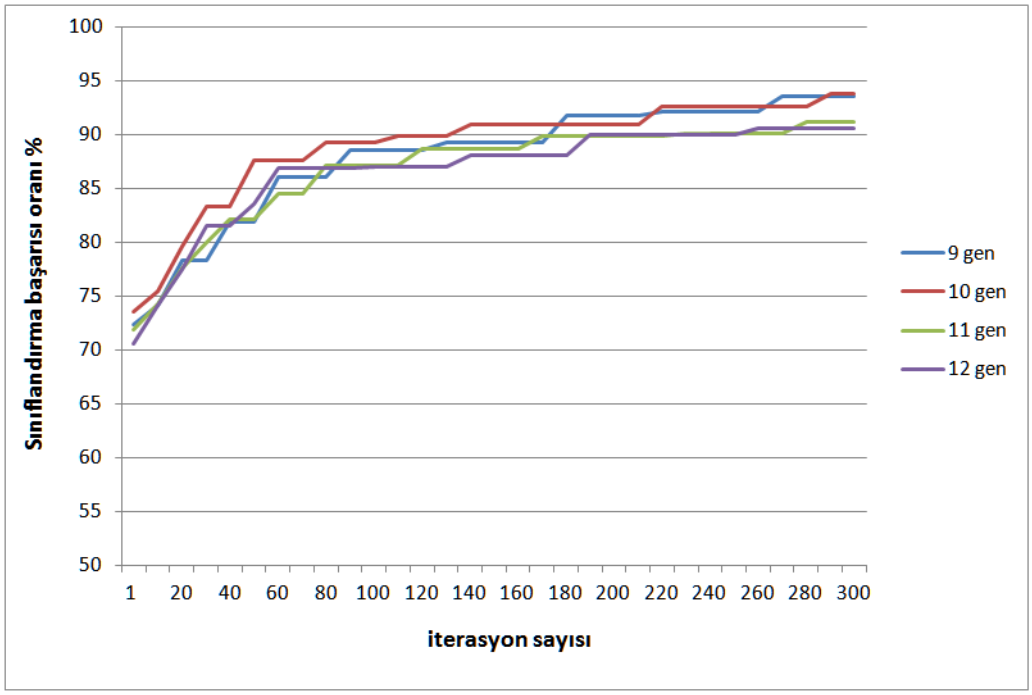


Şekil 5.11. Seçilen genler ve sınıflandırma başarısı oranları

II. aşamada, yaklaşık 300 gen bilgisinden oluşan veri kümesinde genetik algoritma ile belirlenen genlerin sınıflandırma başarısı - iterasyon eğrileri Şekil 5.12'de görüldüğü gibi elde edilmiştir. 7 gen alt kümesi ile %96,15 sınıflandırma başarısı elde edilirken, 22 ve daha yüksek gen alt kümesi ile sınıflandırma başarısı %80 ve daha aşağı elde edilmiştir.



a) En yüksek sınıflandırma başarısı gösteren 5 ile 8 gen alt kümesinin sınıflandırma başarısı-iterasyon eğrileri



b) En yüksek sınıflandırma başarısı gösteren 9 ile 12 gen alt kümesinin sınıflandırma başarısı-iterasyon eğrileri.

Şekil 5.12. Yüksek sınıflandırma başarısı gösteren gen alt kümelerine ait sınıflandırma başarısı - iterasyon eğrileri.

Şekil 5.12'de görüldüğü gibi, genetik algoritma ile gen alt kümesinin seçiminde başlangıçta sınıflandırma başarısı hızla yükselirken sonraki iterasyonlarda sınıflandırma başarısı sabit kalmıştır. 300 ve üstü iterasyonda sonucun çok değişmediği gözlemlenmiştir.

Şekil 5.11'de görüldüğü gibi en yüksek sınıflandırma başarısı 7 gen alt kümesi ile %96,15 olarak elde edilmiştir. Çizelge 5.9'da ise elde edilen en yüksek sınıflandırma başarısı oranları görülmektedir.

Çizelge 5.9. GA_{DVM} ile belirlenen gen alt kümeleri ve sınıflandırma başarıları

Belirlenen gen alt kümesi	Sınıflandırma başarısı
5 gen	% 89,74
6 gen	% 89,74
7 gen	**% 96,15**
8 gen	% 92,30
9 gen	% 92,30
10 gen	% 93,58
11 gen	% 91,02
12 gen	% 88,46

Çizelge 5.10'da GA_{DVM} ile belirlenen 7 gene ait FKS, t-skor ve WTS skor değerleri görülmektedir. Bu çizelgede verilen *Accession değeri*, her bir genin sahip olduğu özel kimlik numarasıdır.

Çizelge 5.10. Belirlenen 7 gene ait skor değerleri

Accession değeri	FKS	t-Skor	WTS
NM_000918	1,5607	2,9303	13,663
NM_003504	1,4326	2,6747	12,513
NM_006101	1,7343	3,4126	15,308
Contig173	1,3312	2,6774	11,56
NM_006783	1,907	3,3968	16,25
AL137718	1,1707	2,2094	8,7278
NM_002019	1,9834	3,9293	17,476

Çizelge 5.11'de belirlenen bu genlerin FKS, t-skor ve WTS ile elde edilen skor sıralamaları ve Çizelge 5.12'de ortalama ve standart sapma değerleri görülmektedir. Sonuçlar Çizelge 5.1 ve Çizelge 5.2 ile karşılaştırıldığında, ortalama ve standart sapma değerlerinin düştüğü gözlemlenmiştir. Filtreleme yöntemi ile belirlenen en yüksek skorlu niteliklerin, GA ile belirlenen niteliklerden farklı olduğu açıkça görülmektedir. Diğer bir ifadeyle GA ile belirlenen 7 gen, filtreleme yöntemlerinde en yüksek skora sahip nitelikler değildir. Ancak bu nitelikler ile en yüksek sınıflandırma başarısı elde edilmiştir.

