T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TRANSKRİPTOM VERİ SETİ ÜZERİNDE DERİN ÖĞRENME YÖNTEMİ İLE KLASİK VERİ MADENCİLİĞİ YÖNTEMLERİNİN SINIFLAMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Merve KAŞIKCI

Biyoistatistik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2019

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TRANSKRİPTOM VERİ SETİ ÜZERİNDE DERİN ÖĞRENME YÖNTEMİ İLE KLASİK VERİ MADENCİLİĞİ YÖNTEMLERİNİN SINIFLAMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Merve KAŞIKCI

Biyoistatistik Programı
YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI
Prof. Dr. Erdem KARABULUT

ANKARA
2019

HACETTEPE ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TRANSKRİPTOM VERİ SETİ ÜZERİNDE DERİN ÖĞRENME İLE KLASİK VERİ MADENCİLİĞİ YÖNTEMLERİNİN SINIFLAMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Merve KAŞIKCI

Danışman: Prof. Dr. Erdem KARABULUT

Bu tez çalışması 05.08.2019 tarihinde jürimiz tarafından "Biyoistatistik Programı" nda yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	*Prof. Dr. Pınar ÖZDEMİR (Hacettepe Üniversitesi)*
Tez Danışmanı:	*Prof. Dr. Erdem KARABULUT (Hacettepe Üniversitesi)*
Üye:	*Doç. Dr. Serdal Kenan KÖSE (Ankara Üniversitesi)*

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

28 Ağustos 2019

Prof. Dr. Diclehan Orhan

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- ○ Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. [1]
- ● Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 6. ay ertelenmiştir. [2]
- ○ Tezimle ilgili gizlilik kararı verilmiştir. [3]

28.08.2019

M. Kaşıkcı

Merve KAŞIKCI

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

(1) *Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

(2) *Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

(3) *Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Prof. Dr. Erdem KARABULUT danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

M. Kaşıkcı

Merve KAŞIKCI

TEŞEKKÜR

Yüksek lisans eğitimim ve tez dönemim boyunca bilgi ve deneyimleriyle her zaman yol gösteren, yardım ve rehberliğini esirgemeyen danışman hocam Prof. Dr. Erdem KARABULUT'a, bu tezin hazırlandığı tüm süreçte yanımda olan, motive eden, gerek analizlerin yapılması, gerekse tez yazım sürecinde değerli bilgilerini benimle paylaşan, analizlerin yapılabilmesi için gerekli donanım ve yazılımlara ulaşmamı sağlayan, emeklerinin karşılığını asla ödeyemeyeceğim Sayın Dr. Erdal COŞGUN'a, yüksek kapasiteli işlem gücüne sahip bir bilgisayara erişim sağlayarak analizlerin gerçekleştirilmesine destek olan Microsoft Genetik Ekibi'ne, önerileri ve bilgileriyle her konuda yardımcı olan Biyoistatistik ve Biyoinformatik Anabilim Dallarında bulunan tüm hocalarım ve asistan arkadaşlarıma, bu süreçte destek olan ailem ve arkadaşlarıma çok teşekkür ediyorum.

ÖZET

Kaşıkcı, M., Transkriptom Veri Seti Üzerinde Derin Öğrenme Yöntemi ile Klasik Veri Madenciliği Yöntemlerinin Sınıflama Performanslarının Karşılaştırılması, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Programı Yüksek Lisans Tezi, Ankara, 2019. Bu tez çalışmasında akciğer kanseri ve böbrek kanserine ait RNA dizileme verileri kullanılarak, kanser alt türlerinin sınıflandırılmasında klasik veri madenciliği yöntemleri ve Derin Öğrenme yöntemi kullanılmıştır. Çalışmada kullanılan klasik veri madenciliği yöntemleri Yapay Sinir Ağları, Rastgele Orman ve Destek Vektör Makineleri'dir. Sınıflama yöntemlerinin performansları doğruluk, Kappa, F ölçütü gibi ölçüler kullanılarak karşılaştırılmıştır. Akciğer kanseri veri seti iki sınıflı ve sınıf dağılımları dengeli bir veri seti iken böbrek kanseri veri setinde üç sınıf vardır, sınıflardaki gözlem sayıları dengesizdir. Sınıflamada kullanılan gen setleri, farklı filtreler uygulanarak elde edilmiştir. Böylece, farklı özellikte veri setlerinde ve farklı filtrelerde sınıflama yöntemlerinin performansları incelenmiştir. Her sınıflama yöntemi için, yöntemlere ait belirli parametrelerin alabileceği değer aralıkları belirlenmiş ve eğitim setleri üzerinde denenerek uygun parametre seçimi gerçekleştirilmiştir. Çalışmada kullanılan veri setlerinde, klasik veri madenciliği yöntemlerine göre daha derin bir yapıya sahip olan Derin Öğrenme yöntemi, genel olarak başarılı bir performans göstermiştir. Performans karşılaştırmada kullanılan ölçüler bakımından tek katmanlı klasik Yapay Sinir Ağı, diğer yöntemlere göre daha düşük değerler almıştır.

Anahtar Kelimeler: RNA dizileme, kanser, veri madenciliği, sınıflama yöntemleri, Derin Öğrenme

Tezi Destekleyen Kuruluş: Microsoft Genetik Ekibi

ABSTRACT

Kaşıkcı, M., Comparison of Classification Performance for Deep Learning Method and Classical Data Mining Methods on Transcriptome Data Set, Hacettepe University Graduate School of Health Sciences Master Thesis in Biostatistics, Ankara, 2019. In this thesis, Artificial Neural Networks, Random Forest, Support Vector Machines, which are classical data mining methods, and Deep Learning method were used to classify the cancer subtypes. The performances of these methods were compared by using some performance comparison measures like accuracy, Kappa and F measure. For this reason, two different RNA sequencing data sets were used. The first data set is the lung cancer data set which has two classes. It is a balanced data set in terms of class size. The other data set is renal cancer data set. This data set contains three classes and the number of observation in these classes are uneven. Gene sets used in the classification were obtained by using different filters. Therefore the performances of the classification methods in different data sets and filters were examined. For each classification method, specific parameters were optimized and the most appropriate parameters were selected. Deep Learning method which has a deeper structure compared to classical data mining methods, showed a successful performance on the data sets used in this study. In terms of the measurements used in performance comparison, the classical single-layer Artificial Neural Network has lower values compared to other methods.

Key Words: RNA sequencing, cancer, data mining, classification methods, Deep Learning

Organization Supporting the Thesis: Microsoft Genomics Team

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

CART	Sınıflama ve Regresyon Ağacı
CNTK	Microsoft Cognitive Toolkit
cDNA	Tamamlayıcı DNA
DNA	Deoksiribo Nükleik Asit
DÖ	Derin Öğrenme
DVM	Destek Vektör Makineleri
EAKA	Eğri altında kalan alan
GSA	Gen Belirleyici Analiz
fdr	Yanlış bulgu oranı
mRNA	Mesajcı RNA
ReLU	Düzleştirilmiş Doğrusal Birim
RNA	Ribo Nükleik Asit
RO	Rastgele Orman
rRNA	Ribozomal RNA
RTF	Radyal Tabanlı Fonksiyon
Tanh	Hiperbolik Tanjant Fonksiyonu
TCGA	The Cancer Genome Atlas
tRNA	Taşıyıcı RNA
YSA	Yapay Sinir Ağları

ŞEKİLLER

TABLOLAR

1. GİRİŞ

Teknolojinin gelişmesi ile birlikte, verilere ulaşılabilirlik ve elde edilen verilerin büyüklüğü artmakta, buna bağlı olarak da depolama ve işleme kapasiteleri giderek gelişmektedir. Elde edilen büyük ölçekli ve karmaşık yapıdaki verilerin analiz edilmesinde klasik yöntemler yetersiz kalabilmektedir. Bu nedenle klasik yöntemlerin belirli özelliklerinin birleştirilmesi veya geliştirilmesi ile daha derin mimarilere sahip yöntemler geliştirilmektedir. Bu yöntemlerden biri olan Derin Öğrenme, hem sınıf etiketi olmayan verilerde boyut azaltma, kümeleme gibi amaçlarla kullanılabilmekte hem de sınıf etiketi olan verilerde, uygun sınıf etiketi atanarak sınıflama modellerinin geliştirilmesinde kullanılmaktadır. Yapılan çalışmalarda, genellikle kullanıldığı amaçlar için başarılı sonuçlar verdiği görülmektedir. Bu durum daha farklı alanlarda uygulanması için araştırmacılara cesaret vermektedir.

Sağlık ve biyoinformatik alanlarında, toplumun belirli konulardaki algısını ortaya koymak, hastalıkların tespit edilmesinde farklı yöntemler geliştirmek gibi amaçlarla veriler toplanmakta ve depolanmaktadır. Hastalıkların teşhislerinin konulmasında biyomedikal görüntüleme ve sinyal işleme verileri sık olarak kullanılmaktadır. Bu verileri kullanarak yapılmak istenen hasta-sağlıklı veya hastalığın alt türlerinin sınıflandırılmasında derin öğrenme algoritmaları başarılı sonuçlar vermiştir. Hastalıkların araştırılmasında önemli veri kaynaklarından biri genetik verilerdir. Omik türü veri olarak da adlandırılan veriler sayesinde hastalıklar gen, transktipt, protein, metabolom bazında incelenebilmektedir. Gelişen teknoloji ile birlikte, bu alanlarda daha çok ve daha kesin verilere ulaşılabilmektedir. Son yıllarda, hastalık ve gen ilişkisinin incelenmesinde, RNA dizileme yöntemi tercih edilen bir yöntemdir.

Bu çalışmanın amacı, farklı kanser türlerini içeren RNA dizileme verilerinden elde edilen gen ifade seviyelerini, kanser alt türleri için, derin öğrenme ile sınıflamak, seçilen belirli klasik veri madenciliği yöntemlerinin sonuçlarını da karşılaştırma için kullanmaktır. Çalışmada kullanılan veri setlerinde bulunan genlerin hepsinin sınıflamaya dahil edilmesi hem işlem süresini uzatacağından, hem de performansları düşüreceği düşünüldüğünden öncelikle boyut azaltma işlemi gerçekleştirilmiştir. Burada iki aşamalı bir yol izlenmiş, öncelikle kanser alt türleri arasında anlamlı olarak farklılık gösteren genler seçilmiş, ardından yanlış bulgu oranı filtre olarak

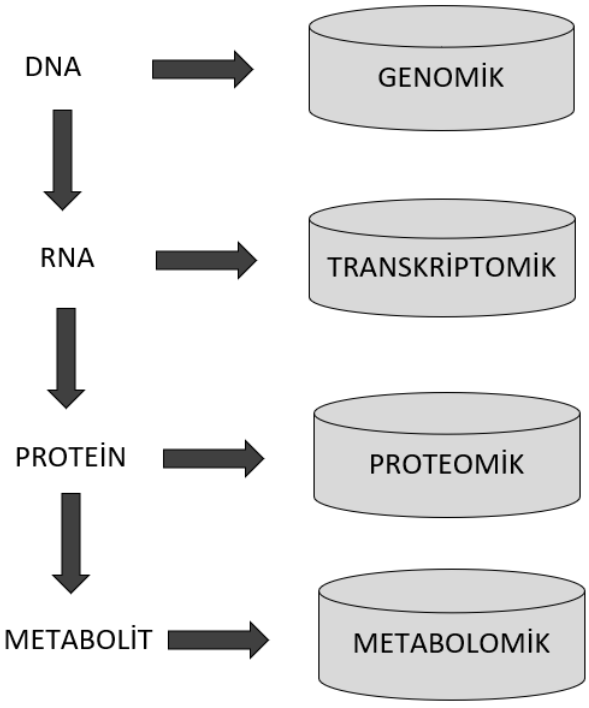
kullanılmıştır. Dengeli ve dengesiz sınıf dağılımlarına sahip iki veri seti seçilmiş, böylelikle sınıflama yöntemlerinin performanlarının farklı filtreler ve farklı özellikteki veri setlerinde nasıl olacağı gözlemlenmek istenmiştir.

Çalışma 5 ana bölümden oluşmaktadır. Bölüm 2'de çalışmada kullanılan veri seti olan transkriptom veri setinden, verinin elde edilmesinden, sınıflamaya uygun hale getirilme sürecinden, sınıflamada kullanılan yöntemlerden ve performansları karşılaştırmada kullanılan ölçülerden bahsedilmiştir. Bölüm 3' de veri setleri hakkında bilgi verilmiştir. Hangi sınıflama yöntemlerinde, hangi parametrelerin seçildiğine açıklık getirilmiştir. Bölüm 4'de sınıflama analizlerinin sonuçları her veri seti ve boyut azaltma yöntemlerine göre, tablo ve grafiklerle özetlenmiştir. Bölüm 5'de sınıflama yöntemlerinin performansları karşılaştırılmış, alanyazında yer alan benzer çalışmaların sonuçlarından bahsedilmiştir. Son bölüm olan Bölüm 6'da çalışmada ulaşılan genel sonuçlardan ve ileride yapılacak çalışmalarda nelerin amaçlandığından söz edilmiştir.

2. GENEL BİLGİLER

2.1. Omikler (Omics)

Omik kavramı genomik, transkriptomik, proteomik, metabolomik, epigenomik, lipidomik, nutrigenomik gibi pek çok bilim dalının son eki olan "-omik" bölümünden türetilmiş olup tüm bu alanları içeren bir kavram olarak kullanılmaktadır. Omik bilimi genel olarak, omik ekinin eklendiği moleküler terime ilişkin biyolojik moleküllerin kapsamlı bir şekilde incenlenmesini konu alır (1). Günümüzde 1000'den fazla alt türü olsa da omik alanında en çok karşılaşılan çalışmalar genomik, transkriptomik, proteomik ve metabolomik alanına ait çalışmalardır. Genomik alanında elde edilen veriler genlerin anlık dizilişlerini içermekte iken transkriptomik, proteomik ve metabolomik türü veriler ise gen ürününün biyolojik fonksiyonu ortaya koymaktadır. Omik türü verilerin elde edilmesini sağlayan araçlar sayesinde aynı anda gen, transkriptom, protein, metabolit sayılarının ölçülmesi sağlanır, böylece kısa sürede büyük ölçekli veriler elde edilir (Şekil 2.1).



Şekil 2.1. Genomik, transkriptomik, proteomik ve metabolomik ilişkisi.

2.1.1. Genomikler (Genomics)

Genom, bir organizmanın oluşumu ve varlığını sürdürebilmesi için gerekli olan kalıtsal bilgiyi, başka bir ifade ile organizmanın sahip olduğu DNA dizilerinin tamamını ifade eder. Genomik, bir organizmaya ait genomun incelenmesini konu alır. İnsan genomunu oluşturan DNA baz çiftlerinin tam ve doğru olarak sıralanmasını sağlanması, genlerin yerlerinin belirlenmesi, genetik hastalıklar ile genlerin ilişkisini araştırmak gibi amaçlarla 1990'lı yılların başında İnsan Genom Projesinin başlaması ile bağımsız bir alan haline gelen genomik, omik çalışmalarının başlangıcı kabul edilir. Bu alanında yapılan çalışmalar nükleotid dizileri, genomun yapısı, yapısal farklılıklar, gen ifade düzeyleri gibi özelliklerin tanımlanması, ölçülmesi ve karşılaştırılmasını kapsamaktadır (5). Genomik analizler genetik farklılıkların ortaya çıkartılmasında önemli bir rol oynar. Özellikle hastalıklara ilişkili genlerin ve yolakların tanımlanması için yapılan genomik analizler hastalık mekanizmasının anlaşılmasında, yeni biyobelirteçler tanımlanarak hastalığın erken ve doğru teşhis edilmesinde önemli bilgiler sağlar (3). Ancak transkripsiyon ve translasyon sonrasında meydana gelen değişiklikler nedeniyle DNA'nın en son ortaya çıkacak biyolojik etkisini yalnızca genomik analizler ile tahmin etmek zordur.

2.1.2. Transkriptomikler (Transcriptomics)

Bir organizmanın kalıtsal tüm bilgisini taşıyan çift sarmallı DNA'nın sarmallarından biri kalıp olarak kullanılıp belirli bir bölgedeki bilgi kopyalanarak tek sarmallı yapıya sahip, genetik bilgi akışını sağlayan RNA molekülü oluşturulur. Bu işleme transkripsiyon adı verilir. DNA'dan farklı olarak RNA oluşumu sırasında Timin nükleotidinin yerini Urasil nükleotidi alır. RNA çeşitleri genel olarak kodlayan ve kodlamayan olarak iki grupta toplanabilir. Kodlayan RNA, mesajcı RNA (mRNA) adı verilen, DNA'daki kalıtsal bilgilerden gerekli proteinlerin üretilmesinde görev alan, kalıp olarak kullanılan biyomoleküllerdir. Taşıyıcı RNA (tRNA) ve ribozomal RNA (rRNA), en çok bilinen kodlamayan RNA çeşitlerindendir. tRNA protein sentezi için gerekli aminoasitleri taşımak ile görevli iken rRNA proteinlerin üretildiği yer olan ribozomların yapısında yer alır ve hücrede en fazla bulunan RNA çeşididir.

Transkriptom, bir hücrede, dokuda ya da organizmada, belirli bir zamanda yer alan, genom tarafından üretilen mRNA gibi kodlanmış (RNA'nın %1-4'ünü kodlanan kısım oluşturur) ve rRNA, tRNA gibi kodlanmamış (RNA'nın %95'inden fazlasını oluşturur) RNA moleküllerinin tümüdür. Transkriptomik kavramı ise belli bir zamanda, belli bir organizmada bulunan transkriptlerin tamamının incelenmesini içerir. Transkriptomik alanında transkriptler hem niteliksel (ilgilenilen birimde hangi transkriptlerin yer aldığı, ekleme ve RNA düzenleme bölgelerinin tanımlanması gibi) hem de niceliksel (her bir transkriptin ne kadarının ifade edildiği/kullanıldığını) olarak incelenir.

RNA, DNA ve protein arasında işlev gören bir ara bir molekül olduğu için transkriptomik alanında yapılan çalışmalar genellikle gen ifadeleri ve proteinler ile ilgilidir. mRNA'lar genomda bulunan belirli genlerle eşleştiği için genotip ve ifade seviyeleri arasında bağlantı kurulabilir. Transkriptlerin incelenmesi ile elde edilen RNA profili hücre ve doku tipleri arasındaki işlevsel farklılıkları, genlerin işlevi, ifade seviyesi ve aralarındaki bağlantı hakkında bilgi verdiği için herhangi bir durum ya da hastalıkta etkili olan genlerin bulunmasına katkı sağlar (3). mRNA ifade seviyelerinin incelenmesinde kullanılan, 1990'lı yıllarda ortaya çıkan DNA mikrodizi teknolojisi ve daha güncel bir teknoloji olan RNA dizileme gibi yeni nesil dizileme tekonolojileri ifade profillerinin çıkartılmasını sağlar. Transkriptomik, ilgilenilen organizmaya ait transkript türlerinin tamamının tanımlanmasını, transkripsiyon yapılarını belirlemeyi, hastalık ya da büyüme gibi farkı durumlara bağlı olarak değişen ifade seviyelerini ölçmeyi hedeflemektedir (6). Yapılan çalışmalar genomun %80'inin transkript edilirken, yani RNA molekülü oluşturmak için kopyalanırken, yaklaşık olarak %3'ünün proteinleri kodladığını göstermiştir (2). Transkriptomik çalışmaları sayesinde RNA'nın kodlayan kısmına ilişkin, önceden bilinene göre daha karmaşık bir yapı olduğu gösterilip yeni izoformların tanımlanması sağlanmıştır. Aynı zamanda RNA'nın kodlamayan kısımlarının da incelenmesini sağladığı için kodlayıcı olmayan RNA dizilerinin de önemli pek çok görevde yer aldığını ortaya çıkarmıştır. Ishii ve ark. (7) kodlanmayan RNA bölgelerinin miyokard enfarktüsü, Gupta ve ark. (8), kanser, Moran ve ark. (9) diyabet, Knoll ve ark. (10) endokrin sistemi ile ilişkilerini ortaya koyan çalışmalar yapmıştır.

Translasyon, transkripsiyon ile RNA'ya kopyalanan genetik bilginin protein molekülüne dönüşmesi sürecini ifade etmektedir. Transkriptomik çalışmalar, tüm transkriptomların analizine olanak sağlar, ancak translasyon sonrası meydana gelebilecek değişikliler de protein ifadesini etkileyebilir. Bu nedenle, protein seviyesi ile ilgili çalışmalarda transkriptomları incelemek yeterli olmayabilir.

2.1.3. Proteomikler (Proteomics)

Belirli bir anda, bir hücrede ya da dokuda bulunan proteinlerin tümü proteom olarak adlandırılmaktadır. Proteomik ise proteomu tümüyle inceleyerek proteinlerin yapılarını, yerlerini, miktarlarını, işlevlerini, diğer protein ve moleküllerle ilişkilerini inceler. Belirli bir hücre, doku ya da organizmanın proteomu ile ilgili araştırma yapılacağı zaman, proteomun dinamik bir yapıda olduğu unutulmamalıdır. Proteom dizilendiği andaki ortamı yansıtmaktadır. Protein sentezi için, DNA ve mRNA'nın dört nükleotitli yapıları, daha karmaşık yapıda olan 20 çeşit amino asitten gerekli olanlara dönüşür. Farklı proteinlerin oluşması yalnızca ilgili genlerden kopyalanan dizilerle ilgili değildir. Translasyon sonrası yaşanan değişikliklerin etkisiyle de protein yapısı farklılaşabilir, hücre içinde yer değiştirebilir, sentezlenebilir, bozulabilir. Tüm olasılıklar değerlendirildiğinde, herhangi bir genomun sonsuz sayıda proteoma neden olabileceği görülmektedir (13). Bu nedenle proteomların incelenmesi, genom ve transkriptlere kıyasla daha karmaşıktır.

2.1.4. Metabolomikler (Metabolomics)

Hücresel düzenleme sürecinin son ürünü olan yağ asitleri, amino asitler, karbonhidratlar, vitaminler ve hormonlar gibi küçük moleküller metabolit; belirli bir anda, belirli bir yerdeki metabolitlerin tümü ise metabolom olarak adlandırılmaktadır. Eşzamanlı olarak metabolitlerin tanımlanması, miktarlarının belirlenmesi metabolomiğin alanına girer. Ölçülen metabolitlerin seviyeleri ya da göreceli oranları, metabolik işlev hakkında bilgi verir. Normal aralığın dışında değerlerin olması genellikle hastalık ile ilişkilendirilir (2). Metabolitlerin belirlenmesi ve analizi için kullanılan tekniklerin şu an için geliştirilmeye ihtiyaç duyması ve tespit edilebilir metabolitlerin tamamının biyolojik rolünün hala tam anlaşılamamış olması, metabolomik çalışmaların kısıtları arasında sayılmaktadır (4).

2.1.5. Transktriptom Verisinin Elde Edilmesi

Transkriptom verisine ulaşmak ve analiz etmek için melezleme (hibridizasyon) ve yeni nesil dizileme bazlı yaklaşımlar başta olmak üzere çeşitli teknolojiler geliştirilmiştir. Melezleme yöntemine dayalı mikrodiziler, 1990'lı yıllardan beri kullanılan, kullanışlı yöntemler olup yerini daha yeni ve daha pahalı olan yeni nesil dizileme yöntemlerine bırakmaktadır.

2.1.5.1. Mikrodizi (Microarray)

Mikrodiziler, aynı anda binlerce genin incelenmesini sağlayan araçlardır. Belirli DNA dizilerinin cam, naylon veya silikon gibi katı bir yüzeye tutturulması ile oluşturulurlar (3). Gen ifadelerinin ölçülmesi ile ifade profillerinin oluşturulması, gen ifadelerindeki değişikliklerin saptanmasının yanı sıra DNA'da oluşan mutasyonların belirlenmesi amacıyla da kullanılmaktadır. Mikrodizi analizi ile elde edilen gen ifadesi ölçümleri, referans bir örnek ile karşılaştırılarak, hücre tiplerinde, dokularda, gelişim ya da hastalık durumlarında nasıl değişiklik gösterdiği ile ilgili önemli bilgiler verir. Hastalık fenotipinin araştırıldığı transktiptomik çalışmalarda, mikrodizi analizinin aşamaları aşağıdaki gibi sıralanabilir:

1. Araştırılan hastalığı taşıyan ve taşımayan biyolojik örneklerdeki tüm mRNA molekülleri çıkartılır.
2. Çıkartılan tek iplikli yapıdaki mRNA'lardan, ters transkriptaz enzimi kullanılarak tek iplikli, kodlanmayan bölgeleri çıkartılmış, tamamlayıcı DNA (cDNA) molekülleri üretilir.
3. Bu moleküller, çalışmanın türüne göre bir veya iki renk kullanılarak, floresan boya ile işaretlenir.
4. Örnek diziler ve onlara karşılık gelen referans diziler melezleştirilerek çift iplikli cDNA'lar elde edilir.
5. Transkriptler floresan boya ile işaretlendiği için ışık yoğunluğu, gen ifadesinin ölçümü olarak değerlendirilir. Elde edilen değerler mutlak değil, göreceli bolluk değerleridir (11). Elde edilen sayısal değerler analiz edilerek anlamlı olarak farklı ifade edilmiş genler belirlenir.

Mikrodiziler transkriptom analizini mümkün hale getiren kullanışlı araçlar olarak, önemli bilgilerin kazanılmasını sağlamıştır. Mikrodizilerden sonra ortaya

çıkan, RNA dizileme gibi yeni nesil dizileme teknolojileri, mikrodizilere göre daha maliyetli araçlar olsalar da daha detaylı bilgilere ulaşılabilmesi, genom hakkında önceden elde edilmiş bilgilere dayanmadıkları için araştırmacılar mikrodiziler yerine yeni nesil dizileme teknolojilerine giderek daha fazla yönelmektedirler (4).

2.1.5.2. RNA Dizileme (RNA Sequence)

Son yıllarda yüksek verimliliğe sahip yeni nesil dizileme teknolojileri, biyolojik çalışmalarda yaygın olarak kullanılmaya başlanmıştır. İlgilenilen hücre ya da dokudaki transkriptlerin eş zamanlı olarak, hızlı ve detaylı incelenmesine olanak sağlayan RNA dizileme teknolojisi, mikrodizilerde olduğu gibi gen ifade seviyelerinin farklı durumlarda nasıl değiştiğini belirlemek için kullanılmaktadır. İnsan genomunda yaklaşık olarak 20.000 gen kodlayan 3 milyar DNA baz çifti bulunmaktadır. Kodlama bölgeleri genomun tamamının %1-2'sini oluştururken kalan % 98–99'u kodlama yapmayan bölgelerden oluşur (4). Yalnızca kodlayan değil, kodlamayan transkriptlerin de varlığı ve miktarlarının saptanması ile transkriptom profilindeki değişikliklerin hastalık, büyüme gibi durumlar üzerindeki etkileri araştırılabilir. RNA dizileme yönteminin aşamaları aşağıdaki gibi özetlenebilir:

1. RNA dizileme yöntemi, belirli hücre ya da dokulardaki tüm RNA moleküllerinin çıkartılması ile başlar.
2. Ardından, araştırmanın amacına göre, kodlayan veya kodlamayan RNA çeşitlerinden araştırılmak istenen seçilir ve ayrılır. Bu işlem RNA zenginleştirmesi olarak da ifade edilmektedir. Örneğin, amaç kodlayan RNA olan mRNA'ların incelenmesi ise, mRNA tüm RNA'nın çok küçük bir bölümünü kapladığı için diğer RNA türleri çıkartılarak, toplam RNA içerisinde mRNA'nın oranı arttırılmış olur.
3. İlgilenilen RNA türünden cDNA üretilir. RNA dizileme yönteminde, mikrodizilerin aksine doğrudan cDNA dizisi belirlenir (6).
4. Elde edilen cDNA'lar referans diziye göre sıralanır, okuma uzunlukları da aynı şekilde sıralanarak gen ifadeleri elde edilmiş olur.

RNA dizileme, mikrodiziye göre birkaç önemli avantaja sahiptir. Bunlardan ilki, RNA dizilemenin yalnızca bulunan transkriptleri tespit etmekle sınırlı olmamasıdır. Henüz belirlenememiş genomik dizilere sahip organizmalar için RNA dizileme yöntemi kullanılabilir. Mikrodizilerin aksine kodlanmayan bölgeleri de

inceleme fırsatı verir. RNA dizilemede arka plan sinyali veya gürültüsü, mikrodiziye göre daha düşüktür. Bu durum dizilerin daha yüksek bir doğrulukla sıralandığını göstermektedir (6)

RNA dizileme ile elde edilen ifade seviyelerinin üst sınırı yoktur. RNA dizileme, transktiptlerin algılanabileceği çok daha geniş bir aralık sunmaktadır. Örneğin, Nagalakshmi ve arkadaşlarının (12) yaptığı bir çalışmada transkript ifadelerinin 8000 katı içeren bir aralığı olduğu saptanmıştır. Mikrodiziler, çok düşük veya çok yüksek düzeylerde ifade edilen genlerde bu kadar duyarlılığa sahip değillerdir (6).

2.1.5.3. Farklı İfade Edilmiş Genlerin Bulunması Analizi

Farklı özelliklere sahip örneklerden elde edilen genlerin ifade seviyeleri arasında fark olduğu gözleniyorsa ve gözlenen bu fark istatistiksel olarak da anlamlı ise ilgili genler farklı ifade edilmiş genler olarak adlandırılır. Kullanılacak istatistiksel yöntemler mikrodizilerde sayısal yoğunluk değerlerine dayanmaktadır. RNA dizileme verilerinde ise okuma sayılarının dağılımları temel alınarak analiz edilir (14). Kesikli sayısal değişken olan okuma sayılarının dağılımı genellikle Poisson ya da Negatif Binom dağılımlarından biri ile modellenmektedir. Dağılımlar ile ilgili bilgi vermeden önce, teknik kopya ve biyolojik kopya kavramlarını açıklamak iyi olacaktır. Teknik kopya, aynı örnekten alınan dizileri farklı koşullarda, birden fazla kez analiz etmeyi içerir. Genellikle, yapılan analizlerin tekrarlı ölçümleri elde edilerek kullanılan araçların tutarlılığını incelemek amaçlanmaktadır. Biyolojik kopyalar ise çalışmanın konusuna uygun olarak, farklı biyolojik örneklerin ifade seviyelerinin ölçülmesi, örnekler arasındaki biyolojik farklılığın incelenmesi amacıyla alınmaktadır (15).

Dizileme deneylerinden elde edilen genlerin, belirli bir gen havuzundan rastgele seçilen genler olduğu düşünülebilir. Okuma sayıları n, beklenen okuma sayısı λ ile gösterilirse, okuma sayılarının dağılımı, $f(n,\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$ şeklinde Poisson dağılımı gösterir (16). Poisson dağılımında ortalama ve varyans eşit olup λ'ya eşittir. Teknik kopyaların yer aldığı dizileme çalışmalarında, aynı örnek üzerinden analizler yapıldığı için varyasyonun çok yüksek olması beklenmez. Bu nedenle Poisson dağılım, varyans kısıtı olsa da kullanılabilir. Ancak biyolojik kopyaların kullanıldığı birçok çalışmada varyansın ortalamadan daha yüksek çıktığı görülmüştür. Bu sorun

aşırı yayılım (overdispersion) olarak da ifade edilmektedir. Örnekler arası biyolojik varyasyonun yüksek olduğu düşünülen durumlarda Poisson dağılımı yeterli olmayabilir. Böyle durumlarda Poisson dağılımı yerine, varyansın geniş bir aralıkta değer alabildiği Negatif Binom dağılımı tercih edilebilir. Negatif Binom dağılımında varyans ve ortalama ilişkisi, dağılım parametresine ek bir yayılım parametresi (α) eklenerek şöyle gösterilebilir: $\sigma^2 = \mu + \alpha\mu^2$. Farklı ifade edilmiş genlerin bulunmasında ilk olarak, her bir gen için uygun dağılım parametreleri kestirilir (17). Her gen için farklı örneklerdeki ifade seviyeleri karşılaştırılarak farklı ifade edilmiş genler bulunur. Genellikle olabilirlik oranı (likelihood ratio) ya da Walt testi sonuçlarına göre, belirlenen 1. tip hata düzeyinden düşük p değerine sahip genler farklı ifade edilmiş genler olarak kabul edilir. Farklı ifade edilmiş genlerin bulunması için çok sayıda yöntem ve yazılım bulunmaktadır. Bu alanda, R programında yer alan DESeq (19), DESeq2 (17), EdgeR (18), limma (20) paketleri sıkça kullanılmaktadır. Paketlerdeki yazılımlar genellikle aynı yaklaşıma sahiptir. Regresyona dayalı modeller kurarak her gen için örnekler arası ifade farkını tahmin ederler. Daha sonra bu farkın sıfıra eşit olduğu yokluk hipotezini test eden istatistiksel bir test ile sonuca varılır (21).

Omik türü veriler, yüksek boyutlu veriler olarak değerlendirilir. Karşılaştırılan örneklerde farklı olarak ifade edilen genler bulunduktan sonra çeşitli filtreleme yöntemleri kullanılabilir. Çalışmaya dahil edilecek genlerin seçiminde temel bileşenler analizi ve bağımsız bileşen analizi, boyut azaltma amaçları ile kullanılabilmektedir. Gen ifade seviyeleri bakımından en yüksek toplam, ortalama veya varyansa sahip belirli sayıdaki genlerin seçimi de özellik seçiminde kullanılan yöntemlerdendir. Yanlış bulgu oranı (false discovery rate, fdr) yaklaşımı, RNA dizileme verilerinde filtre olarak sıkça kullanılmaktadır. Yanlış pozitif, gerçekte örnekler arasında ilgili gen ifadesi bakımından fark yokken, fark olduğunun kabul edilmesidir. Yanlış bulgu oranına göre genleri filtrelemek, yanlış pozitif oranını kontrol altına almayı sağlar.

Örneklerde gen ifadelerinin farklı olup olmadığı istatistiksel bir testle analiz edildikten sonra, yanlış bulgu oranı, gerçekte aralarında fark olan ve test sonucu farklı kabul edilen genlerin sayısının, farklı kabul edilen tüm genlerin sayısına oranı alınarak bulunur. Filtreleme için yanlış bulgu oranı yaklaşımında, p değerleri kullanılarak her gen için q değeri hesaplanır. Toplam gen sayısı m olarak gösterilecek olursa, m adet

gen test sonucu hesaplanan p değerine göre küçükten büyüğe doğru sıralanır. Sıra r, her sıraya karşılık gen için p değeri p_r ile gösterilecek olursa, q_r değeri,

$$q_r = m * \left(\frac{p_r}{r}\right) \quad (2.1.)$$

formülü ile bulunur. Elde edilen q değerleri de p değerlerinin yanına eklenir. Ardından belirlenen yanlış bulgu oranı, q değerleri için önem düzeyi olarak kullanılır (22). Örneğin, yanlış bulgu oranı 0,05 olarak belirlenirse, q değeri 0,05'ten küçük olan genler seçilerek filtreleme yapılmış olacaktır.

2.1.5.4. Normalleştirme

Transkriptom verilerinde normalleştirme, sıralama derinliği gibi teknik işlemlerden, gen uzunluğu, guanin-sitozin içeriği gibi örneklerden kaynaklanabilecek yanlılıkları en aza indirme amacı ile uygulanmaktadır. Transkriptom verilerinde gerekli olan okuma sayısı, ilgilenilen en az miktardaki RNA türü tarafından belirlenir. İfade düzeyi düşük olan genleri tanımlamak, farklı durumlar arasındaki çok küçük miktardaki kat değişikliklerini belirlemek, yeni transkript tespit etmek gibi amaçlarla, verinin sıralama derinliği, yani dizileri okuma sayısı arttırılabilir. Bu durum ifade seviyesi yüksek olan genlerin daha yüksek algılanması gibi bir yanlılığa neden olabilir. Sıralama derinliğine karar verilirken referans olabilecek kılavuzlar, deneyin türü ve amacı, transkriptomun büyüklüğü ve karmaşıklığı göz önünde bulundurulmalıdır (21). Başka bir varyasyon kaynağı olan gen uzunluğu da ölçüm sonuçlarına yanlılık katabilmektedir. Daha uzun olan genler, boyutlarındaki farklılıktan dolayı, daha kısa diziye sahip genlere göre daha yüksek okuma sayılarına, diğer bir deyişle ifade seviyelerine sahip olabilirler (25). Risso ve arkadaşları (26), guanin-sitozin içeriğinin de yanlılığa neden olabileceğini belirtmişlerdir. Buna göre guanin-sitozin oranı yüksek genlerin farklı ifade edilmiş genler olarak bulunma olasılıkları yüksektir. Normalleştirme, bu gibi yanlılık faktörlerinin etkisini azaltarak, örneklerin karşılaştırılabilir olmasını amaçlamaktadır. Transkriptom verilerinin normalleştirilmesine yönelik pek çok yöntem geliştirilmiştir. Yöntemler genellikle verilerin hesaplanan bir normalleştirme faktörüne göre ölçeklendirilmesini temel alır. Sık kullanılan bazı yöntemler şunlardır:

Toplam Sayı (Total Count) Normalleştirmesi: İlgili örneğe ilişkin okuma sayıları, tüm okuma sayılarına bölünerek normalleştirme faktörü hesaplanır. Transktiptom verisinde yer alan toplam gen sayısı G, toplam örnek sayısı m, ilgilenilen gen g, ilgilenilen örnek j, okuma sayısı (gen ifadesi) K, normalleştirme faktörü d_j^{TC} ile gösterilecek olursa, bu yöntem için normalleştirme faktörü aşağıdaki gibi hesaplanır:

$$d_j^{TC} = \frac{\sum_{g=1}^{G} K_{gj}}{\sum_{g=1}^{G}\sum_{j=1}^{m} K_{gj}} \qquad (2.2.)$$

Toplam sayı üzerinden normalleştirme yapılması ifade düzeyi yüksek olan genler nedeniyle yanlılık oluştuğundan ve farklı örnekler arası gen listesinin de farklı olabileceğini hesaba katmadığı yönleriyle eleştirilmektedir (21).