Çizelge 5.11. Belirlenen 7 genin FKS, t-skor ve WTS hesaplamasına göre sıra değerleri

Accession değeri	FKS	t-Skor	WTS
NM_000918	271	314	246
NM_003504	370	438	349
NM_006101	182	175	166
Contig173	461	434	444
NM_006783	121	180	127
AL137718	644	719	886
NM_002019	102	97	89

Çizelge 5.12. Belirlenen 7 gene ait skor ortalama ve standart sapma değerleri

	FKS	t-Skor	WTS
μ	1,5886	3,0329	13,643
σ	0,3012	0,5807	3,001

Şekil 5.11'de görüldüğü gibi 7 gen ile en yüksek başarı oranı %96,15 elde edilmiştir. 20 gen ve üstünde sınıflandırma doğruluk oranının düştüğü gözlemlenmiş ve sonraki değerlerde de değişme görülmemiştir. Seçilen 7 gen için sınıflandırma başarısı 10-kat çapraz geçerleme ile test edilmiştir. Böylece daha güvenilir hata kestirimi gerçekleştirilmiştir. Çizelge 5.13'te belirlenen genler için eğri altında kalan alan (AUC) değeri ile Doğru Pozitif (DP), Yanlış Pozitif (YP), Doğru Negatif (DN) ve Yanlış Negatif (YN) değerleri verilmiştir.

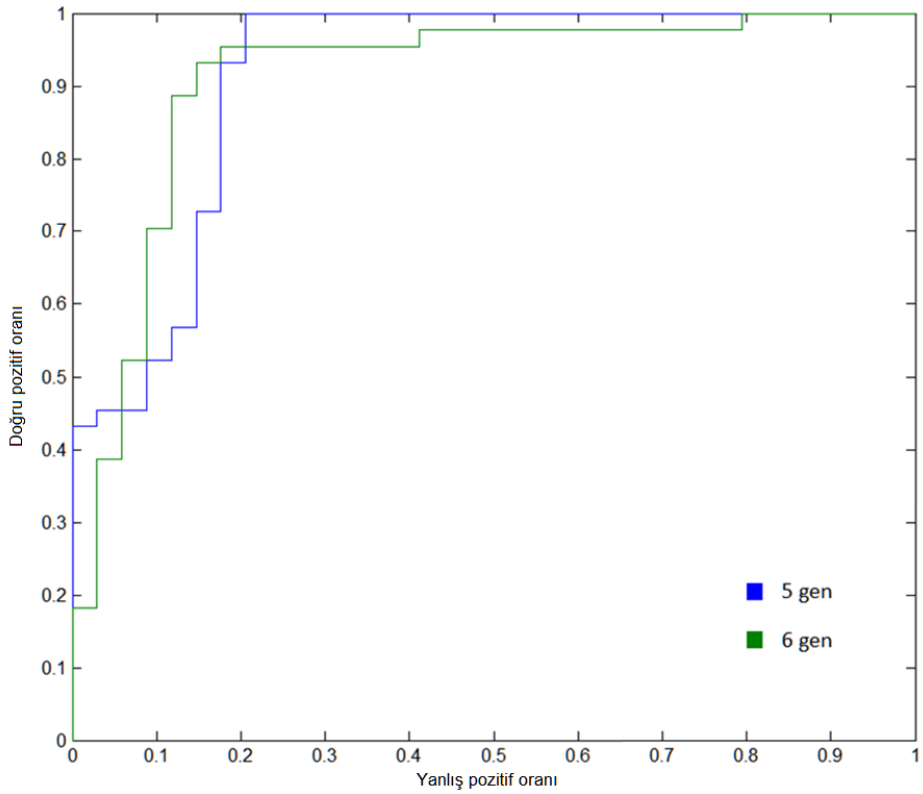
Çizelge 5.13. Belirlenen genler için AUC değeri ile Doğru Pozitif (DP), Yanlış Pozitif (YP), Doğru Negatif (DN) ve Yanlış Negatif (YN) değerleri

Seçilen genler	AUC	DP	YN	DN	YP
5 gen	0,9345	41	3	29	5
6 gen	0,9425	40	4	30	4
7 gen	**0,9606**	**42**	**2**	**33**	**1**
8 gen	0,9318	44	0	28	6
9 gen	0,9311	44	0	28	6
10 gen	0,9476	44	0	29	5
11 gen	0,9211	44	0	27	7
12 gen	0,9144	44	0	25	9

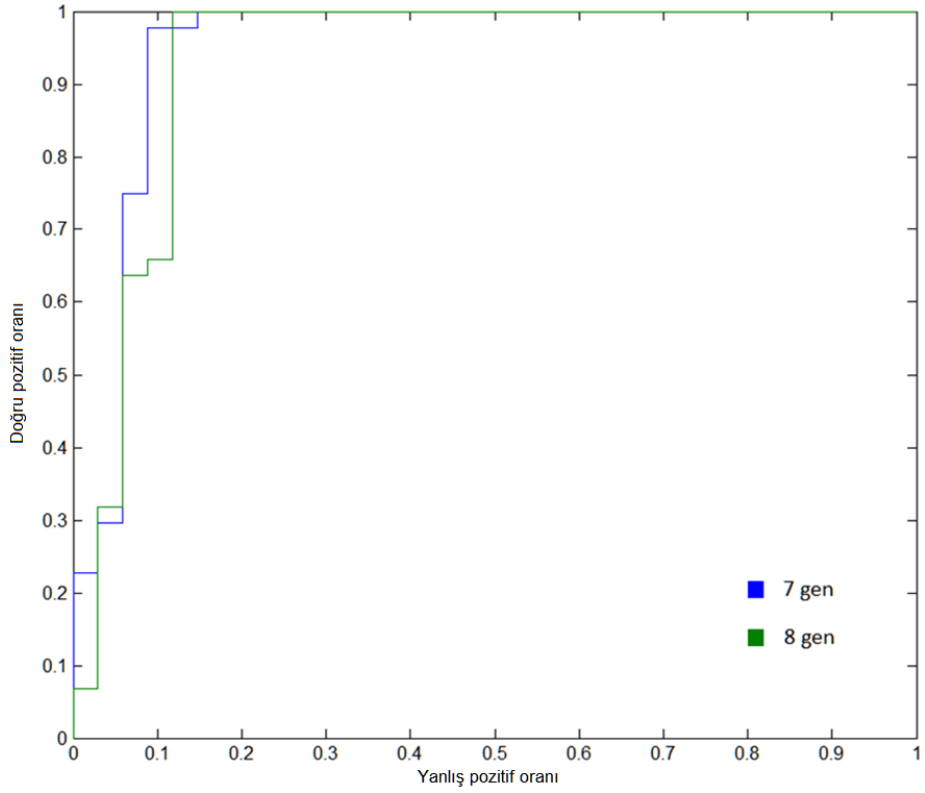
Alıcı İşletim Karakteristiği (Receiver Operating Characteristics - ROC) ve Eğri Altında kalan Alan (Area Under the Curve-AUC), sınıflandırıcı performansını test etmek için sıklıkla kullanılan birer yöntemdir.