Üst Çeyreklik (Upper Quartile) Normalleştirmesi: Sıfır değerine sahip transkriptler veri setinden çıkartılarak, kalan değerlerin 75. yüzdelik değerleri üzerinden normalleştirme yapılır (24).

$$d_j^{UQ} = UQ\left(\frac{K_{gj}}{\sum_{g=1}^{G} K_{gj}}\right) \qquad (2.3.)$$

Bu yöntem de toplam sayı normalleştirmesinde olduğu gibi ifade seviyesi yüksek genlerden etkilenmektedir.

M Değerlerinin Kırpılmış Ortalaması (Trimmed Mean of M Values) Normalleştirmesi: Bu yöntemde normalleştirme faktörü, önceden hesaplanan, ilgilenilen örnek j ile referans alınan bir örneğe (r) ifade seviyelerinin logaritmik oranları olan M değerleri ve ağırlık (w) değerleri yardımıyla elde edilir. Veri setinde yer alan genlerin çoğunun farklı ifade edilmiş genler olmadığı hipotezini temel alır (24). N_j ilgilenilen örneğe ilişkin toplam okuma sayısı, N_r referans örneğe ilişkin toplam okuma sayısı, A genlere ilişkin mutlak ifade seviyeleri ve $K_{gj}, K_{gr} > 0$ olmak üzere $M,\ A$ ve w değerleri aşağıdaki formüller kullanılarak hesaplanır:

$$M_{gj} = log_2((\frac{K_{gj}}{N_j})/(\frac{K_{gr}}{N_r})) \qquad (2.4.)$$

$$A_{gj} = \frac{1}{2} log_2((\frac{K_{gj}}{N_j}) * (\frac{K_{gr}}{N_r})) \quad (2.5.)$$

Ağırlık değerleri hesaplanmadan önce veriler M ve A değerlerine göre belirli oranlarda kırpılırlar (27). Bu sayede çok düşük veya yüksek ifade seviyelerine sahip genler veri setinden çıkartılmış olur.

$$w_{gj} = \frac{N_j - K_{gj}}{N_j * K_{gj}} + \frac{N_r - K_{gr}}{N_r * K_{gr}} \quad (2.6.)$$

Kırpılma işlemi yapılmamış genler G' ile gösterilsin, bu durumda normalleştirme faktörü aşağıdaki gibi hesaplanır:

$$log_2(d_j^{TMM}) = \frac{\sum_{g=1}^{G'} w_{gj} M_{gj}}{\sum_{g=1}^{G'} w_{gj}} \quad (2.7.)$$

Üst çeyreklik ve M değerlerinin kırpılmış ortalaması normalleştirme yöntemleri, R programında yer alan edgeR paketinde yer almaktadır (18).

Ortanca Oranı Normalleştirmesi (Median Ratio Normalization): Bu yöntemde TMM yönteminde olduğu gibi genlerin çoğunun farklı ifade edilen genler olmadığı hipotezine dayanır (24). Her gen için tüm örneklerdeki geometrik ortalama alınarak referans örneği oluşturulur. Ardından yine her gen için ilgilenilen örneğin referans örneğe oranı hesaplanır. Son olarak oranların ortancası alınarak ilgili örnek için normalleştirme faktörü hesaplanmış olur. R programında yer alan DESeq2 paketi bu yöntemi kullanmaktadır (17).

$$d_j^{MED} = median_g \frac{K_{gj}}{\left(\prod_{v=1}^{m} K_{gv}\right)^{1/m}} \quad (2.8.)$$

Walczak ve arkadaşları (24) üç adet RNA dizileme verisinde farklı normalleştirme yöntemlerini uygulamışlar ve bu yöntemleri yanlılık, varyans, duyarlılık ve seçicilik gibi bazı kriterlere göre sıralamışlardır. Medyan oranı normalleştirmesi çoğu kriterde diğer yöntemlerin önüne geçmiştir.

2.1.5.5. Dönüşüm

Özellikle hastalık mekanizmasının araştırılmasını amaçlayan çalışmalarda, RNA dizileme ile elde edilen gen ifadesi verilerine sınıflama, kümeleme gibi yöntemler uygulanabilmektedir. Bu yöntemlerin uygulanabilmesi için verinin normalleştirilmiş olması yeterli olmayabilir. Çünkü RNA dizileme verilerinde ifade seviyeleri çok geniş bir aralıkta dağılabilmektedir (12). Uygulanabilecek en kolay dönüşüm logaritmik dönüşümdür. Dönüşüm uygulanacak gen ifadesi x ile gösterilsin. Gen ifadesi sıfıra eşit olabileceğinden, logaritmik dönüşüm genellikle nomalleştirilmiş değerler üzerinden $log_2(x+1)$ şeklinde yapılmaktadır. Logaritmik dönüşüm ile, dönüştürülmemiş veriye göre daha az çarpık bir dağılım ve daha az sayıda aşırı değer içeren bir veri elde edilmektedir. Ancak bu yaklaşım ile ifade seviyesi düşük olan genlere ağırlık verildiği, bu durumun da gen ifadelerinin varyanslarının kontrol edilememesine neden olduğu için farklı yöntemler geliştirilmiştir (17).

Düzenlenmiş Logaritmik Dönüşüm (Regularized Logarithm Transformation): Logaritmik dönüşüm ile düşük ifadeye sahip genlerin yayılmasını önlemek için geliştirilmiş bir yöntemdir. Her bir gen için örnekler arası hesaplanan varyans, düzenlenmiş logaritmik dönüşüm ile dengelenir. Düzenleme, genlere ilişkin logaritmik kat değişiklikleri kullanılarak hesaplanan ek değerler üzerinden yapılır. (29). Dönüşüm sonrası elde edilen j. örneğe ilişkin i. gen x_{ij} ile gösterilsin.

$$x_{ij} = \beta_{i0} + \beta_{ij} \qquad (2.9.)$$

olarak gösterilebilir. β_{i0} normalleştirilmiş değerlere standart log_2 dönüşümü uygulanmış değerleri göstermektedir. β_{ij} logaritmik kat değişiklikleri kullanılarak oluşturulan düzeltme terimidir. Burada her gen için hesaplanan logaritmik kat değişikliğinin ½'si eklenir. Bu şekilde baştaki vasyans değeri ile düzenlenmiş logaritmik dönüşümle hesaplanan son varyans üzerinde düzeltme teriminin etkisi azaltılmış olur (17).

Varyans Dengeleyici Dönüşüm (Variance Stabilization Transformation): Ortalamadan bağımsız varyansa sahip değişkenlerin üretilmesi amacıyla, Anders ve Huber [30] tarafından geliştirilmiştir. Bu yöntemde genlerin ifade seviyelerinin ortalamaları ve varyansları aşağıdaki şekilde modellenmiştir:

$$Var(\mu_j) = \sigma_j^2 = \mu_j + \alpha_i \mu_j^2, \quad \alpha_j = \alpha_0 + \alpha_1/\mu_j \qquad (2.10.)$$

$\alpha_0$ve α_1 dağılım parametreleri olup genelleştirilmiş doğrusal modeller kullanılarak tahmin edilmektedir. Varyans dengeleyici dönüşüm uygulanmış ifade değerleri, modellenmiş olan ortalama ve varyans ilişkisi kullanılarak,

$$x_{ij}^{vst} = \int_0^{x_{ij}} \frac{1}{Var(\mu_j)} d\,\mu_j \qquad (2.11.)$$

formülü ile hesaplanabilir. Bu şekilde dönüşüm uygulanan veride ifadelerin varyansları ortalamadan yaklaşık olarak bağımsızdır.

Düzenlenmiş logaritmik dönüşüm ve varyans dengeleyici dönüşüm R programında, DESeq2 paketi kullanılarak uygulanabilir. Varyans dengeleyici dönüşüm için hazırlanan vst fonksiyonu, düzenlenmiş logaritmik dönüşüm için hazırlanan rlog fonksiyonuna göre daha hızlı işlem yapmaktadır. rlog fonksiyonu küçük veri setlerinde (n<30) ve sıralama derinliğinin geniş aralıklı olduğu durumlarda daha iyi çalışma eğilimde iken vst fonksiyonu orta ve büyük ölçekli veri setlerinde tercih edilmektedir (31).

2.2. Sınıflama Yöntemleri

2.2.1. Rastgele Orman

Rastgele Orman (RO) yöntemi 2000'lerin başında Brieman (32) tarafından *Bagging* yöntemi ile her düğümde rastgele değişken seçimi (33) birleştirilerek geliştirilmiş bir algoritmadır. Sınıflama ve regresyon amaçlarıyla kullanılmaktadır. Karar ağaçlarının genel olarak yüksek varyansa ve düşük yanlılığa sahip oldukları kabul gören bir yargıdır. Buradan karar ağacı modellerinde genellikle doğruluğu yüksek sonuçların elde edildiği ancak aynı veri setinden alınan farklı veri örneklerine ait model sonuçları arasında değişkenliğin oldukça yüksek olduğu söylenebilir. RO yönteminde birden çok karar ağacı kullanılarak, tek karar ağacı kullanılmasında oluşabilecek yanlı ve yüksek değişkenliğin önüne geçilmesi amaçlanmaktadır (32).

RO yöntemi temel olarak Sınıflama ve Regresyon Ağacı (*Classification and Regression Tree* - CART) yöntemini kullanmaktadır. Karar ağaçlarında kullanılan her bir açıklayıcı değişken düğümleri oluşturur. Kullanılacak açıklayıcı değişkenlere karar

verilirken, bilgi kazancı değerleri hesaplanarak hangi değişkenlerin ağacın oluşturulmasında yer alacağına karar verilir. Seçilen değişkenlere göre ağacın dallarının oluşması, düğümlerin bölünmesi gerçekleşir. Bu şekilde, her bölünmede daha homojen alt gruplar oluşur. RO metodunun CART algoritmasına göre farkı, bu bölünme sırasında oluşacak düğümler için iki aşamalı bir randomizasyon prosedürü uygulamasıdır. *Bootstrap* yöntemi kullanılarak oluşturulan randomizasyonun yanı sıra, *bagging* işleminde olduğu gibi tüm değişkenleri kullanarak bir ağaç düğümünü bölmek yerine oluşturulan her ağacın düğümlerinde, değişkenlerin rastgele bir alt kümesinden seçim yaparak bu değişkenleri düğüm için en iyi bölmeyi bulmak amacıyla aday olarak değerlendirir.

Genel olarak RO yöntemi ile yapılacak sınıflama işlemlerinde şu aşamalar uygulanmaktadır:

1. Verilerin eğitimi aşamasında ilk olarak oluşturulacak ağaç sayısı belirlenmelidir. Karar verilen sayıdaki karar ağaçları kullanılacak veri setinden *bootstrap* yöntemi ile oluşturulur. Burada her bir ağacın farklı örnekler üzerinde eğitilmesiyle, tek bir ağacın, belirli bir eğitim verisi kullanmasıyla ortaya çıkacak yüksek varyanslılık sorununu çözmek amaçlanmaktadır.

2. Oluşturulan ağaçların her düğümünde, bölme işlemi için rastgele seçilen m sayıda değişken kullanılır. Genellikle veri setinde yer alan toplam değişken sayısı p, rastgele seçilecek değişken sayısını ise m ile gösterilmekte olup sınıflama işlemleri için m=$\sqrt{\text{p}}$ olarak alınmaktadır.

Buna bağlı olarak eğitim aşamasında rastgele seçilen değişkenler içinde en iyi bilgi veren değişkenler dallara ayırmada kullanılacak değişken olarak seçilir. Dallara ayırmada genellikle kriter olarak Gini indeksi kullanılmaktadır. Gini indeksinin hesaplanacağı değişken için dikkate alınacak her düğüm t, sınıf sayısı Q ile gösterilecek olursa Gini indeksi aşağıdaki formül yardımıyla hesaplanabilir:

$$G(t) = 1 - \sum_{k=1}^{Q} p^2(k|t) \qquad (2.14.)$$

Gini indeksi temeline dayanan değişken önem ölçüsü olarak Gini indeksindeki ortalama düşüş hesaplanır. Gini indeksindeki ortalama düşüş, her bir değişkenin

düğümlerinin homojenliğine ve ortaya çıkacak ağaç yapılarındaki yapraklara ne ölçüde katkıda bulunduğunun bir ölçütüdür. Düğümlerdeki homojenliğin (düşük varyasyonun) yüksek olmasını sağlayan değişkenler Gini indeksinde daha yüksek bir düşüşe sahiptir.

Karar ağacı algoritmalarında, farklı yöntemlerle ağacın bölünmesi gerçekleştirilmektedir. Bu yöntemlerden diğeri de seçilen değişkenlere göre, entropi kullanılarak bilgi kazancı hesaplamaktır. Entropi, 0 ile 1 aralığında değer almakta olup, belirsizliğin bir ölçüsüdür. Entropinin 0 değerini alması tüm örneklerin aynı sınıf değerini aldığını gösterir. n örnek sayısını, p_i olasılık dağılımlarını göstersin, bu durumda entropi aşağıdaki formül kullanılarak hesaplanır:

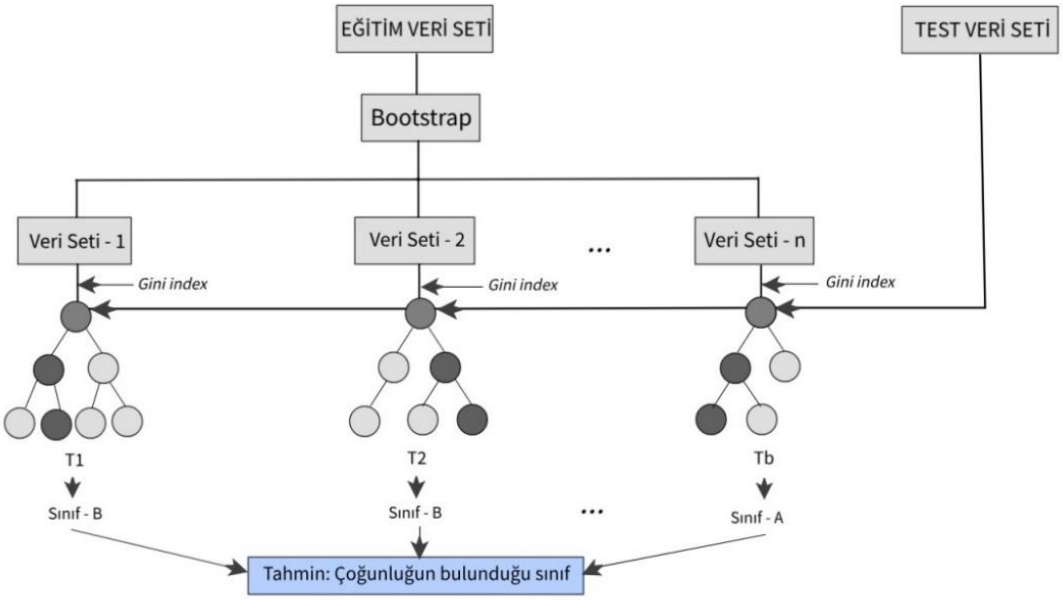
$$Entropi = -\sum_{i=1}^{n} p_i log_2(p_i) \qquad (2.12.)$$

Bilgi kazancı, bir veri setinin seçilen özelliğe göre ayrılmasından sonra elde edilen entropi düşüşü ile ilgilidir. En yüksek bilgi kazancına sahip değişken, dallara ayırmada en homojen dalların oluşmasını sağladığı için tercih edilir. Bölünmede kullanılacak değişken X, bölünme öncesi entropi $Entropi(T)$, bölünme sonrası oluşacak dallar için hesaplanan entropi $Entropi(T,X)$ ile gösterilecek olursa,

$$Bilgi\ Kazancı(X) = Entropi(T) - Entropi(T,X) \qquad (2.13.)$$

formülü ile bilgi kazancı hesaplanır.

3. Daha önce kullanılmamış olan test verisi tahmininde kullanılacak model için, regresyon modelinde oluşturulan ağaçlardan elde edilen tahmin sonuçlarının ortalaması alınırken sınıflama için çoğunluk oyu (majority voting) dikkate alınarak tahminler yapılır. İlgili tahmin için oluşturulan ağaçlarda en yüksek oran hangi sınıfa ait ise o sınıfa atama yapılır. Bootstrap ile oluşturulacak toplam ağaç sayısına B, oluşturulan ağaçların her birine ise T_b diyelim (b=1,..,B). b. RO ağacı için sınıf tahmini $\hat{C}_b(x)$ ile gösterilecek olursa yeni veri için sınıf tahmini $\hat{C}_b(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$ şeklinde yapılmaktadır (Şekil 2.2).



Şekil 2.2. Rastgele Orman algoritması.

2.2.1.1. Avantajları

- Optimize edilmesi gereken parametre (hiperparameter) sayısı çok fazla olmadığından kullanımı kolay bir algoritmadır (34).
- Doğrusal olmayan sınıflama görevini etkin bir şekilde yürütür.
- Makine öğrenimindeki en büyük sorunlardan biri olan aşırı uyum sorunu, modelde yeterince ağaç varsa kolayca çözülebilir bir sorundur.
- Özellikle büyük veri setlerinde, açıklayıcı değişken sayısının fazla olduğu durumlarda ve açıklayıcı değişkenlerdeki eksik gözlemlerle başa çıkabilmektedir (35).
- Her karar ağacında tüm değişkenler yerine rastgele seçilen değişkenler kullanıldığından, bazı değişkenlerde kayıp gözlem olmasından diğer sınıflama yöntemlerine göre daha az etkilenir.

2.2.1.2. Dezavantajları

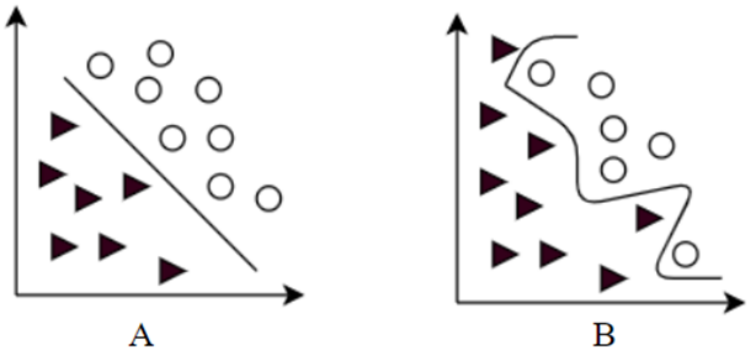
- RO algoritması ağaçları paralel çalıştırılabilme özelliğinden dolayı hızlı çalışabilse de ağaç sayısının çok fazla olduğu durumlarda tahminleri oluşturmak zaman alabilmektedir.

- Yorumlanabilme açısından tek bir karar ağacına göre daha zor yorumlanır (36).
- Farklı sayıda düzeyleri olan kategorik değişkenlerin olduğu veri setlerinde değişken seçimi açısından daha fazla düzeye sahip değişkenler lehine karar verebilir (37). Bu sorunu çözmek için kısmi permütasyon ve yansız ağaçlar gibi yöntemler kullanılmaktadır.

2.2.2. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), Boser, Guyon ve Vapnik tarafından geliştirilen bir sınıflama yöntemidir. Vapnik ve Lerner tarafından, 1963 yılında "Pattern Recognition Using Generalized Portrait Method" isimli makale ile yöntem tanıtılmıştır. DVM başlangıçta ikili sınıflara ayırabilen bir hiper düzlemin belirlenmesi için geliştirilen bir yöntem olsa da kullanımı çok sınıflı ve doğrusal ayrılamayan modeller için de genişletilmiş ve kullanılmaya başlanmıştır. Sınıf sayısının ikiden fazla olduğu veri setlerinde birine karşı biri (*one versus one*) ya da birine karşı hepsi (*one versus all*) yöntemleri kullanılarak çok sınıflı veriler sınıflandırılmaktadır (38). Yüksek boyutlu verilerle başa çıkma kabiliyeti ve esnek yapısı nedeniyle genetik verilerde ve diğer disiplinlerde yaygın olarak kullanılmaktadır.

DVM çekirdek (*kernel*) yöntemini temel olarak aldığından veriler noktalar aracılığıyla temsil edilir. Modelde yer alan bir nokta, yüksek boyutlu özellik alanındaki bir çekirdek fonksiyonu kullanılarak yer değiştirebilir. Bu sayede doğrusal olarak ayrılabilen modeller için tasarlanan yöntemler kullanılarak doğrusal ayrılamayan modeller için karar sınırları oluşturulabilir (Şekil 2.3).



Şekil 2.3. Doğrusal olarak ayrılabilen (A) ve ayrılamayan (B) veriler.

2.2.2.1. Doğrusal Sınıflama

Sınıf etiketinin y ile gösterildiğini, -1 ve 1 değerlerinden birini aldığını varsayalım. x_i veri setindeki n boyutundaki sınıflara ilişkin özelliklerin yer aldığı giriş vektörünü ($\{(x_i, y_i)\}_{i=1}^{n}$) göstersin. DVM'de hataların karesini minimuma indirebilecek ayırıcının belirlenmesi hedeflenir. İki sınıflı ve doğrusal olarak ayrılabilen bir veri için öncelikle en düşük hataya sahip paralel iki vektör seçilir, vektörler ile ayrılan bu düzlemler arasındaki uzaklığın maksimum olması istenir. Ardından bu iki vektörün ortasında yer alan bir nokta belirlenir. Bu nokta $w^t x = \sum_{i=1}^{n} w_i x_i$ olarak tanımlanır. Doğrusal bir sınıflandırıcı, doğrusal bir ayırıcı fonksiyonu temel alır. Burada $f(x) = w^t x + b$ fonksiyonu kullanılarak doğrusal ayırma işlemi gerçekleştirilir. Formülde yer alan w ağırlık vektörü iken b yanlılık olarak da adlandırılan hata terimidir. b arttıkça hiper düzlem başlangıç noktasından uzaklaşmaktadır. Veri setine eklenen yeni x örneği için ayırıcı fonksiyon yardımıyla sınıf atama işlemi gerçekleştirilir. Bu yöntemde sınıflama,

$$w^t x + b \geq 0 \text{ ise } y_i = 1 \qquad (2.13.)$$

$$w^t x + < 0 \text{ ise } y_i = -1 \qquad (2.14.)$$

şeklinde gerçekleştirilir.

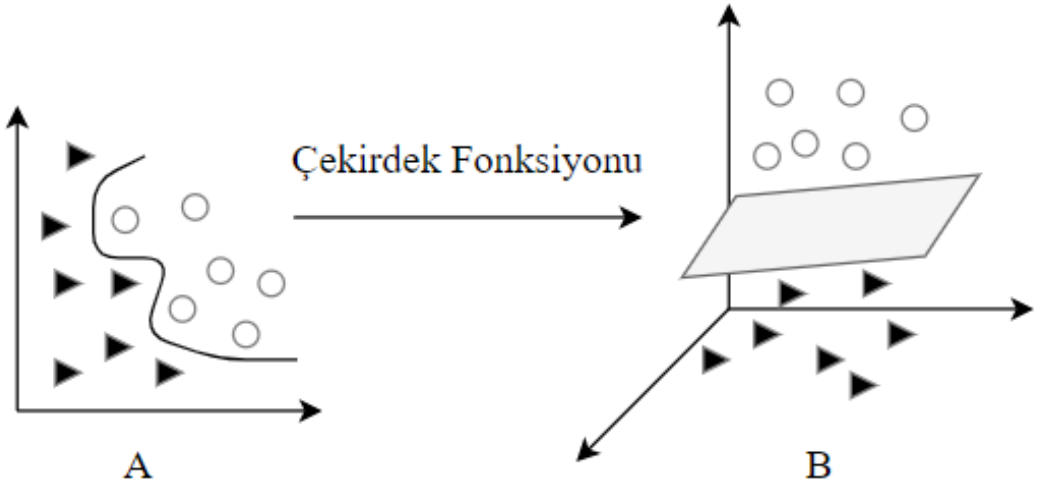
2.2.2.2. Doğrusal Olmayan Sınıflama:

Sınıflama sürecinde karşılaşılan zorluklardan biri farklı eğilimler gösteren verilerin dağılışı nedeniyle verilerin doğrusal olarak ayrılamamasıdır. Böyle durumlarda, DVM çekirdek fonksiyonlar kullanarak orijinal veri uzayını daha yüksek boyuta sahip yeni bir uzaya dönüştürerek verilerin kolay ayrılabilir bir yapıya sahip olmasını sağlar. Doğrusal olmayan durumlar için hiper düzlem fonksiyonu genel olarak şu şekilde yazılmaktadır:

$$f(x) = \alpha_i y_i K(x_i, x_j) + b \qquad (2.15.)$$

α Lagrange çarpanı iken çekirdek fonksiyonu $K(x_i, x_j)$ olarak gösterilmektedir (Şekil 2.4). Başlıca kullanılan dört çekirdek fonksiyonuna ilişkin formüller aşağıda

Tablo 2.1'de yer almaktadır. Her çekirdek fonksiyonun kendine ait parametreleri bulunmaktadır.



Şekil 2.4. Doğrusal olarak ayrılamayan (A) ve çekirdek fonksiyonu ile ayrılabilir hale getirilen (B) veriler.

Sigmoid ve RTF yapay sinir ağları temeline dayanmaktadır. Sigmoid fonksiyonun iki katmanlı perseptron ile eşdeğer olduğu Lin ve arkadaşları belirtmiştir (39). RTF, diğer çekirdek fonksiyonlarına göre daha hızlı öğrenme özelliği ile ön plana çıkmaktadır. RTF, doğrusal çekirdek fonksyionunun aksine doğrusal olmayan ilişkilerle de başa çıkabilmektedir. RTF fonksiyonu ile kurulan modellerde, polinomial çekirdek fonsiyonuna göre ayarlanması gereken parametre sayısı daha azdır, model karmaşıklığı açısından daha sade olduğu söylenebilir (40).

Tablo 2.1. Yaygın olarak kullanılan çekirdek fonksiyonları.

Çekirdek Fonksiyonu	**Formül**
Doğrusal Fonksiyon	$K(x_i, x_j) = x_i^t x_j$
Polinomial Fonksiyon	$K(x_i, x_j) = (\alpha x_i^t x_j + b)^d$
Sigmoid Fonksiyon	$K(x_i, x_j) = \tanh(\sigma x_i^t x_j + r)$
Radyal Tabanlı Fonksiyon (RTF)	$K(x_i, x_j) = e^{(-\frac{\|x_i - x_j\|^2}{\sigma^2})}$

2.2.2.3. Avantajları

- Sınıflama performansı yüksektir, aşırı uyum sorunu fazla yaşanmaz (41).
- Çekirdek fonksiyonları sayesinde veriler hakkında pek bilgi yokken de iyi sonuçlar verir.
- Çekirdek fonksiyonu belirlendikten sonra sinir ağları gibi karmaşık modellere göre optimize edilmesi gereken daha az sayıda parametre olması nedeniyle kullanımı/uygulaması daha kolaydır.

2.2.2.4. Dezavantajları

- Veriye uygun çekirdek fonksiyonu ve parametreleri seçmek her zaman kolay değildir (41).
- Son model ve değişkenlerin ağırlıkları ile ilgili yorum yapmak zordur.

2.2.3. Yapay Sinir Ağları

Yapay sinir ağları (YSA) sınıflama ve örüntü tanımlamayı sağlayan, biyolojik sinir ağlarının yapısından esinlenilerek geliştirilmiş bir yöntemdir. Biyolojik bir sinir hücresi soma, akson, dentrit ve sinapslardan oluşur. Yapay sinir ağlarında kullanılan terimler ile bunların biyolojik sinir ağlarınındaki karşılıklıları Tablo 2.2'de sunulmuştur.

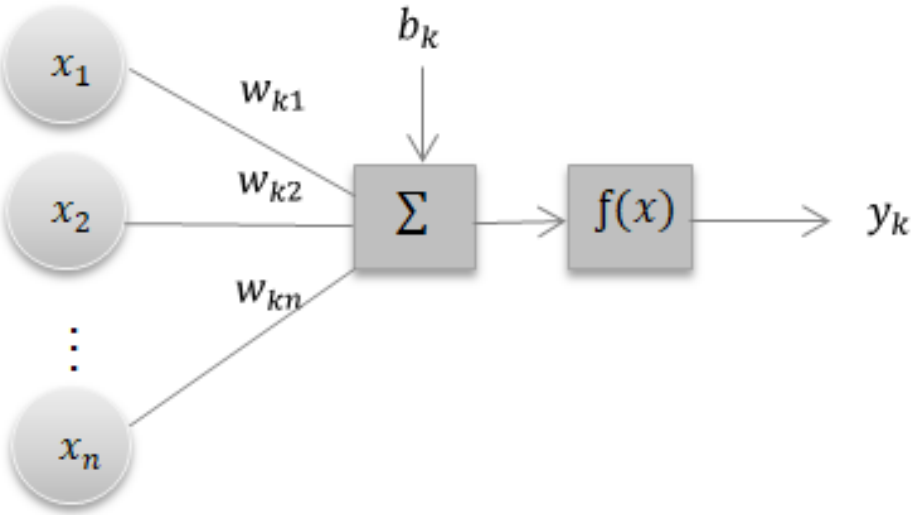
Tablo 2.2. Biyolojik sinir ağı ile yapay sinir ağı benzerliği.

Biyolojik Sinir Ağı	**Görevi**	**Yapay Sinir Ağı**
Soma	*Sinir hücresinin gövde kısmı olup çekirdeği ve diğer kimyasal yapıları içerir.*	Nöron
Dentrit	*Nörona diğer nöronlardan gelen elektrokimyasal sinyalin somaya ulaştırılmasını sağlayan yapıdır.*	Girdi
Akson	*Sinyali nörondan diğer nöronlara ileten sinir hücresinin uzantısıdır*	Çıktı
Sinaps	*İki nöronun dentriti ya da nöronun sinir hücresi olmayan kas, salgı bezi gibi hücrelere bağlanmasını sağlayan özelleşmiş bağlantı noktalarıdır.*	Ağırlık

Biyolojik sinir ağında nöron bilgisi dentrit yoluyla ilgili nörona getirilir, soma kısmında bilgi işlenir, akson üzerinden iletilir. YSA'da benzer olarak bilgi yapay bir

nörona ağırlıklandırılmış girdiler olarak gelir, bu girdiler toplanır, oluşan yanlılık dikkate alınarak aktivasyon fonksiyonuna göre hesaplamalar yapılır (Şekil 2.5). Bu işlemlerin sonucunda işlenmiş olan bilgi çıktılar üzerinden iletilir. YSA'ya ait model aşağıdaki şekilde yazılabilir:

$$y_k = f(\sum_{i=0}^{n} w_{ki} x_i + b_k) \tag{2.16.}$$



Şekil 2.5. Yapay sinir ağı yapısı.

Yaygın olarak kullanılan bazı aktivasyon fonksiyonları şu şekildedir:

Sigmoid Fonksiyonu:

$$\mathrm{f(x)} = \frac{1}{1 + \exp(-\mathrm{x})} \tag{2.17.}$$

Doğrusal olmayan, 0 ve 1 değerleri arasında çıktı üreten bir aktivasyon fonksiyonudur. Uygulanması ve anlaşılması kolay olduğu için sıkça kullanılmaktadır. Sigmoid aktivasyon fonksiyonu gizli katmanlardan giriş katmanlarına geri yayılma sırasında keskin gradyan değerlerine sahip olma, sıfır merkezli olmayan çıktılar üretme gibi dezavantajlara sahiptir. Bu dezavantajlar, eğitim sırasında gradyan değerinin güncellenmesinde farklı yönlere gidilebilmesine neden olmaktadır. Yaşanılan bu dezavantajların bir kısmının çözülmesi için Hiperbolik Tanjant Fonksiyonu (Tanh) gibi farklı aktivasyon fonksiyonları önerilmiştir.

Hiperbolik Tanjant Fonksiyonu:

$$\mathrm{f(x)} = \tanh(\mathrm{x}) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad (2.18.)$$

Sıfır merkezli, −1 ile +1 aralığında çıktı üreten bir fonksiyondur. Sigmoid aktivasyon fonksiyonu ile birçok yönden benzese de çıktı alanının daha geniş olması nedeniyle çok katmanlı sinir ağlarını modellemede Sigmoid fonksiyonuna göre avantaj sağlamaktadır, ancak Sigmoid fonksiyonunda görünen gradyan sorunu bu fonksiyon için de geçerlidir. Tanh fonksiyonunun temel avantajı sıfır merkezli çıktı üreterek geri yayılma işlemine yardım etmesidir.

Softmax Fonksiyonu:

$$\mathrm{f(x)} = \frac{\exp(\mathrm{x})}{\sum_{\mathrm{i=1}}^{\mathrm{k}} \exp(\mathrm{x})} \qquad (2.19.)$$

Olasılıkların toplamı 1 olacak şeklide çıktıların 0 ile 1 arasında değer aldığı bir fonksiyon türüdür. Sigmoid fonksiyonu ile arasındaki temel fark Sigmoid fonksiyonunun ikili sınıflama, Softmax fonksiyonunun ise çok sınıflı sınıflama için kullanılmasıdır.

Düzleştirilmiş Doğrusal Birim Fonksiyonu (Rectified Linear Unit -ReLU)

$$f(x) = \max(0, x) = \begin{cases} x_i, & x_i \geq 0 \\ 0, & x_i < 0 \end{cases} \qquad (2.20.)$$

Derin öğrenme uygulamaları için yaygın olarak kullanılmaktadır. Yalnızca ilgili nöron çıktısı pozitifse aktivasyona izin verir. Bu şekilde diğer aktivasyon fonksiyonlarına göre daha hızlı bir şekilde hesaplamalar yapılır. Doğrusal bir fonksiyona yakın bir işlev görmesi modelin optimizasyonu için avantaj sağlar. ReLU'nun bir dezavantajı Sigmoid fonksiyonuna göre daha kolay olarak aşırı uyum gösterebilmesidir. Bunun önlenmesi için seyreltme (*dropout*) yöntemi ile kullanılması önerilmektedir. Bu şekilde düzenlenen ağlara uygulanması model performansını iyileştirmektedir.

2.2.4.1. Avantajları

- Eksik bilgiler ile çalışabilme becerisine sahiptir. YSA, eğitim aşamasından sonra verilerin eksik bilgi içerdiği durumlarda bile sonuç verebilmektedirler (42).
- Benzer olayları yorumlayarak öğrenir ve ona göre karar alır (42).
- Aynı anda birden çok işi yapabilecek sayısal güce sahiptir (44).

2.2.4.2. Dezavantajları:

- Yapının belirlenmesinde özel bir kural olmadığından uygun ağ yapısının saptanması zor olabilir. Uygun yapı deneme-yanılma ya da deneyim ile belirlenir (42).
- Cevap aramak için kuralların ya da kriterlerin açık olmadığı durumlarda kullanılabilir. Kara kutu denilebilecek sorunları genellikle çözerler ama sorunları nasıl çözdüğünü açıklamak zor olabilir (43).
- Kolaylıkla aşırı uyum gösterme eğilimdedir (44).

2.2.4. Derin Öğrenme

Günümüzde her türlü şekil ve boyuttaki verilerin giderek büyümesi ve kullanılabilirliğinin artması, büyük verinin klasik yazılım araçları ve teknolojileri kullanılarak yönetilmesini ve analiz edilmesini zorlaştırmaktadır. Bu verilerin ele alınması, derin öğrenmenin (DÖ), özellikle büyük verilerde bol miktarda toplanan etiketli ve etiketsiz verilerin üstesinden gelme yeteneği ile desteklenebilir. DÖ'nin önemli bir avantajı, sınıf etiketi olmayan, kategorize edilmemiş büyük miktarda denetimsiz verilerin analizi ve öğrenilmesi için etkin bir araç olmasıdır (45). Diğer sınıflama yöntemlerinin aksine ayrıca değişken seçimi yapılmasına gerek yoktur.

Doğal sinyallerin işlenmesi için insan beyni mekanizmalarına ilişkin biyolojik gözlemlerden esinlenen DÖ yöntemi:

1. Görsel nesne ve ses tanıma/algılama,
2. Metin madenciliği
3. İlaç geliştirme
4. Genetik

gibi birçok araştırma alanındaki başarılı performansından dolayı kullanımı giderek artmaktadır. Özellikle "gerçek zamanlı" veri analizi gereken durumlarda hız ve kesinlik bakımından tercih edilmektedir.