Çizelge 5.13'te açıkça görüldüğü gibi belirlenen 7 gen alt kümesinin sınıflandırma performansı diğer gen alt kümelerine göre oldukça yüksek çıkmıştır. Özellikle negatif sınıf değeri yani 34 hasta içinde yalnızca 1 hata ile sınıflandırma gerçekleştirilmiştir. 44 pozitif örnek için yalnızca 2 hatalı sonuç elde edilmiştir.

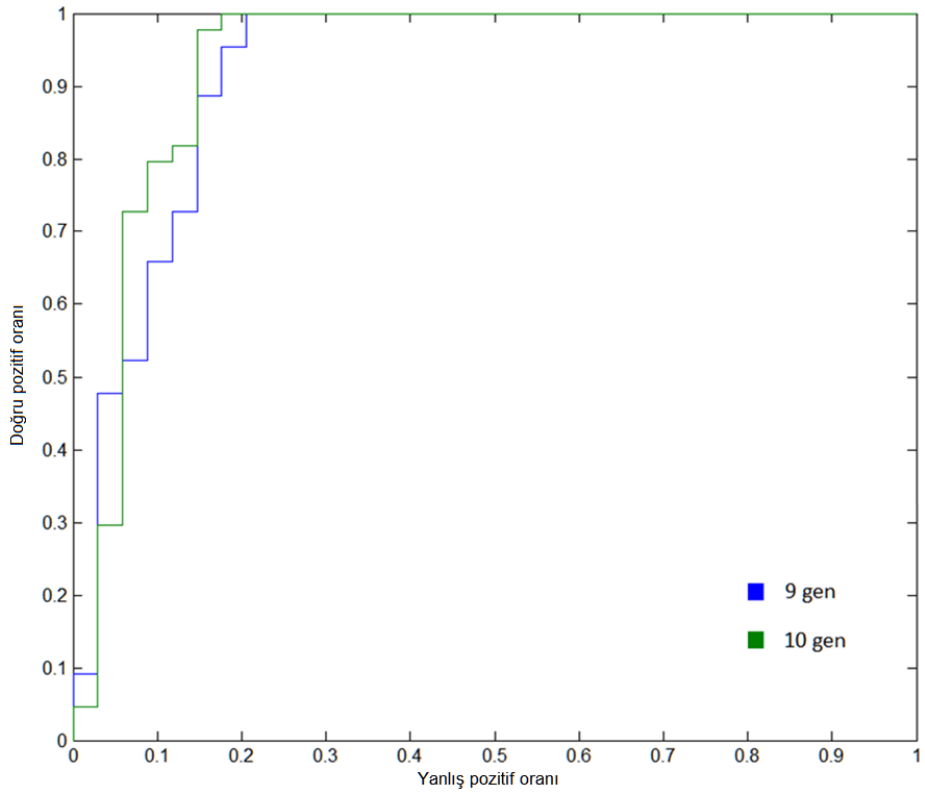
Belirlenen genlere ait ROC eğrileri Şekil 5.14, Şekil 5.15, Şekil 5.16 ve Şekil 5.17'de görüldüğü gibi elde edilmiştir. Elde edilen eğriler incelendiğinde 7 gen alt kümesi için AUC değeri oldukça yüksektir.



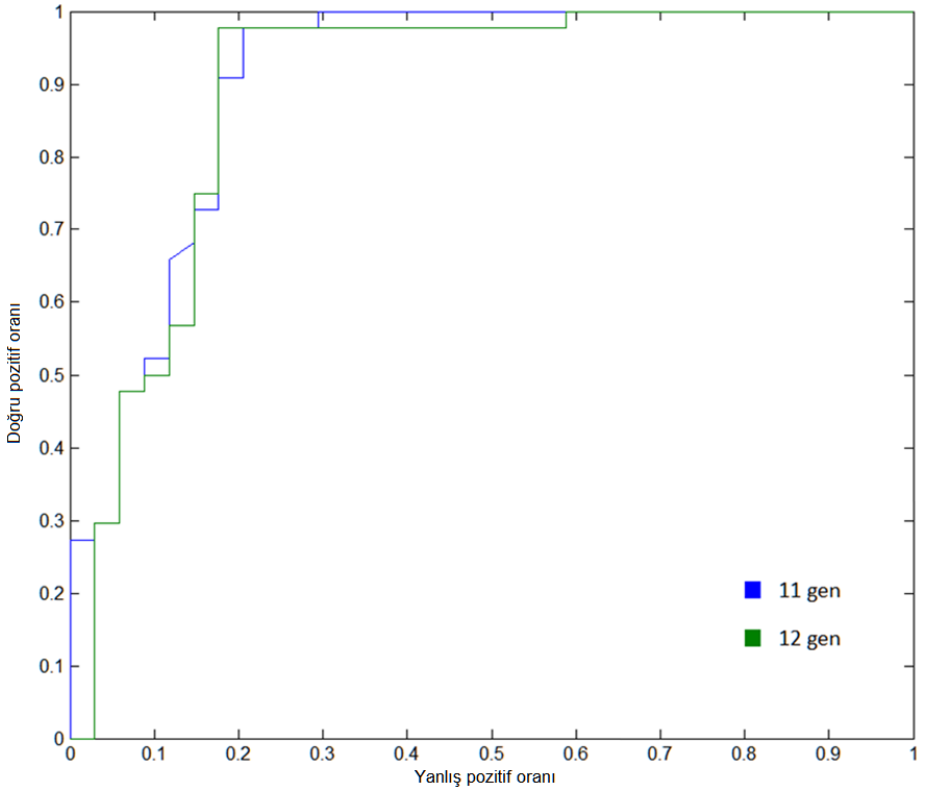
Şekil 5.13. Belirlenen 5 ve 6 gen alt kümesi için ROC eğrileri



Şekil 5.14. Belirlenen 7 ve 8 gen alt kümesi için ROC eğrileri



Şekil 5.15. Belirlenen 9 ve 10 gen alt kümesi için ROC eğrileri



Şekil 5.16. Belirlenen 11 ve 12 gen alt kümesi için ROC eğrileri

5.6. İrdeleme

Meme kanseri sıklıkla görülen ve ölümlere yol açan ciddi bir rahatsızlıktır. Doğru ve erken tanı hastalığın tedavisinde çok önemlidir. Günümüzde kullanılan TNM kanser evreleme sistemi, doğru evreleme yapmak için yeterli değildir. Hastalığın temelinde yatan genetik faktörlerin belirlenmesi, hastalığın daha doğru evrelendirilmesini ve dolayısıyla daha etkin tedavi yöntemlerinin uygulanmasını sağlayabilir.

Bu çalışmada meme kanserinde etkin rol alan genlerin tespit edilebilmesi için yeni bir metot önerilmektedir. Bu amaçla meme kanseri hastalarına ait gen ifade verilerinden faydalanılmıştır. Gen ifade verisi, mikrodizi olarak adlandırılan ve on binlerce gen ifade düzeyini aynı anda ölçmeye yarayan teknoloji ile elde edilir.

Mikrodizi teknolojisi pahalı ve uygulama problemleri olan bir teknolojidir. Aynı zamanda az örnekleme (hasta) ve büyük miktarda niteliğe (on binlerce gen) sahip olması sebebiyle bu verilerin bilgisayar desteği olmaksızın analizi de mümkün değildir.

Nitelik indirgeme yöntemleri gen ifade verilerinin analizinde sıklıkla kullanılmaktadır. Buna rağmen çoğu kez uygun nitelik alt kümesini bulmayı garanti etmemektedir. Önerilen metot filtreleme yöntemi ve genetik algoritma ile birlikte nitelik seçme işlemi yapmaktadır. Metodun gen seçme adımı iki aşamada gerçekleştirilmiştir; ilk aşamada başlangıç veri kümesi indirgenmiş, ikinci aşamada gen seçimi gerçekleştirilmiştir.

FKS, t-skor ve WTS yöntemleri ile gen skorlaması sonucunda elde edilen ilk 2 000 gen alınarak, genetik algoritma yardımıyla en yüksek sınıflandırma başarısı gösteren ve yaklaşık 300 gen bilgisi içeren yeni veri kümesi oluşturulmuştur. Bu veri kümesinde, her bir filtreleme yönteminden yüksek skor almış genler ile birlikte genetik algoritmada yüksek başarı göstermiş genler de yer almıştır.

İkinci aşamada, belirlenen başlangıç veri kümesi içinde sınıflandırma başarısı en yüksek gen veya gen grubunun bulunması amaçlanmıştır. Genetik algoritma için başlangıç popülasyonu 100, iterasyon sayısı 300 olarak belirlenmiştir. Başlangıç popülasyonunun daha fazla arttırılması sonuçları çok değiştirmemiştir. Bunun yanı sıra iterasyon sayısı arttıkça hesaplama zamanının artığı fakat sonuçları etkileyecek önemli bir değişiklik olmadığı görülmüştür.

Genetik algoritma ve destek vektör makinesi ile en yüksek sınıflandırma başarısı gösteren genler belirlenmiştir. Çizelge 5.9'da görüldüğü gibi en yüksek sınıflandırma başarısı %96,15 ile 7 genle elde edilmiştir. Belirlenen bu gen alt kümesi ile 44 pozitif etiketli sınıf içinde 42 örnek doğru tespit edilirken 2 hata yapılmış, 34 negatif etiketli örnek içinde sadece 1 hata yapılarak 33 örnek doğru tespit edilmiştir. Çizelge 5.10 ve Çizelge 5.11'de de görüldüğü gibi belirlenen 7 gen alt kümesi, filtreleme yöntemi ile belirlenen en yüksek skora sahip genler arasında değildir. Ancak bu gen alt kümesi ile en yüksek sınıflandırma başarısı elde edilmiştir. Bu durum, filtreleme yöntemlerinin tek başına uygun nitelik alt kümesini vermede başarılı olamadığını göstermektedir.

Şekil 5.15'te de açıkça görüldüğü gibi belirlenen 7 gen ile gerçekleştirilen sınıflandırıcının performansı diğerlerine göre çok daha üstündür. AUC değerleri karşılaştırıldığında belirlenen 7 gen ile yapılan sınıflandırıcının AUC değeri 0,9606 ile yine en yüksek değer elde edilmiştir. Bu durum seçilen 7 genin diğer gen alt kümelerine göre daha doğru ayrım yaptığını göstermektedir.

Çizelge 5.14'te meme kanseri sınıflandırmasında, aynı gen ifade veri kümesi kullanılarak gerçekleştirilen çalışmaların sınıflandırma başarısı karşılaştırılmıştır. Veer ve arkadaşları [19], korelasyon tabanlı yöntem ile 78 hastanın 65'ini %83 oranında doğru sınıflandırmışlardır. Tan ve Gilbert [77], C4.5, "bagged" ve desteklenmiş karar ağaçları olmak üzere üç farklı makine öğrenme yöntemi ile kanser sınıflandırma çalışması yapmışlardır. Çalışmalarını yedi farklı kanser mikrodizi veri kümesinde test ederek, meme kanserini C4.5 ile %63,16 bagging C4.5 ile %89,47 ve AdaBoost C4.5 ile %89,47 oranında sınıflandırmayı başarmışlardır.

"Bagged" ve desteklenmiş karar ağaçlarının diğer karar ağaçlarından daha iyi bir performans gösterdiğini ortaya koymuşlardır. Michiels ve arkadaşları [78], meme kanseri sınıflandırması için Pearson Korelasyon Katsayısı ile en yüksek skorlu ilk 50 geni belirlemişler, en yakın-ağırlık merkezi yöntemi ile yapılan sınıflandırmada %69 oranında sınıflandırma başarısı elde etmişlerdir. Gevaert ve arkadaşları [79], klinik ve mikrodizi veri kümesi entegrasyonu için; karar entegrasyonu, kısmi entegrasyon ve tam entegrasyon olmak üzere üç yöntem önermişlerdir. Önerilen bu üç yöntemi meme kanseri sınıflandırmasında kullanmışlardır. Belirlenen 70 gen ve bayes ağlarını kullanarak meme kanserini %74 oranında sınıflandırmayı başarmışlardır. Boulesteix ve arkadaşları [80], mikrodizi verilerini değerlendirme için iki adımdan oluşan yeni bir yöntem önermişlerdir. Önerdikleri yöntem ile meme kanserini %70 oranında doğru sınıflandırabilmişlerdir. Chen ve Yang [81] çalışmalarında genetik algoritma ve destek vektör makinesi tabanlı GASVM isimli bir model önermişlerdir. Sınıflandırma doğruluğunu artırmak için klinik ve mikrodizi verilerini birlikte kullanmışlardır. Korelasyon tabanlı belirlenen 70 gen (C70), literatürde yer alan genler (R15), t-Testi ile belirlenmiş genler (T50) ve tüm gen kümesi olmak üzere dört ayrı gen kümesi içinden gen seçimi yapmışlardır. Önerdikleri yöntem ile %94,73 oranında sınıflandırma başarısı elde etmişlerdir.