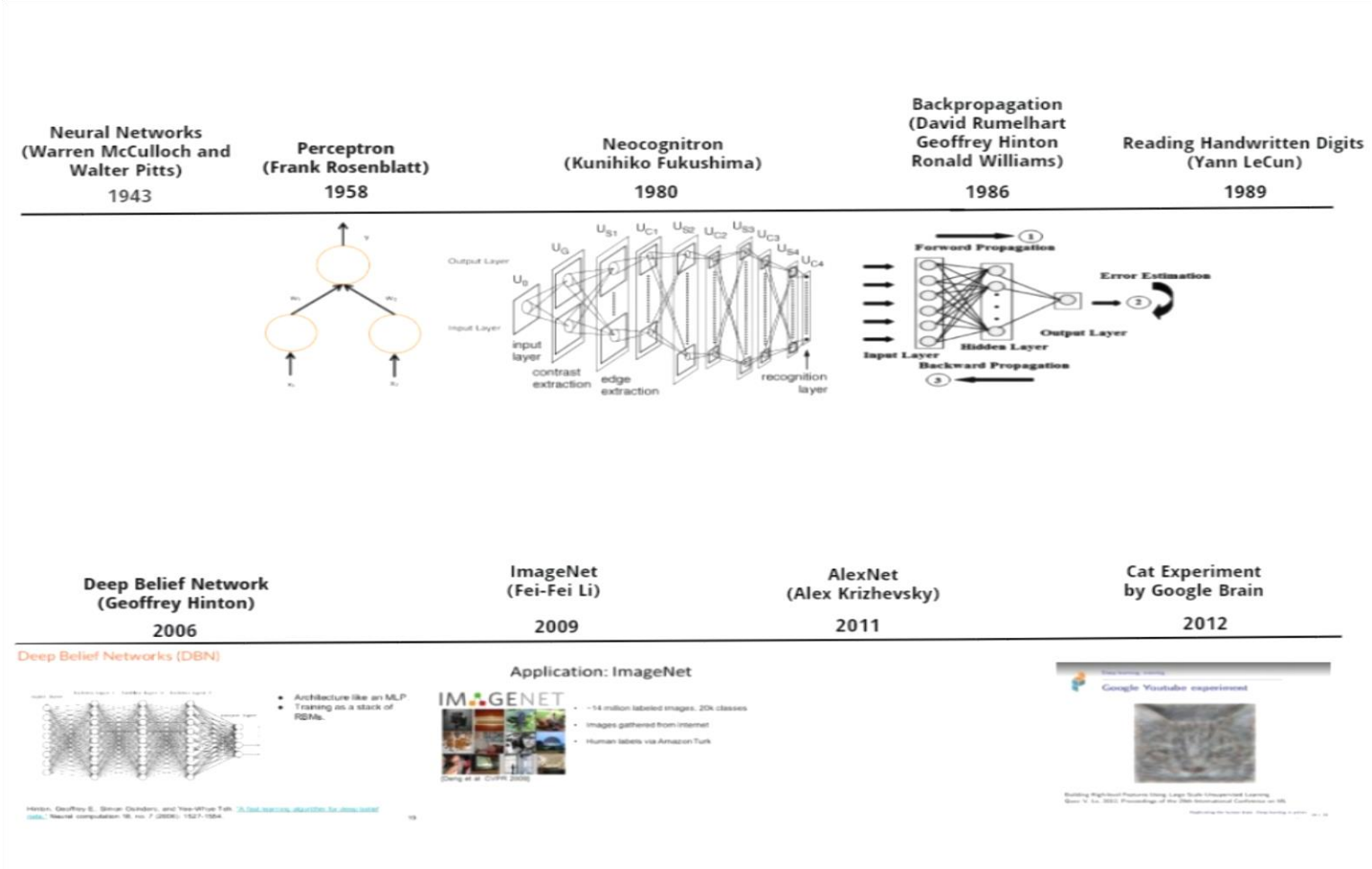
2.2.4.1. Derin Öğrenmenin Tarihçesi

YSA'nın temelleri, McCulloch ve Pitts'in (46) biyolojik öğrenme süreci modellemesi ile atılmıştır. En basit YSA olan, girdi ve çıktı katmanlarından oluşan perseptron (*perceptron),* 1958 yılında Rosenblatt (47) tarafından ortaya konulmuştur. 1980 yılında, Fukushima (48) tarafından geliştirilen ve *Neocognitron* adı verilen çok katmalı yapay sinir ağı modeli, el yazısı tanıma amacıyla geliştirilmiştir. Bu çalışma, genellikle görüntü işleme alanında kullanılan DÖ modeli, Evrişimsel Sinir Ağlarının (*Convolutional Neural Networks*) temeli olarak görülmektedir. Yapay Sinir ağlarının eğitilmesinde kullanılan yöntemlerden biri olan geriye yayılım (*back-propagation*) algoritması 1960'lı yılların başında ortaya çıkmış olsa da tanınırlığı Rumelhart, Hinton ve Williams'ın 1986'daki "Learning Representations by Back-propagating Errors" isimli çalışma ile artmıştır. 1989 yılında ise LeCun tarafından, el yazısı rakamların okunması için geriye yayılım yöntemi ile Evrişimsel Sinir Ağlarının birleştirildiği bir çalışma yayınlanmıştır.

DÖ algoritmaları, 1980'lerde geliştirilmeye başlanan çok katmanlı yapay sinir ağlarının uzantılarıdır. Yapay sinir ağları modelleri yirmi yıl boyunca, uygulamayı kısıtlayan aşırı uyum (*overfitting*) problemleriyle karşı karşıya kalmışken, on yıl kadar önce geliştirilen yeni çıkarsama algoritmaları nedeniyle popülaritesini geri kazanmıştır. Büyük miktarda verinin elde edilebilmesi, geleneksel yapay sinir ağları ile ilgili aşırı uyum probleminin üstesinden gelmeye yardımcı olmuştur. MNIST rakam görüntüsünü sınıflamak için bir oto-kodlayıcı modelinin başarıyla uygulanmasına ilişkin bir makalenin, 2006'daki Neural Computation (49) ve Science (50) dergilerinde yayınlanmasından sonra DÖ, araştırmacıların dikkatini çekmeye başlamıştır.

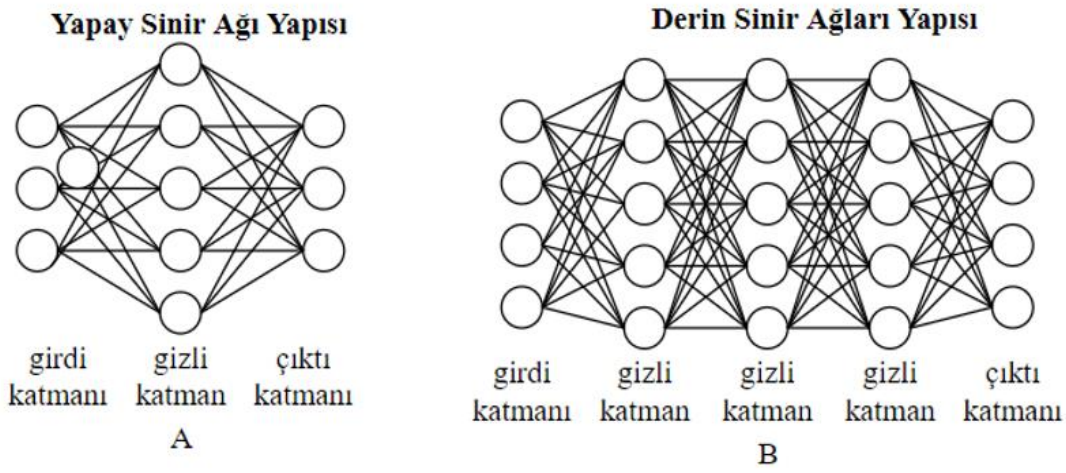
Nesne tanımlama yazılımlarının geliştirilmesi amacıyla oluşturulan ve ImageNet adı verilen görsel veri tabanı 2009 yılında kurulmuştur. Bu veri tabanına milyonlarca nesne resmi eklenmiştir. ImageNet veri tabanında bulunan görsellerin sınıflandırılmasında en başarılı model AlexNet ismi ile geliştirilen Evrişimsel Yapay Sinir Ağı modeli olmuştur. Bu başarı sonrası DÖ araştırmacıların ilgisini çekmeye

başlamış, farklı pek çok alanda kullanılmaya başlanmıştır. 2012 yılında Google Brain ekibi tarafından, "*Cat Experiment*" adı verilen bir projenin sonuçları yayınlanmıştır. Bu çalışmada, denetimsiz öğrenmenin zorluklarının araştırılması amaçlanmıştır. Kedi görsellerinin tanınmasını konu alan bu çalışmada internette yer alan on milyon etiketsiz görüntü verisi kullanılmıştır. DÖ algoritmalarından Evrişimsel Yapay Sinir Ağı modeli etiketsiz veriler üzerinde uygulanmıştır. Bu çalışmada bin adet bilgisayara yayılmış bir sinir ağı modeli kullanılmıştır. Elde edilen sonuçların önceki denemelere göre daha iyi performans gösterdiği belirtilmiştir (Şekil 2.6). Son yıllarda biyoinformatik ve tıbbi görüntü işleme gibi alanlarda DÖ algoritmalarının kullanımı yaygınlaşmaktadır.



Şekil 2.6. Derin öğrenme tarihçesi.

2.2.4.2. Derin Öğrenme Yapısı



Şekil 2.7. Yapay sinir ağı (A) ve derin sinir ağı (B) mimarileri.

YSA'nın geleneksel kullanımından farklı olarak DÖ, birçok gizli nöron ve katmanın mimari bir avantaj olarak yeni eğitim modelleri ile birlikte kullanılmasını gerektirmektedir (Şekil 2.7). Çok sayıda nörona başvurulması mevcut ham verilerin kapsamlı bir şekilde temsil edilmesine izin vermekte iken, ağa daha fazla gizli katman eklemek, gizli katmanların doğrusal olmayan ilişkileri yakalaması nedeniyle daha karmaşık hipotezleri ifade edebilen derin bir mimari oluşturulmasına olanak tanır. Ağın optimum bir şekilde ağırlıklandırılmış olması durumunda, ham veri veya görüntülerin etkin üst düzey temsilleri elde edilir.

DÖ yönteminde geriye yayılım (*backpropagation*) ile parametrelerin güncellenmesi sağlanır. Geriye yayılım işlemi sırasında güncelleme, zincir kuralı (*chain rule*) olarak da adlandırılan, geriye doğru türev alarak farkın bulunması (gradyan düşüşünün hesaplanması) ve bulunan fark değeri ile öğrenme hızı (*learning rate*) parametresinin çarpılması, çıkan sonucun ağırlık değerlerinden çıkarılarak yeni ağırlık değerinin hesaplanmasıyla yapılmaktadır. DÖ algoritmaları temel olarak ardışık katmanların derin mimarilerini içermektedir. Amaç, veriyi çoklu dönüştürme katmanlarından geçirerek karmaşık ve soyut olarak, hiyerarşik bir biçimde öğrenmektir. Her tabakanın çıkışı bir sonraki tabakaya girdi olarak sağlanır. Gizli katmanlarda y = f(x,w) şeklindeki doğrusal fonksiyonda matris çarpımı yapılıp nöronların ağırlığı hesaplandıktan sonra, çıktı doğrusal olmayan bir değere

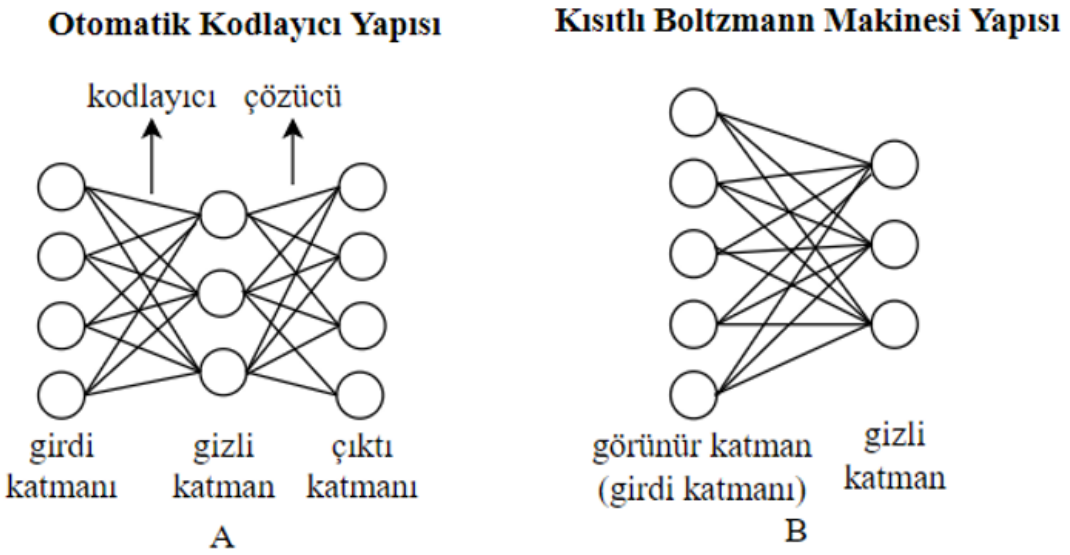
dönüştürülür. DÖ yöntemleri doğrusal olmayan yapıya sahip problemlerin çözümünde diğer yöntemlere göre daha etkili olduğu için, DÖ yöntemleriyle çözülmeye çalışılan problem genelde doğrusal olmayan bir problemdir. Matris çarpımı sonucu elde edilen değerin doğrusal olmayan hale dönüştürülmesi aktivasyon fonksiyonları ile yapılmaktadır. İşlemler sonucu elde edilen son gösterim, girdi verisinin doğrusal olmayan bir fonksiyonudur.

Derin öğrenme modellerinin eğitimi sırasında büyük boyutlu veri setlerinde verilerin tümünü aynı anda analize dahil etmek süre, hız ve bellek açısından maliyetlidir. Bu sorunun üstesinden veri setini yığın (batch) olarak adlandırılan parçalara bölerek gelinebilir. Her yığın için algoritmadaki tüm adımlar uygulanır, bu adımların kaç adet yığın için tamamlanması iterasyonu gösterir. Veri setinin tamamının görülme sayısı *epoch* olarak adlandırılır (51). Örnek olarak, 1000 adet eğitim verisi bulunan bir veri setinde yığın büyüklüğü 200 olarak belirlenirse, bu veri setindeki örneklerin tamamının bir kere (bir *epoch*) eğitim aşamalarından geçmesi için 5 iterasyon gereklidir.

2.2.4.3. Derin Öğrenme Algoritmaları

Yapay sinir ağları, kurulan model göre kodlayıcı (*encoder*) ve çözücü (*decoder*) olarak adlandırılan birimleri içermektedir. Girdi katmanında bulunan nöronlardan $(x_1, x_2, \dots, x_n)$ kodlayıcı fonksiyonu yardımı ile gizli katmanlarda yer alacak nöronlar $(h_1, h_2, \dots, h_n)$ hesaplanır. Oluşturulan bu gizli katmanlardaki nöronlar, çözücü fonksiyonu tarafından çıktı katmanındaki nöronların hesaplanmasında $(y_1, y_2, \dots, y_n)$ kullanılır.

Derin sinir ağları, ikiden fazla gizli katmandan oluşur, böylece fazla sayıdaki ve karmaşık yapıdaki verileri, tek gizli katmanlı yapay sinir ağlarına göre daha yüksek doğrulukta sınıflandırabilmektedir. Derin sinir ağları dışında, kullanılan temel yapı ve amaca yönelik olarak farklı derin öğrenme mimarileri de geliştirilmiştir. Bu derin öğrenme modelleri oluşturmak için genellikle temel yapı taşı olarak kullanılan iki tane öğrenme algoritması vardır: Otomatik Kodlayıcılar (Autoencoders) ve Kısıtlı Boltzmann Makineleri (Restricted Boltzmann Machines) (Şekil 2.8).



Şekil 2.8. Otomatik kodlayıcı (A) ve kısıtlı Boltzmann makinesi (B) yapıları

Otomatik Kodlayıcılar (Autoencoder)

- Bağımlı değişken etiketini atamak değil, giriş vektörünü yeniden oluşturmak üzere eğitilir, yöntem bu nedenle denetimsizdir.
- Giriş verileri yüksek boyutluysa, tek bir gizli katmanı olan bir Otomatik Kodlayıcı tüm veriyi temsil etmek için yeterli olmayabilir. Bu durumda alternatif olarak, bir otomatik kodlayıcı mimarisi oluşturmak için birçok Otomatik Kodlayıcı paralel veya seri olarak yerleştirilebilir.

Kısıtlayıcı Boltzmann Makineleri (Restricted Boltzmann Machines - RBM)

- Kısıtlama, aynı katmanın birimleri arasında hiçbir etkileşim bulunmaması ve bağlantıların yalnızca farklı katmanlardan birimler arasında olmasından kaynaklanmaktadır.
- Görünür değişkenlere gömülü olan istatistiksel yapı gizli değişkenler tarafından yakalanabilir.

Derin sinir ağları dışında, farklı veri türleri ve amaçlara yönelik geliştirilmiş bazı DÖ mimarileri şunlardır:

Yığınlı Otomatik Kodlayıcılar (Stacked Auto Encoders): Pek çok otomatik kodlayıcının birlikte kullanılması ile oluşturulmaktadır. Özellik çıkartma ve boyut

azaltma amaçları ile sınıf etiketi olmayan, denetimsiz verilerde kullanılmaktadır (62, 76).

Derin İnanç Ağları (Deep Belief Network): Birden çok sınırlı Boltzmann makinesinin bir arada kullanılması ile oluşur. Her alt ağın gizli katmanı, bir sonraki RBM'nin görünür katmanına bağlıdır. Derin İnanç Ağları, veri etiketlerini kullanmadan, eğitim verilerinin ortak olasılık dağılımını öğrenen olasılıksal bir model olduğu için genellikle boyut indirgeme veya kümeleme gibi denetimsiz öğrenme amaçları için kullanılmaktadır. Etkili ve gözlemlenmemiş başlatma noktaları sağlayarak doğrusal olmayan parametre kestirimi problemlerinin karmaşıklığını kontrol edebilir (45). Bunun yanı sıra, denetimli öğrenme yöntemlerinden biri olan sınıflama için de kullanılabilmektedir.

Evrişimsel Sinir Ağları (Convolutional Neural Network): Genellikle iki boyutlu resim verilerinde uygulanan, görsel nesne tanımlamada sık olarak kullanılan, beynin görsel korteksinden ilham alan bir DÖ algoritmasıdır. Görsel kortekste, basit ve karmaşık hücreler olmak üzere iki temel hücre tipinde söz edilebilir. Basit hücreler, görsel uyarıcıların alt bölgelerindeki ilkel yapılara tepki gösterir ve karmaşık hücreler, daha karmaşık formları tanımlamak için bilgileri basit hücrelerden sentezler [52]. Bu algoritmanın avantajı, nöron sayısı ve eğitilmesi gereken parametre sayısı çok fazla olan, yüksek derecede korelasyona sahip birimler içeren görüntü verisi gibi verilerin çok boyutlu girdileriyle baş edebilmesidir. Ağlar kıvrım olarak adlandırılan birimleri içerir, girdi alanının küçük bölgelerine kıvrım birimleri ile birlikte çeşitli filtreler eklenerek karmaşık işlemlerin gerçekleştirilmesi kolaylaştırılır.

Tekrarlayan Sinir Ağları (Recurrent Neural Network): Sıralı bilgileri kullanmak için tasarlanmışlardır. Giriş verileri ardışık olarak işlendiğinden, döngüsel bağlantının bulunduğu gizli ünitelerde tekrar tekrar hesaplama yapılır. Dolayısıyla, geçmiş bilgi, gizli birimlerde örtülü olarak saklanır. Çıktının önceki hesaplamalara bağlı olduğu uygulamalarda kullanışlı bir yöntemdir. Tüm adımlar arasında aynı ağırlıkları paylaşır. Tekrarlayan Sinir Ağları'nın avantajı, sıralı olayları hafızasında tutabilmesi, zaman bağımlılıklarını modelleyebilmesidir. Özellikle Doğal Dil İşleme (Natural Language Processing) uygulamalarında büyük başarı göstermektedir [53].

2.2.4.4. Avantajları

- Çok sayıda gizli katman içermesi karmaşık problemlerin çözümünde avantaj sağlamaktadır (45).
- Değişken seçimi yapılmadan veriye doğrudan uygulanabilir, etiketli ve etiketsiz verilerde kullanılabilir.

2.2.4.5. Dezavantajları

- Derin sinir ağları modeli kurulması için eniyilenmesi gereken birçok parametre bulunmaktadır, optimum hiper parametreleri belirlemek zor olabilir.
- Derin öğrenme girdi ve çıktı özellikleri arasında karmaşık ilişkileri öğrenir, ancak yapısında bir nedensellik temsili yoktur.
- Teorik temelin açıklanmasında, görselleştirilmesinde ve modelin yorumlanabilirliğinde diğer makine öğrenme yöntemlerine göre zorluklar yaşanmaktadır (53).