Çizelge 5.14. Aynı gen ifade verisi kullanılarak gerçekleştirilen çalışmaların sınıflandırma başarısı karşılaştırması

Yazar	**Metot**	**Sınıflandırma Başarısı**
Veer ve arkadaşları	Korelasyon tabanlı	%83
Tan ve Gilbert	Karar ağaçları (C4.5)	%89,47
West ve arkadaşları	Pearson korelasyon katsayısı	%69
Gevaert ve arkadaşları	Bayes ağları	%74
Boulesteix, Porzelius ve Daumer	Rasgele ormanlar, Kısmi en küçük kareler (PLS)	%70
Chen ve Yang	GADVM	%94,73
Yapılan tez çalışması	Nitelik indirgeme ve GA-DVM	%96,15

6. SONUÇ

Meme kanseri kadınlarda en sık görülen ve ölüme neden olan kanser türlerinde ilk sırada yer almaktadır. Diğer kanser türlerinde olduğu gibi erken teşhis ve doğru tedavi hastanın iyileşmesi ve yaşam kalitesinin artırılmasında çok önemlidir.

Günümüzde meme kanserinin everelendirilmesinde TNM sistemi kullanılmaktadır. Ancak bu sistem kanser everelendirmesinde yetersiz kalmakta ve bazen hastaların gereksiz cerrahi müdahalelere veya ağır ilaç tedavilerine maruz kalmalarına neden olmaktadır. Hastalığın temelinde yatan genetik faktörlerin belirlenmesi doğru teşhis ve evrelendirme için oldukça önemlidir.

Gelişen teknoloji ile birlikte on binlerce gen ifade bilgisi aynı anda ölçülebilmektedir. Ancak gen ifade verileri çok az örnekleme sahipken çok büyük miktarda gen bilgisi içerir. Bu genlerin pek çoğu ilgisiz ya da gürültü olarak adlandırılan gen bilgileridir. İlgisiz genlerin atılması ya da etkin genlerin bulunması ciddi bir problemdir. Bu sebeple gen analizlerinin bilgisayar desteği olmaksızın yapılması imkansızdır.

Gen ifade verilerinin indirgenmesinde istatistiksel yöntemler çoğu kez başarısız olmaktadır. Veri madenciliği, gen ifade verilerinin analizinde başarılı bir şekilde kullanılmaktadır. Bu çalışmada, meme kanserinde genetik faktörlerin belirlenmesi amacıyla veri madenciliği teknikleri kullanılarak gen seçimi gerçekleştirilmiştir.

Yapılan çalışma üç adımdan oluşmaktadır: İlk adımda gen ifade verisinde yer alan boş ve hatalı kayıtların düzeltilmesi gerçekleştirilmiştir. Boş ve hatalı kayıtlar sınıflandırma performansını doğrudan etkilemektedir. Bu adımda, gen ifade verisinin az örneklem içermesi sebebiyle hatalı ve boş kayıtlar silinmemiş, bunun yerine tahmine dayalı değer ataması yapılmıştır. Tahmini değer ataması için Öklid kullanılmıştır. Böylece örnek sayısı azaltılmadan veri kümesi gen seçimi için hazırlanmıştır.

Çalışmanın ikinci adımında gen seçimi için genetik algoritma ve destek vektör makinesi kullanılmıştır. Gen ifade verisi yaklaşık 25 000 gen bilgisinden oluşmakta buna karşın sadece 97 hasta bilgisi bulunmaktadır. Genetik algoritma, nitelik seçme işlemlerinde sıklıkla kullanılmakta, ancak arama uzayının çok büyük olması genetik algoritmanın başarısını ve çalışma süresini olumsuz etkilemektedir. Bu sebeple arama uzayının daraltılması gerekmektedir. Bu amaçla filtreleme yöntemleri ile gen ifade veri kümesi indirgenerek başlangıç veri kümesi elde edilmiştir.

Daha önce yapılan genetik algoritma ile nitelik seçme çalışmalarında görülmüştür ki veri kümesi ya doğrudan kullanılmış ya da filtreleme yöntemlerinden biri tercih edilerek yeni arama uzayı elde edilmiştir. Yapılan çalışmada sadece bir filtreleme yöntemine bağlı kalınmamış, üç ayrı filtreleme yöntemi denenmiş ve bunun sonucunda yüksek skor elde edilen tüm genler yeni arama uzayına dahil edilmiştir. Böylece genetik algoritma performansı arttırılmıştır.

Filtreleme için gen analizlerinde başarıyla uygulanan Fisher Korelasyon Skorlama, t-Skor ve WTS yöntemleri kullanılmıştır. Her bir yöntemin arkasından ayrı ayrı genetik algoritma kullanılarak en yüksek sınıflandırma başarısı gösteren gen grubu seçilmiştir. Yaklaşık 300 genden oluşan bu yeni veri kümesi, ikinci adımda gen seçiminde kullanılmak üzere başlangıç veri kümesi olarak belirlenmiştir. Başlangıç veri kümesinde yer alan genlerin, filtreleme yöntemi ile belirlenen genler arasında en yüksek skora sahip olan genler olmadığı görülmüştür. Bu durum, filtreleme yöntemlerinin gen ifade verilerinde nitelik seçme için tek başına yeterli olamayacağını göstermektedir. Filtreleme yöntemleri, en uygun alt nitelik kümesini vermeyi garanti etmese de genetik algoritma için başlangıç veri kümesini belirlemede önemli bir adım olmuştur.

Yapılan çalışmada belirlenen 7 gen alt kümesi ile meme kanseri hastalarında %96,15 doğruluk oranında sınıflandırma başarısı elde edilmiştir. Önerilen metot, literatürdeki diğer yöntemlere göre, daha az gen ile daha yüksek sınıflandırma başarısı göstermesi nedeniyle daha başarılı bir gen seçimi gerçekleştirmektedir. Sınıflandırma performansı, 10 kat çapraz doğrulama ve ROC eğrileri ile test edilmiştir. Belirlenen 7

gen alt kümesi ile yapılan sınıflandırma sonucunda AUC değeri 0,9606 olarak elde edilmiştir. Bu durum sınıflandırma performansının oldukça başarılı olduğunu göstermektedir.

Tespit edilen 7 gen ile yüksek doğrulukta meme kanseri evrelendirmesi yapılabilir. Ayrıca yeni gen çipleri tasarlanabilir. Daha az gen daha az maliyetli gen çipi tasarımı anlamına gelmektedir. Kanser gibi ciddi hastalıklara sebep olan genlerin belirlenmesi, bu hastalıkların tedavisinde ve yeni ilaçların geliştirilmesinde büyük bir önem arz etmektedir. Bu nedenle bu alanda yapılan çalışmalar önem kazanmaktadır.

Bu çalışmada meme kanserine neden olan etkin genler tespit edilmiştir. Önerilen metot ile diğer kanser türlerinde de hastalığa neden olan genler tespit edilebilir. Hastalıkların temelinde yatan genetik faktörlerin belirlenmesi bu tür ciddi hastalıkların tanı ve tedavisine katkı sağlayabilir.

Filtreleme yöntemleri, uygun gen altkümesini vermeyi garanti etmese de arama uzayının daraltılmasında kullanılabilir. Bu çalışma kapsamında filtreleme ile elde edilen sonuçlar, IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı'nda [82] yayımlanmıştır. Önerilen metot ile ve gen seçimini gerçekleştirdiğimiz çalışma, Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi'nde [83] yayımlanmıştır.

KAYNAKLAR

1. İnternet : GLOBOCAN 2008 Cancer Incidence, Mortality and Prevalence Worldwide in 2008 "Breast Cancer Incidence, Mortality and Prevalence Worldwide in 2008 Summary" http://globocan.iarc.fr (2012).

2. Phillips, G., Sharkey, P., Mors, S., "Mining lung cancer patient data to assess healthcare resource utilization", ***Expert Systems with Applications***, 35 (4):1611-1619 (2008).

3. Chang, C.L., Chen, C.H., "Applying decision tree and neural network to increase quality of dermatologic diagnosis", ***Expert Systems with Applications,*** 36 (2): 4035-4041 (2009).

4. Lin, K.S., Chien, C.F., "Cluster analysis of genome-wide expression data for feature extraction", ***Expert Systems with Applications***, 36 (2): 3327-3335 (2009).

5. Walker, P.R., Smith, B., Liu, Q.Y., Famili, A.F., Valde, J.J., Liu, Z., Lach, B., "Data mining of gene expression changes in Alzheimer brain", ***Artificial Intelligence in Medicine***, 31: 137-54 (2004).

6. Chen, Y., Zhao, Y., "A novel ensemble of classifiers for microarray data classification", ***Applied Soft Computing,*** 8: 1664-1669 (2008).

7. Tseng, M.H., Liao, H.C., "The genetic algorithm for breast tumor diagnosis-The case of DNA viruses", ***Applied Soft Computing***, 9: 703-710 (2009).

8. An, J., Chen, Y.P.P., "Finding rule groups to classify high dimensional gene expression datasets", ***18th International Conference on Pattern Recognition,*** Hong Kong, 119-1199 (2006).

9. Armañanzas, R., Inza, I., Larrañaga, P., "Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers", ***Computer Methods and Programs in Biomedicine***, 91 (2): 110-121 (2008).

10. Das, R., Turkoglu, I., Sengur. A., "Diagnosis of valvular heart disease through neural networks ensembles", ***Computer Methods and Programs in Biomedicine***, 93: 185-191 (2009).

11. Giannakeas, N., Fotiadis, D.I., "An automated method for gridding and clustering-based segmentation of cDNA microarray images", ***Computerized Medical Imaging and Graphics***, 33: 40-49 (2009).

12. Zhang, X., Song, X., Wang, H., Zhang, H., "Sequential local least squares imputation estimating missing value of microarray data", ***Computers in Biology and Medicine***, 38: 1112-1120 (2008).

13. Horng, J.T., Wub, L.C., Liu, B.J., Kuo, J.L., Kuo, W.H., Zhang, J.J., "An expert system to classify microarray gene expression data using gene selection by decision tree", ***Expert Systems with Applications***, 36: 9072-9081 (2009).

14. Li, L., Jiang, W., Li, X., Moser, K.L., Guo, Z., Du, L., Wang, Q., Topol, E., Wang, Q., Rao, S., "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset", ***Genomics***, 85:16-23 (2005).

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

16. Uriarte, R., D. ve Andrés, S.,A., "Gene selection and classification of microarray data using random forest", ***BMC Bioinformatics***, 7: 1-13 (2006).

17. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. ve Levine, A. J., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", ***Cell Biology***, 96: 6745-6750 (1999).

18. Dudoit, S., Fridlyand, J ve Speed, T., P., "Comparison of discrimination methods for the classification of tumors using gene expression data", ***American Statistical Association***, 97: 77-87 (2002).

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

20. Yıldırım, A, Bardakçı, F., Karataş, M., Tanyolaç, B., "Moleküler Biyoloji", ***Nobel Yayın***, Ankara, 55-600 (2010).

21. Wang, J., Zaki, M., Toivonen, H., Shasha, D., "Data Mining in Bioinformatics", ***Springer***, USA, 3-8 (2005).

22. Huerta, E.B., Duval, B., Hao, J.K., "A hybrid GA/SVM approach for gene selection and classification of microarray data", ***EvoWorkshops***, ***LNCS 3907, Springer-Verlag Berlin Heidelberg***, 34-44 (2006).

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

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

25. Antonie, M.L., Zaiane, O.R., Coman, A., "Application of data mining techniques for medical image classification", ***Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco***, USA (2001).

26. Khan, J., Wei, J.S., Ringnér, M., Saal, L.H:, Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", ***Nature Medicine***, 7 (6): 673-679 (2001).

27. Parka, C., Koob, J.Y., Kimc, S., Sohnb, I., Leeb, J.W., "Classification of gene functions using support vector machine for time-course gene expression data", ***Computational Statistics & Data Analysis***, 52: 2578-2587 (2008).

28. Ao, S.L., "Data Mining and Applications in Genomics", ***Springer***, USA, 30-33 (2008).

29. Alpaydın, E., "Yapay Öğrenme", ***Boğaziçi Üniversitesi Yayınevi***, İstanbul, 10-45,90-107,262 (2011).

30. Kadirkamanathan, V., Girolami, M., Sanguinetti, G., Niranjan, M., "Pattern Recognition in Bioinformatics", ***Springer***, USA, 282 (2009).

31. Slaby, A., "ROC analysis with Matlab", ***Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces***, Croatia, 25-28 (2007).

32. Sokolova, M., Lapalme, G.," A systematic analysis of performance measures for classification tasks", ***Information Processing and Management***, 45:427-437 (2009)

33. Han, J., Kamber, M., "Data Mining: Concepts and Techniques Second Edition", ***Morgan Kaufmann Publisher***, San Francisco, 282-286, 337-385 (2006).

34. Setiono, R., Liu, H., "Effective Data Mining Using Neural Networks", ***IEEE Transactions On Knowledge And Data Engineering***, 8 (6): 657-961 (1996).