2.2.4.6. Sağlık Alanında Derin Öğrenme Çalışmaları

DÖ yöntemi kullanımı biyoistatistik, biyoinformatik, genetik, biyomedikal görüntü ve sinyal işleme gibi alanlarda giderek yaygınlaşmaktadır. Görsel nesne tanımadaki başarısı nedeniyle radyografik, retinal ve manyetik rözanans görüntüleme gibi yöntemlerden elde edilen biyomedikal görüntü verileri ile elektrokardiyografi, elektroensefalografi gibi sinyal verilerinin kullanıldığı pek çok çalışma bulunmaktadır. Suk ve Shen (54) 2013 yılında yaptıkları çalışmada, manyetik rözenans görüntüleme ve pozitron emisyon tomografisi verilerini kullanarak, Alzheimer ve hafif kognitif bozukluk hastalıklarının sınıflandırılmasında DÖ yöntemini kullanmışlardır. Hua ve ark. (55) tomografi verileri kullanarak akciğer kanserinin, Havaei ve ark (56) beyin tümörünün teşhis edilmesinde, Cao ve ark. (57) tüberküloz teşhisinde, medikal görüntüleme veya sinyal verilerini kullanarak DÖ yöntemini uygulamışlardır.

Proteomik alanında protein sınıflaması (58), ilaç geliştirme (59) gibi amaçlarla DÖ kullanan çalışmalar bulunmaktadır. Transkriptomik alanında DÖ modelleri genellikle sınıflama amacı ile ilgili değil, ilgili transkriptlerin uygun bir şekilde temsil edilmeleri sağlanarak, özellik çıkartmak, verinin boyutunu azaltmak amacıyla

uygulanmışlardır. Ibrahim ve ark. (60) kanser teşhisinde mikroRNA moleküllerinin ifade seviyelerini kullanmışlar, kullandıkları transktiprom veri setlerinde DÖ yöntemlerinden Derin İnanç Ağlarını, denetimsiz öğrenme yöntemi olarak, uygun transkriptlerin seçilerek verinin boyutunu azaltmak amacı ile uygulamışlardır. Chaudhary ve ark. (61) karaciğer kanseri mRNA ve mikroRNA dizileme verileri ile çalışmışlardır. Karaciğer kanseri alt gruplarında hangi transkriptlerin olması gerektiğine DÖ yöntemlerinden biri olan Yığınlı Otomatik Kodlayıcılar kullanarak karar vermişlerdir. Bu şekilde etiketsiz veriyi etiketli hale getirmişlerdir. Ardından da sınıf etiketleri bulunan veriyi DVM ile sınıflandırmışlardır. Benzer yöntemleri Fakoor ve ark. (62) 2013 tarihli çalışmalarında mikrodizi verisi üzerinde uygulamışlardır. Urda ve ark. (63) iki sınıf etiketine sahip RNA dizileme verileri üzerinde DÖ ve LASSO regresyon modelini kullanmışlardır.

DÖ yöntemi, RNA dizileme çalışmalarında genellikle sınıf etiketi bulunmayan verilere sınıf etiketi atama ya da boyut azaltma gibi amaçlar için kullanılmıştır. Alanyazında, RNA dizileme verilerinin DVM, RO gibi klasik veri madenciliği yöntemleri ile veya topluluk sınıflandırıcıları olarak da ifade edilen, birden fazla yöntemden belirli özellikler alınarak yeni oluşturulan sınıflama yöntemleri ile sınıflandırıldığı çalışmalar bulunmaktadır, ancak DÖ yönteminin sınıflamada kullanıldığı çalışma sayısı oldukça azdır.

2.3. Performans Ölçüleri

Sınıflama yöntemlerinin başarıları karşılaştırılmak istendiğinde dikkate alınabilecek bazı ölçüler vardır. Bu çalışmada kullanılan performans ölçüleri hakkında, Tablo 2.3.'deki hata matrisi dikkate alınarak bilgi verilmiştir.

Tablo 2.3. İkili sınıflama problemlerinde kullanılan hata matrisi.

	Gerçek Sınıf	
Kestirilen Sınıf	Pozitif	Negatif
Pozitif	A=Doğru Pozitif (DP)	B=Yanlış Pozitif (YP)
Negatif	C=Yanlış Negatif (YN)	D=Doğru Negatif (DN)

Doğruluk (Accuracy): Sınıflandırıcının genel etkinliğinin bir ölçüsü olan doğruluk, doğru sınıflandırılmış örnek sayısının, toplam örnek sayısına oranıdır.

Dengesiz dağılımlarda kullanılması uygun değildir, hatalı yorumlar yapılmasına neden olabilir.

$$\frac{DP + DN}{DP + DN + YP + YN} \tag{2.21.}$$

Kesinlik (Precision): Doğru sınıflandırılmış pozitif örnek sayısının, pozitif tahmin edilen toplam örnek sayısına oranıdır. Sağlık alanında pozitif kestirim değeri olarak adlandırılmaktadır.

$$\frac{DP}{DP + YP} \tag{2.22.}$$

Sınıf sayısı ikiden fazla olduğunda, l sınıf sayısı olarak alınırsa, her sınıf için hesaplanan kesinlik değerlerinin ortalaması alınarak ortalama bir kesinlik değeri elde edilebilir (64).

$$\frac{\sum_{i=1}^{l} \frac{DP_i}{DP_i + YP_i}}{l} \tag{2.23.}$$

Duyarlılık (Recall): Doğru sınıflandırılmış pozitif örnek sayısının, toplam pozitif örnek sayısına oranıdır. Sınıflandırıcının pozitif örnekleri tahmin etmedeki etkinliğinin bir göstergesidir (64).

$$\frac{DP}{DP + YN} \tag{2.24.}$$

Sınıf sayısı ikiden fazla olduğunda duyarlılık aşağıdaki formül kullanılarak hesaplanabilir:

$$\frac{\sum_{i=1}^{l} \frac{DP_i}{DP_i + YN_i}}{l} \tag{2.25}$$

F Ölçütü (F Measure): Kesinlik ve duyarlılığın harmonik ortalamasıdır. İki performans ölçüsü birlikte değerlendirildiği için sıkılıkla tercih edilen bir ölçüdür. Dengesiz dağılımlarda kullanılması önerilmektedir.

$$\frac{2xKesinikxDuyarlılık}{Kesinlik + Duyarlılık} \quad (2.26.)$$

Kappa: Kappa istatistiği 1960'lı yıllarda Cohen tarafından psikolojik davranış gözlemcileri arasındaki uyumun bir göstergesi olarak tanıtılmıştır. Aynı olayı gözleyen iki veya daha çok gözlemci arasındaki uyum seviyesinin ölçülmesinde kullanılması amaçlanan Kappa istatistiği, farklı yöntemlerin sınıflama performanslarının karşılaştırılmasında da kullanılmaktadır. Bu alanda Kappa istatistiği ilgilenilen yönteme ilişkin sınıflama modelinin tahminleri ile gerçek durum arasındaki uyum seviyesini ölçmek için kullanılmaktadır (65). İki durumlu olaylar için Kappa istatistiği

$$K = \frac{P_0 - P_c}{1 - P_c} \quad (2.27.)$$

formülü ile hesaplanmaktadır. P_0 uyum oranını, P_c şansa bağlı olarak beklenen uyum oranını göstermektedir. Tablo 2.3. yardımıyla P_0 ve P_c aşağıdaki şekilde hesaplanır:

$$P_0 = \frac{A + D}{A + B + C + D} \quad (2.28.)$$

$$P_c = \frac{(A + B)(A + C) + (C + D)(B + D)}{(A + B + C + D)^2} \quad (2.29.)$$

İkiden fazla durum olduğunda P_c'nin hesaplanabilmesi için öncelikle her durum için şansa bağlı beklenen uyum sıklıkları hesaplanıp toplamları alınır. Şansa bağlı olarak beklenen uyum sıklığının toplam gözlem sayısı bölünmesi ile P_c hesaplanmış olur (66).

ROC Eğrisi Altında Kalan Alan (Area Under the ROC Curve): Sınıflandırıcının yanlış sınıflamadan ne kadar kaçınabildiğinin bir göstergesidir. Dikey eksende duyarlılık, yatay eksende 1-seçicilik değerleri bulunan, iki boyutlu ROC grafiğinden elde edilen eğri altında kalan alan, ikili sınıflama problemlerinde sık kullanılan bir performans ölçütüdür (67). Sınıf sayısı ikiden fazla olduğunda, grafik çizilemeyeceği için eğri altında kalan alan hesaplanamamaktadır. Bu durumda Hand ve Till'in (68), ortalama eğri altında kalan alan yaklaşımı kullanılabilir. Sınıf sayısı c

ile gösterilirse, tüm ikili sınıf karşılaştırması için hesaplanan eğri altında kalan alanlardan aşağıdaki formül kullanılarak ortalama bir değer hesaplanabilir.

$$\frac{2}{c(c-1)}\sum Eğri\ altında\ kalan\ alan \qquad (2.30.)$$

3. GEREÇ VE YÖNTEM

3.1. Uygulama

Çalışmada akciğer ve böbrek kanseri olmak üzere iki farklı kanser türüne ait, RNA dizileme yöntemi ile edilmiş gen ifade verileri kullanılmıştır. Veri setlerine The Cancer Genome Atlas (TCGA) veri portalı üzerinden ulaşılabilmektedir. Her iki veri setinde de 20.531 gene ilişkin okuma sayıları bulunmaktadır. Akciğer kanseri veri setinde kanserin iki alt türü, böbrek kanserinde ise üç alt türü bulunmaktadır. Veri setleri ile ilgili bilgiler Tablo 3.1.'de yer almaktadır.

Tablo 3.1. Çalışmada kullanılan veri setleri.

Veri Seti	Kanser Alt Türü	n
Akciğer Kanseri	Akciğer Adenokarsinomu (Lung Adenocarcinoma - LUAD)	576
	Akciğer Skuamöz Hücre Karsinoması (Lung Squamous Cell with Carcinoma - LUSC)	552
Böbrek Kanseri	Papiller Hücreli Böbrek Karsinomu (Kidney Renal Papillary Cell – KIRP)	606
	Berrak Hücreli Böbrek Karsinomu (Kidney Renal Clear Cell - KIRC)	323
	Kromofob Hücreli Böbrek Karsinomu (Kidney Chromophobe Carcinomas - KICH)	91

3.1.1. Farklı İfade Edilmiş Genlerin Bulunması

Kanser alt türleri arasında farklı ifade edilmiş genlerin bulunması için yeni nesil dizileme verilerinin analizinde kullanılan yazılımlardan biri olan Partek Flow yazılımı kullanılarak, iki farklı yöntem ile analiz yapılmıştır. Farklı ifade edilmiş genlerin analizinde, Partek Flow yazılımı ile R programında yer alan DESeq2 paketi kullanılarak analiz yapılabilmektedir. Kullanılan ilk yöntem DESeq2 paketi, ikinci yöntem ise yazılımın kendi yöntemi olan Gen Belirleyici Analiz'dir (Gene Specific Analysis - GSA). Gen Belirleyici Analiz, kullanılan diğer yöntemlere göre daha esnek bir yaklaşım sunmaktadır. Farklı ifade edilmiş genlerin bulunması için kullanılan

yöntemler genellikle Poisson, Negatif Binom gibi belirli bir modele dayanmaktadır. Gen Belirleyici Analiz ise ilgili gene ilişkin özellikleri dikkate alarak farklı dağılımların uygulanabilmesine, varsa etkileşim teriminin eklenebilmesine olanak sağlar. Örneğin ilgilenilen veride birinci gen için Poisson dağılımı daha uygunken 2. gen için Negatif Binom dağılımı daha uygun olabilir. Gen Belirleyici Analiz, genler için en uygun modelleri kullanarak farklı ifade edilmiş gen analizini gerçekleştirir (69).

3.1.2. Filtreleme

Kanser alt türleri arasında farklı ifade edilmiş genlere ilişkin p değerleri bulunduktan sonra, $p<0,05$ olan genler listelenmiştir. Kalan genler yanlış bulgu oranının (fdr) 0,05, 0,02 ve 0,01 olduğu durumlara göre filtrelenerek her iki veri seti için, iki farklı ifade analizi yöntemine göre üç farklı filtreleme yapılmıştır. Farklı filtreler uygulandıktan sonra elde edilen veri setlerine ilişkin gen sayıları Tablo 3.2. ve Tablo 3.3.'de verilmiştir.

Tablo 3.2. Akciğer kanseri veri setinin filtrelenmesi.

Yöntem	Filtre	Gen Sayısı
DESeq2	fdr<0,01	14.876
DESeq2	fdr<0,02	15.245
DESeq2	fdr<0,05	15.861
GSA	fdr<0,01	10.190
GSA	fdr<0,02	10.477
GSA	fdr<0,05	10.864

Tablo 3.3. Böbrek kanseri veri setinin filtrelenmesi.

Yöntem	Filtre	Gen Sayısı
DESeq2	fdr<0,01	8.726
DESeq2	fdr<0,02	9.416
DESeq2	fdr<0,05	10.873
GSA	fdr<0,01	6.072
GSA	fdr<0,02	6.528
GSA	fdr<0,05	7.264

3.1.3. Dönüşüm

Uygulanan filtreler sonucu oluşturulan veri setlerine, sınıflama analizine uygun hale getirilmesi amacıyla, DESeq2 paketinde yer alan vst fonksiyonu kullanılarak

varyans dengeleyici dönüşüm uygulanmıştır. Kullanılan bu fonksiyon verinin normalleştirilmemiş, tam okuma sayıları üzerinden bir dönüşüm gerçekleştirir. Fonksiyon kendi içinde veri normalleştirmesi yaptıktan sonra dönüşüm gerçekleştirmektedir.

3.2. Yöntem

Dönüşüm uygulanan veri setleri, veri setinde yer alan örneklerin rastgele %80'i eğitim, %20'si test setine dahil olacak şekilde bölünmüştür. Akciğer kanseri veri setlerinde, her bir veri setinde 1128 örnek bulunmakta olup eğitim setinde 903, test setinde 225 örnek yer almaktadır. Böbrek kanseri veri setlerinde bulunan 1020 örnek, eğitim setinde 817, test setinde 203 örnek olacak şekilde iki veri setine ayrılmıştır.

Sınıflama analizleri, Amerika Birleşik Devletleri merkezli teknoloji şirketi Microsoft desteği ile gerçekleştirilmiştir. 64 gb belleğe sahip, 8 çekirdekli, 2 NVDIA GEFORCE GTX 1050 GPU'ya sahip özel bir bilgisayar Microsoft Azure bulut bilişim ortamında kullanılmış, bu sayede analizler oldukça hızlı bir şekilde sonuçlandırılmıştır. DÖ modelleri için, çeşitli yapay sinir ağları algoritmalarının uygulanabildiği, Microsoft'un derin öğrenme için geliştirdiği açık kaynak kodlu bir araç olan *Microsoft Cognitive Toolkit* (CNTK) kullanılmıştır. CNTK, C++ veya Python programlama dillerinin kullanılmasına olanak sağlamaktadır. Grafik işlemci üniteleri arasında parallelleştirme ile işlemler hızlı olarak gerçekleştirilebildiği için büyük veri setlerinde oldukça kullanışlıdır (70). DÖ modelleri için uygun gizli katman sayılarına, gizli katmanlarda yer alacak nöron sayılarına, epoch ve yığın boyutu değerlerine, farklı değerler denenip hata oranları karşılaştırılarak karar verilmiştir. Aktivasyon fonksiyonu olarak ReLU kullanılmıştır. YSA modelleri tek gizli katmandan oluşmakta olup, derin öğrenmede karar verilen nöron sayıları kullanılarak oluşturulmuştur. Öğrenme hızı 0,001 olarak alınmıştır.

Veriler sırasıyla RO, DVM, YSA ve DÖ yöntemleri ile sınıflandırılmıştır. Her yöntemden önce, ilgili yönteme ait ön bilgiler yardımıyla parametrelerin alabilecekleri değerlerin aralıkları belirlenmiştir. Ardından aralıktaki belirli noktalar kullanılarak en uygun parametrelere karar verilmiştir. RO yöntemi için R'da bulunan randomForest (71) paketi kullanılmıştır. Eğitim verileri belirli parametreler için tuneRF fonksiyonu ile çalıştırılmıştır. Sınıflama hatası en düşük olan parametreler seçilerek modeller

kurulmuştur. Değişken sayısı p ile gösterilsin, ağacın bölünmesinde seçilecek rastgele özellik sayısı, genellikle sınıflama problemlerinde $\sqrt{p}$; regresyon problemlerinde $p * 0{,}30$ olarak alınmaktadır. Ancak Probst ve arkadaşları (72), büyük ölçekli veriler söz konusu olduğunda sınıflama ve regresyon problemleri için özellik sayısının yüksek alınmasının hata oranını azaltabileceğini belirtmişlerdir. RO yöntemi için ağaç sayısı ve ağaçlar oluşturulurken ağacın bölünüp dallara ayrılmasında rastgele seçilecek özellik sayısı parametreleri üzerinde durulmuştur. Genel olarak tüm veri setlerinde eğitim setinin hata oranın ağaç sayısı (*ntree*) 1000 ve özellik sayısı (*mtry*) $\sqrt{p}$ olduğu durumlarda düşük olduğu görülmüştür.

DVM yöntemini uygulamak için R programında bulunan e1071 (73) paketi kullanılmıştır. Çekirdek fonksiyonu olarak doğrusal olmayan, genellikle en çok kullanılan fonksiyon olan RTF seçilmiştir. Modelde yer alan değişken sayısı arttıkça DVM aşırı uyum sorunu yaşama eğilimindedir. Bu sorunun çözülmesi için uygun *gamma* ve maliyet (*cost*) parametreleri seçilmelidir (74). *Gamma* parametresi, hiper düzlem üzerinde, eğitim setinde yer alan her bir örneğin etkisini gösteren bir ölçüdür. Düşük değerler alması değerlerin uzak, yüksek değer alması ise değerlerin yakın olduğunu gösterir. *Gamma* parametresi çok büyükse, hiper düzlemde yer alan vektörlerin etki alanının yalnızca kendilerini kapsadığı; *gamma* parametresi çok küçükse seçilen herhangi bir vektörün etki bölgesinin tüm eğitim setini kapsadığı söylenebilir (75). Verilerin şeklinin yakalanması için uygun *gamma* parametresinin seçilmesi önemlidir. *Gamma* parametresi genellikle $1/p$ olarak alınmaktadır, ancak veri setine bağlı olarak farklı değerler de alınabilir. *Cost* parametresi, eğitim verileri kullanılarak elde edilen modelin hatası ve vektörler arasındaki uzaklığın ayarlanması arasındaki değişimi kontrol eder. Hata sıklığı ile modelin karmaşıklığını dengelemek için, veri setine uygun olarak belirlenmelidir. e1071 paketinde *cost* parametresi, araştırmacı değiştirmediği sürece 1 olarak alınmaktadır. *Cost* değerinin büyük alınması, vektörler arası uzaklığın az; küçük alınması ise vektörler arası uzaklığın fazla olmasına neden olur. Veri setine uygun *gamma* ve *cost* değerlerinin seçilebilmesi için tuneSVM fonksiyonu kullanarak farklı değerler için hatanın nasıl değiştiğine bakılmıştır. Kullanılan yöntemler için seçilen en uygun parametreler aşağıdaki tablolarda verilmiştir.