35. Lee, Y., Lee, C.K., "Classification of multiple cancer types by multicategory support vector machines using gene expression data", ***Bioinformatics***, 19 (9):1132-1139 (2003).

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

37. Özkan, Y., "Veri Madenciliği", ***Papatya***, İstanbul, 188-195 (2008).

38. Giannopoulou, E.G., "Data Mining in Medical and Biological Research", ***In-Teh***, Croatia,143 (2008).

39. Zhang, P., "Model selection via multi-fold cross validation", ***The Annals of Statistics***, 21:299-313 (1993).

40. Lasko, T.,A., Bhagwat, J., G., Zou, K., H. ve Ohno-Machado, L., "The use of receiver operating characteristic curves in biomedical informatics", ***Journal of Biomedical Informatics***, 38: 404-415 (2005).

41. Çamlıca, H., Dişçi, R., "Tanı testlerinde sınır değerlerinin belirlenmesi", ***Türk Onkoloji Dergisi***, 23(1): 26-33 (2008).

42. Metsis, V., Huang, H., Makedon, F., Tzika, A., "Heterogeneous data fusion to type brain tumor biopsies", ***Proceedings of the 5th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI 2009)***, Greece, 23-25 (2009).

43. Loo, B.L., Roberts, S., Hrebıen, L., Kam, M.,"New Criteria for Selecting Differentially Expressed Genes", ***IEEE Engineering In Medicine And Biology Magazine***, 26 (2): 17 – 26 (2007).

44. Huang, C.L. ve Wang, C.J., "A GA-based feature selection and parameters optimization for support vector machines", ***Expert Systems with Applications***, 31:231-240 (2006).

45. Liu,H. ve Motoda, H., "Computational methods of feature selection", ***Chapman & Hall/CRC, Taylor & Francis Group***, 6000 Broken Sound Parkway NW, 26-27 (2008).

46. Maldonado, S., Weber, R., "A wrapper method for feature selection using support vector machines", ***Information Sciences***, 179: 2208-2217 (2009).

47. Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A.J., "Filter versus wrapper gene selection approaches in DNA microarray domains", ***Artificial Intelligence in Medicine***, 31: 91-103 (2004).

48. Peng, Y., Wu, Z., Jiang, J., "A novel feature selection approach for biomedical data classification", ***Journal of Biomedical Informatics***, 43: 15-23 (2010).

49. Kohavi, R., John, G.H., "Wrappers for feature subset selection", ***Artificial Intelligence***, 97: 273-324 (1997).

50. Şen, Z., "Genetik Algoritma ve En İyileme Yöntemleri", ***Su Vakfı Yayınları***, İstanbul, 22-82 (2004).

51. Ooi, C.H., Tan, P., "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data", ***Bioinformatics***, 19: 37-44 (2003).

52. Li, S., Wu, X., Hu, X., "Gene selection using genetic algorithm and support vectors machines", ***Soft Comput***, 12:693–698 (2008).

53. Hernandez, J.C.H., Duval, B., Hao, J.K, "A genetic embedded approach for gene selection and classification of microarray data", ***EvoBIO 2007***, Springer-Verlag Berlin Heidelberg, 90-101 (2007).

54. Shah, S., Kusiak, A., "Data mining and genetic algorithm based gene/SNP selection", ***Artificial Intelligence in Medicine***, 31: 183-196 (2004).

55. Devillers, J., "Genetic Algorithms in Molecular Modeling", ***Academic Press***, San Diego, 35-64 (1996).

56. Kutluk, T., Kars, A., "Kanser Konusunda Genel Bilgiler", ***Sağlık Bakanlığı Sağlık Projesi Genel Müdürlüğü***, 17 (2001)

57. Akbaş, E., Seyrek, E., Şenli, H.M., Erdoğan, N.E., Helvacı, İ., "Beta adrenerjik reseptör 2 kodon 27 polimorfizmi ile meme kanseri arasındaki ilişkinin araştırılması", ***The Journal of Breast Health***, 7 (1): 5-9 (2011).

58. Somunoğlu, S., "Meme kanserinde risk faktörleri", ***Fırat Sağlık Hizmetleri Dergisi***, 2 (5):3-12 (2007).

59. Bozcuk, A.N., "Genetik", ***Palme Yayıncılık***, Ankara, 5-10 (2011).

60. Bayrak, İ.K., Nural, M.S., Elmalı, M., Baydın, M., "Erkek memesinde infiltratif duktal karsinom: Mamografi, ultrason ve dinamik manyetik rezonans görüntüleri", ***The Journal of Breast Health***, 3 (3): 160,162 (2007).

61. Sherman, C.D., Calma, K.C., Hossfeld, D.K., "Klinik Onkoloji", ***Sağlık Bakanlığı Türk Kanser Araştırma ve Savaş Kurumu***, Ankara, 162-164 (1990).

62. Eroglu, C., Eryılmaz, M.A., Cıvcık, S., Gürbüz, Z., "Meme kanseri risk değerlendirmesi: 5000 olgu", ***International Journal of Hematology and Oncology***, 20 (2):27-33 (2010).

63. Klemi, P.J., Parvinen, I., Pylkkänen, L., Kauhava, L., Räihä, P.I., Räsänen, O., Helenius, H.,"Significant improvement in breast cancer survival through population-based mammography screening", ***The Breast***, 12, 308-313 (2003).

64. Nadkarni, M.S:, Gupta, P.S., Parmar, V.V., Badwe, R.A., "Breast conservation surgery without pre-operative mammography—A definite feasibility", ***The Breast***, 15 (5): 595-600 (2006).

65. Arıbal, E., Tunçbilek, N., Çelik, L., "Türk radyoloji derneği meme radyolojisi çalışma grubu meme kanseri radyolojik tarama standartları", ***The Journal of Breast Health***, 8 (1): 3-10 (2012).

66. Flobbe, K., Nelemans, P.J., Kessels, A.G.H., Beets, G.L ., von Meyenfeldt, M.F., van Engelshoven, J.M.A., "The role of ultrasonography as an adjunct to mammography in the detection of breast cancer: a systematic review", ***European Journal of Cancer***, 38 (8): 1044-1050 (2002).

67. Harms, K., Wittekind, C., "Prognosis of women with pT4b breastcancer: The significance of this category in the TNM system", ***European Journal of Surgical Oncology (EJSO)***, 35 (1); 38–42 (2009).

68. Cherbit, A.L., Gal, M.L., Asselain, B., Neuenschwander, S., "Breastcancer: zones of increased density mammographic features, correlated to clinical TNM and prognosis", ***European Journal of Radiology***, 24 (1): 48-53 (1997).

69. Coburn, N.G., Pearsonb, E.C., Chung, M.A., Law, C., Fulton, J., Cady, B., "A novel approach to T classification in tumor-node-metastasis staging of breast cancer", ***The American Journal of Surgery***, 192 (4): 434-438 (2006).

70. Ferahman, M., "Meme kanserinde güncel TNM evrelemesi", ***İ.Ü. Cerrahpaşa Tıp Fakültesi Sürekli Tıp Eğitimi Etkinlikleri, Meme Kanseri Sempozyum Dizisi***, 54: 87 - 91 (2006).

71. Engin, K., "Meme Kanseri", ***Nobel Kitapevleri***, İstanbul, 81 (2005).

72. Yüksel, B.C., Yıldız, Y., Öztürk, B., Berkem, H., Katman, U., Ozel, H., Hengirmen, S., "Serum tumormarker CA19-9 in the follow-up of patients with cystic echinococcosis", ***The American Journal of Surgery***, 195 (4): 452-456 (2008).

73. Ekmekçi, A., "Gen, Genetik Değişim ve Hastalıklar", ***Gazi Kitapevi***, Ankara, 1-10 (2006).

74. Mohammadi, A., Saraee, M.H., Salehi, M., "Identification of disease-causing genes using microarray data mining and Gene Ontology", ***BMC Medical Genomics***, 4(12): 2-9 (2011).

75. Smith, A.A., "Classification and alignment of gene-expression time-series data", Doktora tezi, ***University Of Wisconsin Madison,*** USA, 14-18 (2009).

76. İnternet : Bioinformatics and Statistics Division of Molecular Carcinogenesis, Netherlands Cancer Institute "Breast Cancer Dataset" http://bioinformatics.nki.nl/data.php (2012).

77. Tan, A.C., Gilbert, D., "Ensemble Machine Learning On Gene Expression Data For Cancer Classification", ***Bioinformatics***, 2 (3) :S75-S83 (2003).

78. Michiels, S., Koscielny, S., Hill, C., "Prediction of Cancer Outcome with Microarrays: A Multiple Random Validation Strategy", ***Lancet***, 365: 488-492, (2005).

79. Gevaert, O., Smet, F.D., Timmerman, D., Moreau, Y., MoorPredicting, B.D., "Predicting The Prognosis Of Breast Cancer By Integrating Clinical And Microarray Data With Bayesian Networks", ***Bioinformatics***, 22(14): e184–e190 (2006).

80. Boulesteix, A.L., Porzelius, C., Daumer, M. “Microarray-Based Classification And Clinical Predictors: On Combined Classifiers And Additional Predictive Value”, ***Bioinformatics***, 24(15), 1698-1706 (2008).

81. Chen, A.H., Yang, C., “The İmprovement Of Breast Cancer Prognosis Accuracy From Integrated Gene Expression And Clinical Data”, ***Expert Systems with Applications***, 39: 4785-4795, (2012).

82. Yıldız, O., Tez, M., Bilge, H.Ş., Akcayol, M.A., Güler, İ., "Meme Kanseri Sınıflandırması İçin Gen Seçimi", ***IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)*** 2012

83. Yıldız, O., Tez, M., Bilge, H.Ş., Akcayol, M.A., Güler, İ. ,”Meme kanseri sınıflandırması için veri füzyonu ve genetik algoritma tabanlı gen seçimi”, ***J. Fac. Eng. Arch. Gazi Univ.*** (2012)

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : YILDIZ, Oktay
Uyruğu : T.C.
Doğum tarihi ve yeri : 01.04.1975 Sorgun
Medeni hali : Evli, 1 çocuklu
Telefon : 0 (312) 582 31 16
e-mail : oyildiz@gazi.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet tarihi
Yüksek lisans	Gazi Üniversitesi / Fen Bilimleri Enst.	2004
Lisans	Gazi Üniversitesi/ Elektronik-Bilgisayar Eğt.	1997

İş Deneyimi

Yıl	Yer	Görev
1999 - 2009	Gazi Üniversitesi Müh. Mim. Fak.	Uzman
2009 -	Gazi Üniversitesi Müh. Fak. Bilg. Müh. Böl.	Öğr.Gör.

Yabancı Dil

İngilizce

Yayınlar

1. Yıldız, O., Tez, M., Bilge, H.Ş., Akcayol, M.A., Güler, İ., "Meme Kanseri Sınıflandırması İçin Gen Seçimi", ***IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)*** (2012).

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

3. Şahin, E., Mutlu, B., Yıldız, O., Öztürk, E.A., "Yapay Zeka Tabanlı Trafik Planlama Uygulaması (YZTTPU)", ***IEEE 20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)*** (2012).

4. Çetinyokuş, T., Dağdeviren, M., Yıldız, O., "Personel Seçiminde Eşleşme Yöntemi Temeline Dayanan Bir Uzman Sistem Yaklaşımı", ***e-Journal of New World Sciences Academy***, 5(4):590-602 (2010).

5. Yıldız, O., Dağdeviren, M. ve Çetinyokuş, T. "İşgören Performansının Değerlendirilmesi İçin Bir Karar Destek Sisteminin Geliştirilmesi Ve Uygulaması", ***Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi***, 23(1):239-248 (2008).

6. Toklu S., Simsek M., Yıldız, O., Bay I., Simsek A., Akcayol M.A.,"A Priority Based Protocol And Load Balancing for Queue Management On Wireless Networks ", ***2008 International Conference on Wireless Networks (ICWN'08),*** Monte Carlo Resort, Las Vegas, Nevada, USA, 14-17 (2008).

7. Dogru I. A., Simsek M., Toklu S., Yıldız, O., Akcayol M.A., "Rule-Based Mobility Management Routing For AD HOC Networks", ***2008 International Conference on Wireless Networks (ICWN'08)***, Monte Carlo Resort, Las Vegas, Nevada, USA, 14-17 (2008).

8. Yıldız, O., Dağdeviren, M. ve Çetinyokuş, T., "Performans Değerlendirme Amaçlı Bir Karar Destek Sisteminin Geliştirilmesi: Bir KOBİ Uygulaması,", ***4. KOBI'ler ve Verimlilik Kongresi***, İstanbul Kültür Üniversitesi, İstanbul, 297-306 (2007).

9. Çetinyokuş, T., Yıldız, O., Dağdeviren, M. ve Gökçen, H., "KOBİ'lerde Personel Seçimine Yönelik Bir Uzman Sistem Yaklaşımı", ***4. KOBI'ler ve Verimlilik Kongresi,*** İstanbul Kültür Üniversitesi, İstanbul, 307-316 (2007).