Tablo 3.4. Akciğer kanseri veri seti için kullanılan parametreler.

Yöntem	Filtre	RO	DVM	YSA	DÖ
DESeq2	fdr<0,01	ntree=1000 mtry=121	gamma=1x10^{-5} cost=10	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch=8 yığın boyutu=65
DESeq2	fdr<0,02	ntree=1000 mtry=123	gamma=1x10^{-5} cost=10	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =8 yığın boyutu=65
DESeq2	fdr<0,05	ntree=1000 mtry=126	gamma=1x10^{-5} cost=10	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =8 yığın boyutu =65
GSA	fdr<0,01	ntree=1000 mtry=101	gamma=1x10^{-5} cost=10	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =8 yığın boyutu =65
GSA	fdr<0,02	ntree=1000 mtry=102	gamma=1x10^{-5} cost=20	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =8 yığın boyutu =65
GSA	fdr<0,05	ntree=1000 mtry=104	gamma=1x10^{-5} cost=10	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =8 yığın boyutu =65

Tablo 3.5. Böbrek kanseri veri seti için kullanılan parametreler.

Yöntem	Filtre	RO	DVM	YSA	DÖ
DESeq2	fdr<0,01	ntree=1000 mtry=93	gamma=1x10^{-4} cost=10	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =10 yığın boyutu =59
DESeq2	fdr<0,02	ntree=1000 mtry=97	gamma=1x10^{-5} cost=20	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =10 yığın boyutu =59
DESeq2	fdr<0,05	ntree=1000 mtry=104	gamma=1x10^{-5} cost=20	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =10 yığın boyutu =59
GSA	fdr<0,01	ntree=1000 mtry=78	gamma=1x10^{-4} cost=20	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =10 yığın boyutu =59
GSA	fdr<0,02	ntree=1000 mtry=81	gamma=1x10^{-4} cost=20	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =10 yığın boyutu =59
GSA	fdr<0,05	ntree=1000 mtry=85	gamma=1x10^{-5} cost=20	nöron sayısı=200	gizli katman sayısı=20 nöron sayısı=200 epoch =10 yığın boyutu =59

4. BULGULAR

Çalışmada, akciğer ve böbrek olmak üzere, iki farklı RNA dizileme verisi kullanılmıştır. Farklı ifade edilmiş genleri bulmak için DESeq2 ve GSA yöntemleri uygulanmış, her yöntem için de üç farklı fdr değeri kullanılarak filtreleme yapılmıştır. Tablo 4.1., 4.2., 4.3. ve 4.4.'de, farklı veri setleri ve filtrelerde sınıflama yöntemlerine ilişkin performans ölçüleri yer almaktadır.

Tablo 4.1. Akciğer kanseri veri setinde DESeq2 tekniğiyle belirlenen farklı ifade edilmiş genler kullanılarak kurulan sınıflama modellerinin performanslarının karşılaştırılması.

Filtre	Yöntem	EAKA	Doğruluk	Kesinlik	Duyarlılık	F Ölçütü	Kappa
fdr<0,01	RO	0,933	0,933	0,910	**0,965**	0,937	0,866
	DVM	0,937	0,938	0,924	0,957	0,940	0,875
	YSA	0,885	0,884	0,916	0,852	0,883	0,769
	DÖ	**0,965**	**0,964**	**0,991**	0,939	**0,964**	**0,929**
fdr<0,02	RO	0,924	0,924	0,895	0,965	0,929	0,849
	DVM	0,937	0,938	0,924	0,957	0,940	0,875
	YSA	0,876	0,876	0,892	0,861	0,876	0,751
	DÖ	**0,964**	**0,964**	**0,957**	**0,974**	**0,966**	**0,964**
fdr<0,05	RO	0,946	0,947	0,919	0,983	0,950	0,893
	DVM	**0,964**	**0,964**	**0,956**	0,956	**0,965**	0,929
	YSA	0,855	0,853	0,910	0,791	0,847	0,853
	DÖ	0,955	0,956	0,949	**0,965**	0,957	**0,956**

Tablo 4.2. Akciğer kanseri veri setinde GSA tekniğiyle belirlenen farklı ifade edilmiş genler kullanılarak kurulan sınıflama modellerinin performanslarının karşılaştırılması.

Filtre	Yöntem	EAKA	Doğruluk	Kesinlik	Duyarlılık	F Ölçütü	Kappa
fdr<0,01	RO	0,875	0,876	0,848	0,922	0,883	0,751
	DVM	0,827	0,827	0,828	0,835	0,831	0,653
	YSA	0,811	0,809	0,867	0,739	0,798	0,619
	DÖ	**0,943**	**0,942**	**0,955**	**0,930**	**0,943**	**0,885**
fdr<0,02	RO	0,902	0,902	0,912	0,896	0,904	0,804
	DVM	0,859	0,858	0,895	0,817	0,855	0,716
	YSA	0,801	0,800	0,843	0,748	0,793	0,601
	DÖ	**0,974**	**0,973**	**1,000**	**0,948**	**0,973**	**0,947**
fdr<0,05	RO	0,893	0,893	0,882	0,913	0,897	0,786
	DVM	0,840	0,840	0,832	0,861	0,846	0,680
	YSA	0,801	0,800	0,850	0,739	0,791	0,601
	DÖ	**0,961**	**0,960**	**1,000**	**0,922**	**0,959**	**0,920**

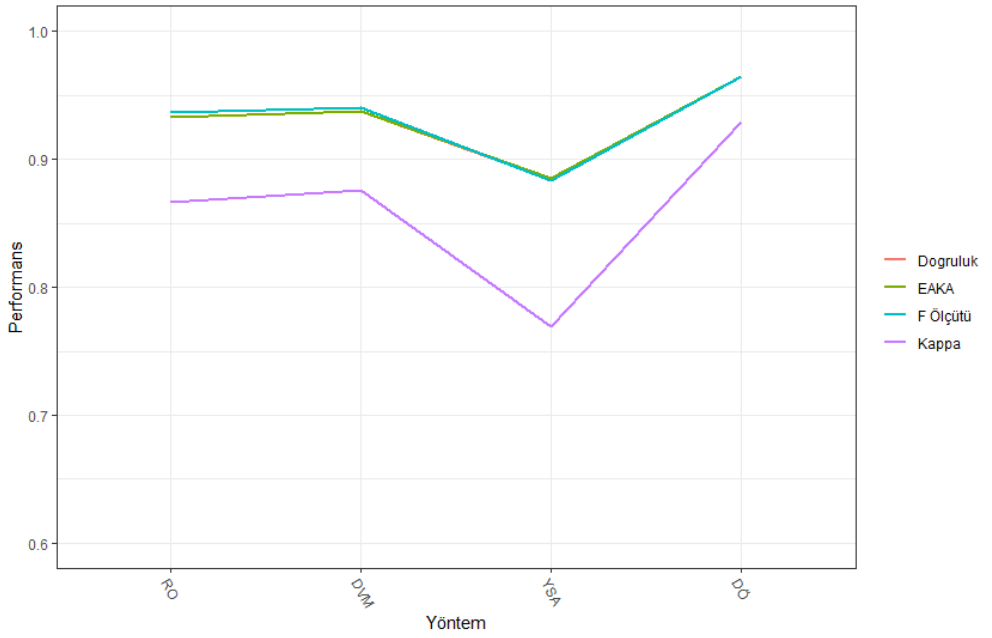
Tablo 4.3. Böbrek kanseri veri setinde DESeq2 tekniğiyle belirlenen farklı ifade edilmiş genler kullanılarak kurulan sınıflama modellerinin performanslarının karşılaştırılması.

Filtre	Yöntem	EAKA	Doğruluk	Kesinlik	Duyarlılık	F Ölçütü	Kappa
fdr<0,01	RO	0,838	0,906	0,851	0,838	0,845	0,825
	DVM	0,885	0,941	0,899	0,894	0,896	0,889
	YSA	0,873	0,867	0,778	0,878	0,799	0,775
	DÖ	**0,973**	**0,966**	**0,907**	**0,978**	**0,935**	**0,938**
fdr<0,02	RO	0,966	0,956	0,936	0,929	0,932	0,917
	DVM	**0,969**	**0,961**	**0,937**	**0,937**	**0,937**	**0,927**
	YSA	0,866	0,857	0,781	0,889	0,795	0,763
	DÖ	0,936	0,936	0,851	0,907	0,872	0,884
fdr<0,05	RO	0,900	0,931	0,922	0,875	0,894	0,871
	DVM	**0,961**	**0,956**	**0,943**	0,923	**0,931**	**0,918**
	YSA	0,934	0,842	0,840	0,873	0,844	0,726
	DÖ	0,948	0,946	0,869	**0,931**	0,893	0,902

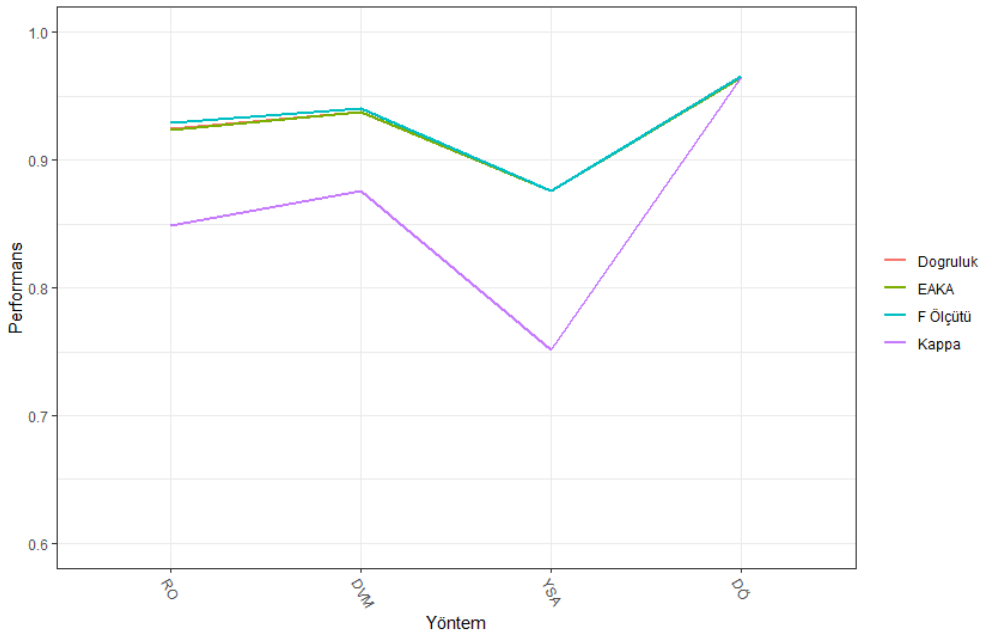
Tablo 4.4. Böbrek kanseri veri setinde GSA tekniğiyle belirlenen farklı ifade edilmiş genler kullanılarak kurulan sınıflama modellerinin performanslarının karşılaştırılması.

Filtre	Yöntem	EAKA	Doğruluk	Kesinlik	Duyarlılık	F Ölçütü	Kappa
fdr<0,01	RO	0,924	0,956	0,946	0,923	0,933	0,917
	DVM	0,955	**0,970**	**0,960**	**0,947**	**0,953**	**0,945**
	YSA	0,925	0,842	0,816	0,846	0,830	0,712
	DÖ	**0,958**	0,936	0,893	0,916	0,904	0,882
fdr<0,02	RO	0,836	0,902	0,846	0,836	0,841	0,816
	DVM	0,882	0,926	0,883	0,883	0,883	0,863
	YSA	0,926	0,857	0,844	0,844	0,843	0,731
	DÖ	**0,960**	**0,946**	**0,933**	**0,916**	**0,923**	**0,897**
fdr<0,05	RO	0,941	0,956	**0,934**	0,916	0,924	0,917
	DVM	0,967	0,956	0,932	0,934	**0,933**	0,918
	YSA	0,914	0,837	0,781	0,823	0,791	0,704
	DÖ	**0,968**	0,956	0,889	**0,968**	0,917	**0,920**

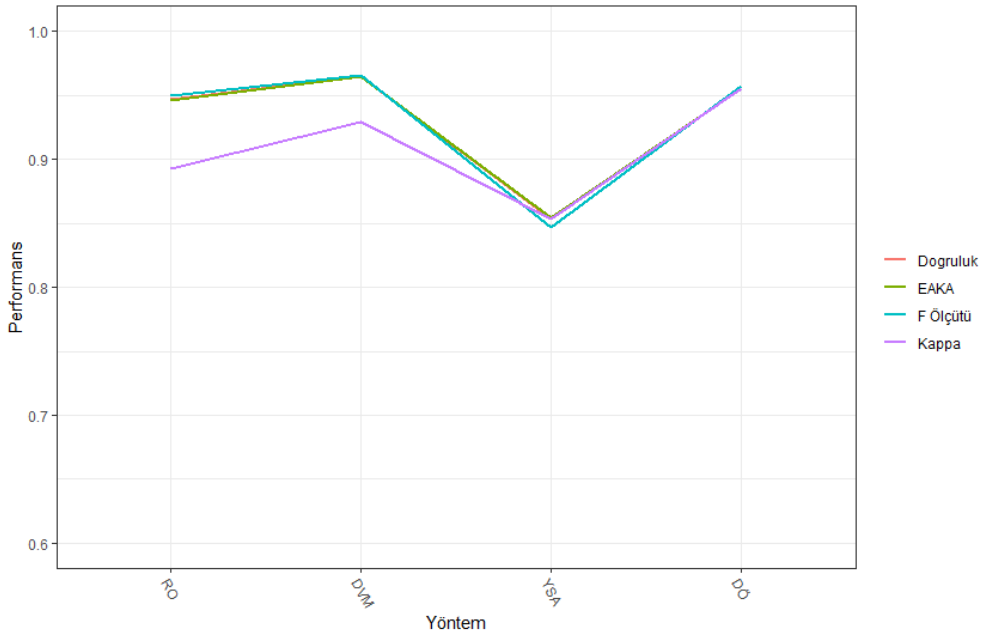
Sonuçları yorumlamada kolaylık sağlaması için doğruluk, EAKA, F ölçütü ve Kappa değerleri aşağıdaki grafiklerle gösterilmiştir (Şekil 4.1. – Şekil 4.12.).



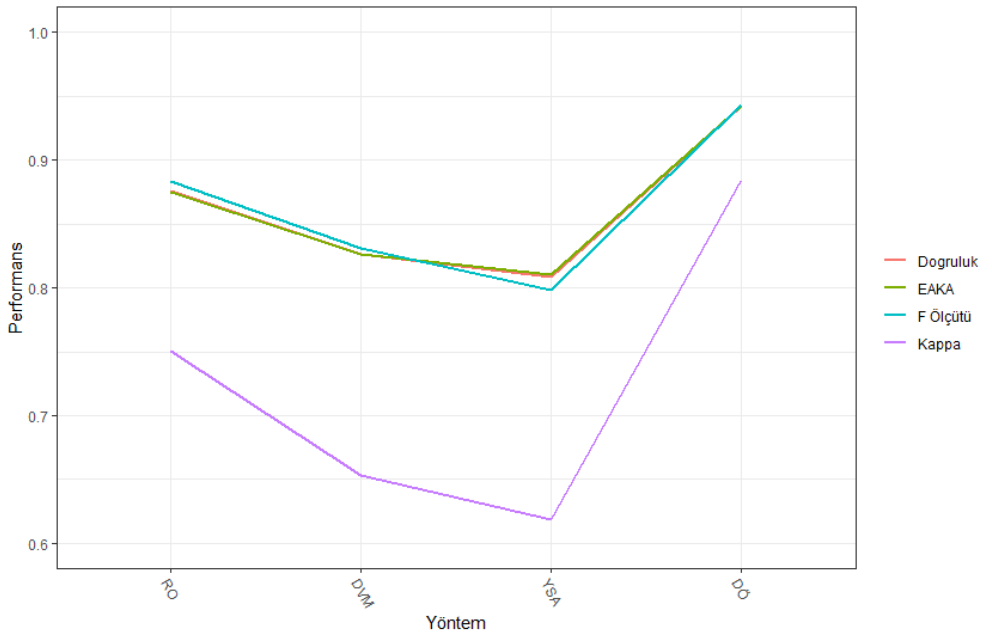
Şekil 4.1. DESeq2 yöntemi ve fdr<0,01 ile filtrelenmiş akciğer kanseri veri seti için sınıflama performanslarının karşılaştırılması.



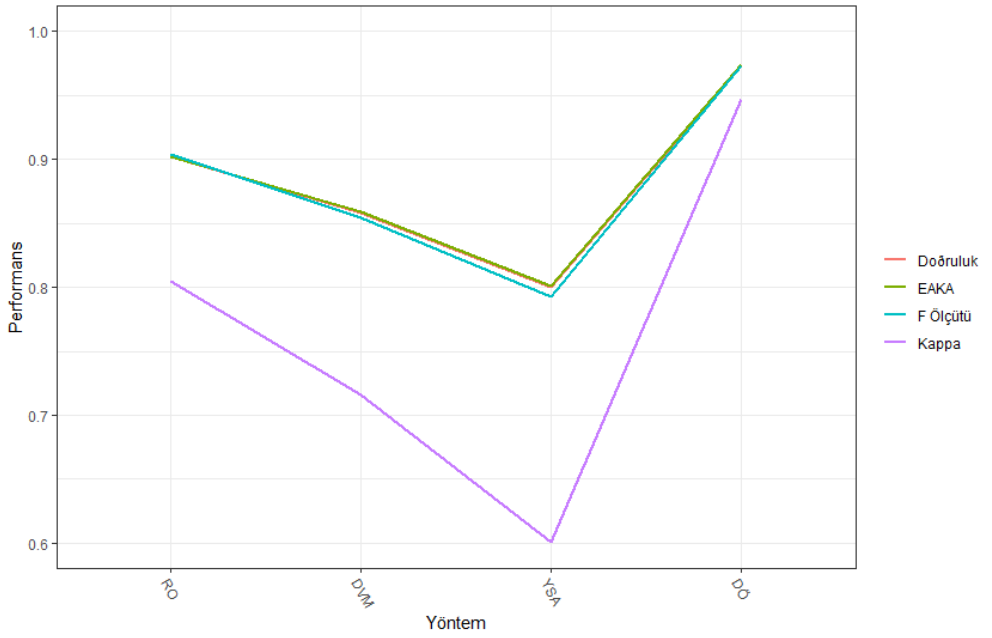
Şekil 4.2. DESeq2 yöntemi ve fdr<0,02 ile filtrelenmiş akciğer kanseri veri seti için sınıflama performanslarının karşılaştırılması.



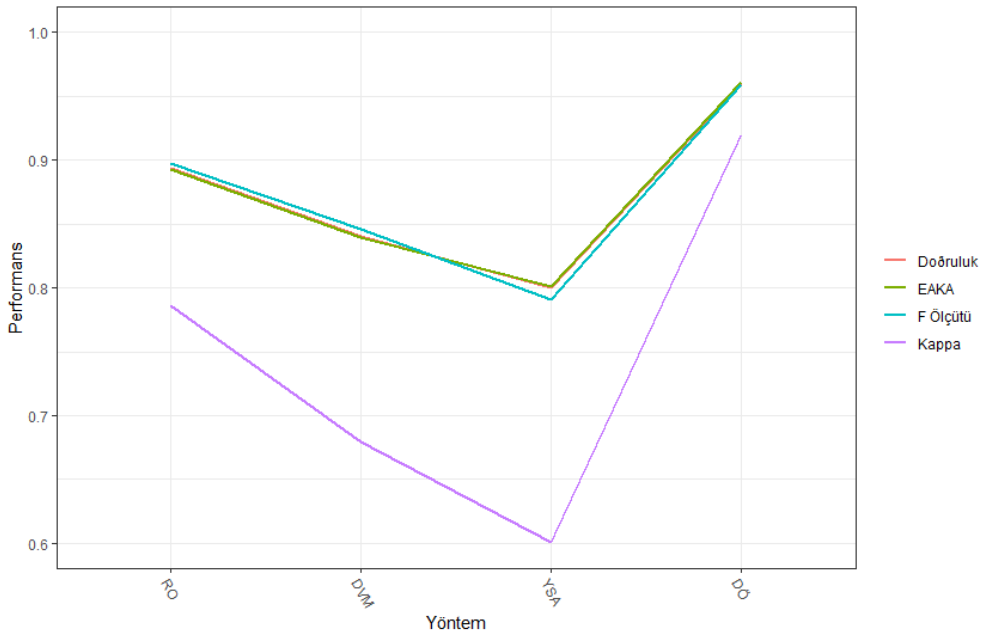
Şekil 4.3. DESeq2 yöntemi ve fdr<0,05 ile filtrelenmiş akciğer kanseri veri seti için sınıflama performanslarının karşılaştırılması.



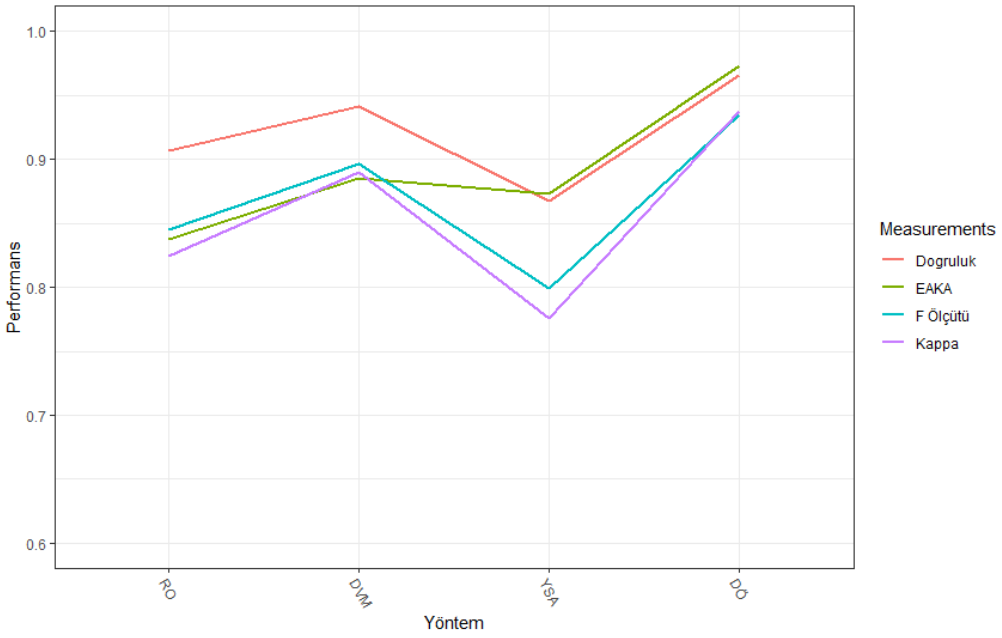
Şekil 4.4. GSA yöntemi ve fdr<0,01 ile filtrelenmiş akciğer kanseri veri seti için sınıflama performanslarının karşılaştırılması.



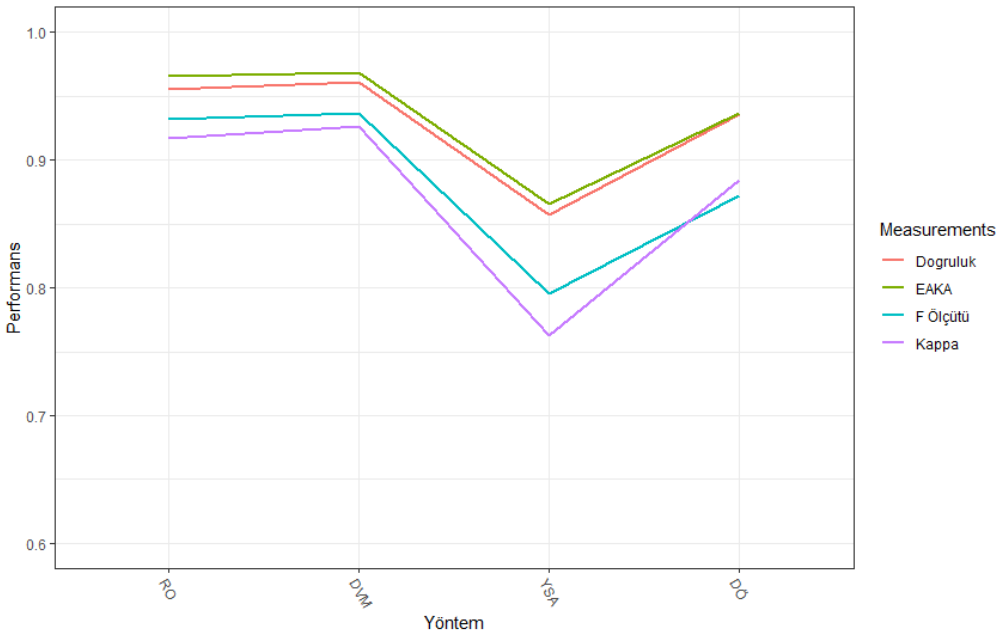
Şekil 4.5. GSA yöntemi ve fdr<0,02 ile filtrelenmiş akciğer kanseri veri seti için sınıflama performanslarının karşılaştırılması.



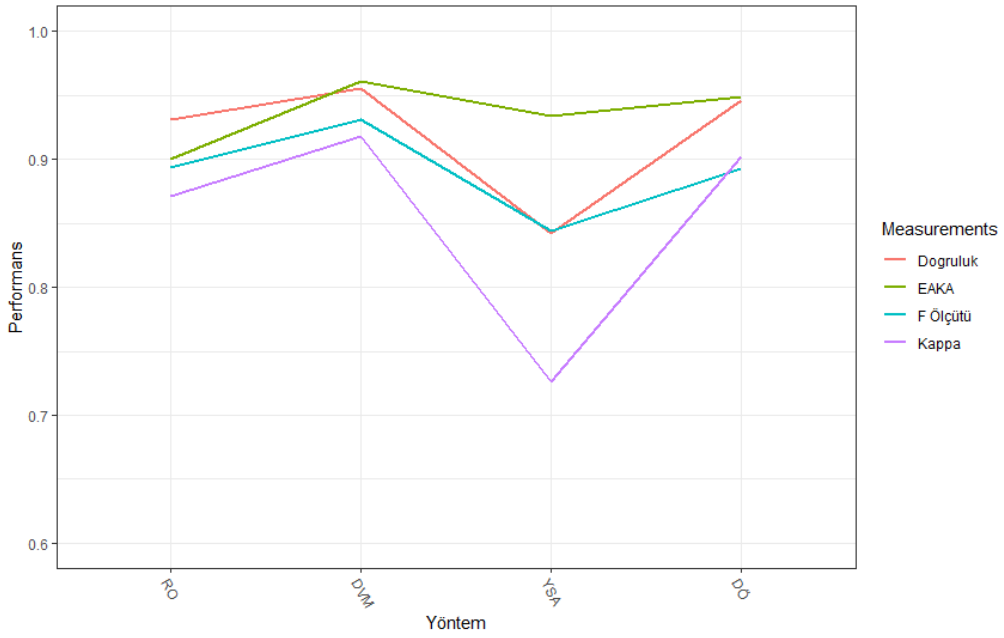
Şekil 4.6. GSA yöntemi ve fdr<0,05 ile filtrelenmiş akciğer kanseri veri seti için sınıflama performanslarının karşılaştırılması.



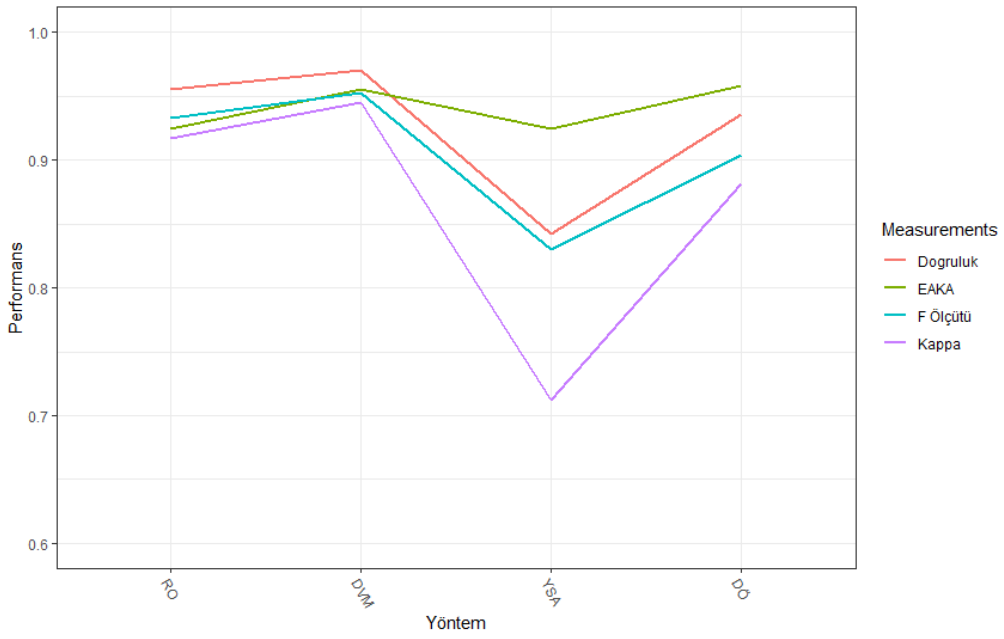
Şekil 4.7. DESeq2 yöntemi ve fdr<0,01 ile filtrelenmiş böbrek kanseri veri seti için sınıflama performanslarının karşılaştırılması.



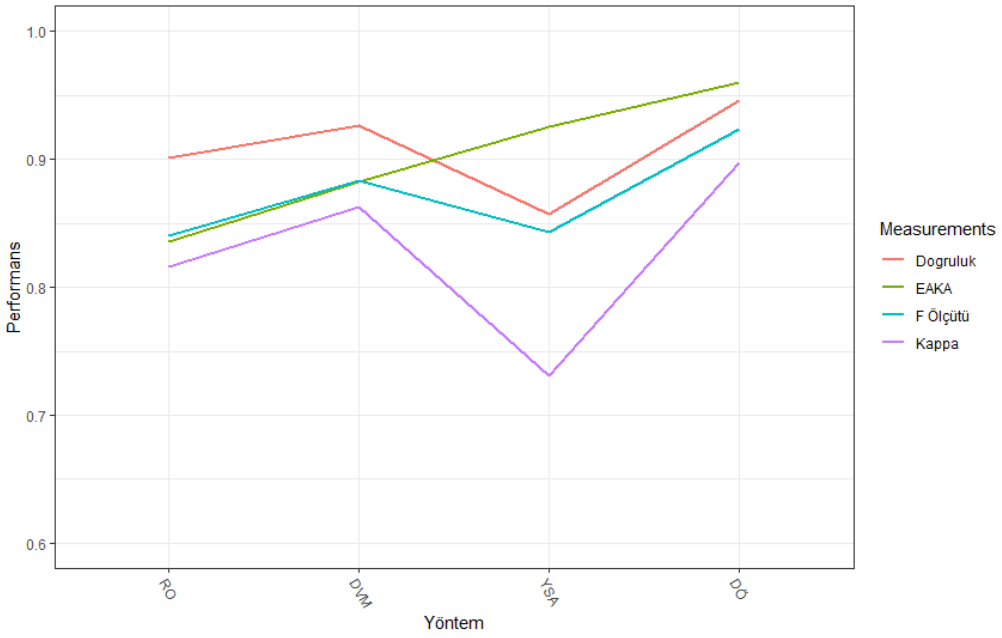
Şekil 4.8. DESeq2 yöntemi ve fdr<0,02 ile filtrelenmiş böbrek kanseri veri seti için sınıflama performanslarının karşılaştırılması.



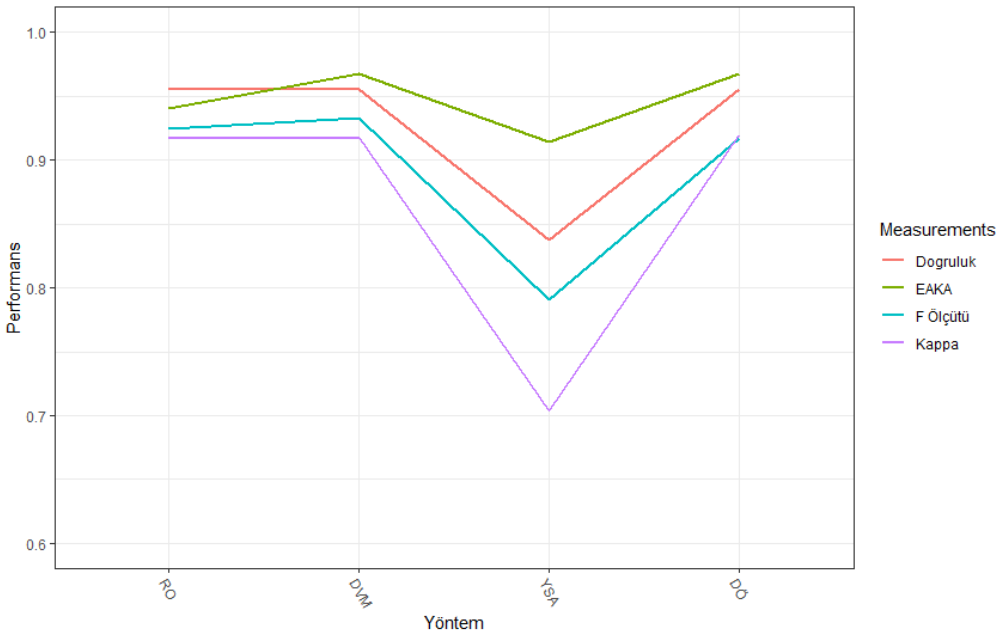
Şekil 4.9. DESeq2 yöntemi ve fdr<0,05 ile filtrelenmiş böbrek kanseri veri seti için sınıflama performanslarının karşılaştırılması.



Şekil 4.10. GSA yöntemi ve fdr<0,01 ile filtrelenmiş böbrek kanseri veri seti için sınıflama performanslarının karşılaştırılması.



Şekil 4.11. GSA yöntemi ve fdr<0,02 ile filtrelenmiş böbrek kanseri veri seti için sınıflama performanslarının karşılaştırılması.



Şekil 4.12. GSA yöntemi ve fdr<0,05 ile filtrelenmiş böbrek kanseri veri seti için sınıflama performanslarının karşılaştırılması.

5. TARTIŞMA

Bu çalışmada iki farklı RNA dizileme verisi kullanılmıştır. İlk veri seti olan akciğer kanseri veri seti, iki alt kanser türüne sahiptir ve verilerin sınıflara dağılımı dengelidir. İkinci veri seti olan böbrek kanserinin ise üç alt türü bulunmakta olup sınıflardaki gözlem sayıları bakımından dengesiz bir dağılıma sahiptir. Her iki veri setinde de 20.531 adet gen bulunmaktadır. Özellik boyutunu azaltmak için iki farklı analiz yapılmıştır. Her veri seti için, kanser alt türleri arasında ifade seviyesi anlamlı olarak farklı bulunan genler seçilerek boyut azaltılmıştır. GSA ile yapılan analizde, DESeq2 ile yapılan analize göre daha az sayıda genin ifade seviyesi anlamlı olarak bulunmuştur. Her iki veri seti için de bu iki yöntemle analiz yapılmıştır. Ardından yanlış pozitiflerin oranının kontrol altına alınmasını sağlayan yanlış bulgu oranı filtre olarak kullanılmıştır. Yanlış bulgu oranı %1, %2 ve %5'den küçük olan genler seçilerek her veri seti ve her analiz yöntemi için, üç farklı filtreye göre üç veri seti elde edilmiştir. Böylece sınıflama yöntemleri her veri setinde kullanılarak sonuçların geçerliliği hakkında yorum yapabilmek amaçlanmıştır. Akciğer kanseri veri seti DESeq2 yöntemi ile analiz edildiğinde, farklı filtreler için elde edilen gen sayıları yaklaşık olarak on beş bin iken GSA yönteminde yaklaşık olarak on bindir. DESeq2 yöntemi ile böbrek kanseri veri setinde ortalama dokuz bin gen seçilmiş iken GSA yöntemi ile 6 bin gen seçildiği söylenebilir. Veri setlerinde özellik çıkarımı bu şekilde sağlandıktan sonra verilere dönüşüm uygulanmıştır. Verilerin dönüşümü için DESeq2 paketinin kullanılması tercih edilmiştir. Bu pakette yer alan rlog ve vst fonksiyonlarından, gözlem sayıları büyük olduğu ve farklı koşullar altında toplamda on iki adet veri seti elde edildiği için, hız açısından daha avantajlı olan vst fonksiyonu tercih edilmiştir. Kullanılan dönüşüm fonksiyonu verilerin normalleştirilmemiş, ham hallerini istemektedir. Veriler bu fonksiyon ile dönüşümden önce normalleştirilir.

Veriler boyut azaltma ve dönüşüm işleminden sonra sınıflama yöntemlerinin uygulanmasına hazır hale gelmişlerdir. Tüm veri setleri RO, DVM, YSA ve DÖ yöntemleri ile sınıflandırılmışlardır. Çalışmada dengeli ve dengesiz sınıf dağılımlarına sahip, farklı filtreler kullanılmış veri setlerinde hangi yöntemlerin daha başarılı performansa sahip oldukları karşılaştırılmak istenmektedir.

Akciğer kanseri veri seti için sınıflamada kullanılacak genler DESeq2 ile yöntemi ve fdr<0,01 olarak seçildiğinde, en yüksek doğruluk, kesinlik, EAKA, F

ölçütü ve Kappa değerlerine sahip sınıflama yöntemi DÖ olmuştur. fdr<0,02 olan genler seçildiğinde de DÖ yönteminin diğer yöntemlere göre daha yüksek değerlere sahip olduğu görülmektedir. fdr<0,05 genler seçildiğinde ise doğruluk, EAKA ve F ölçütünün en yüksek değerlere sahip olduğu yöntem DVM olmuştur. DÖ, DVM'nin ardından ikinci en büyük değerlere sahip olup, Kappa değeri DVM'den daha yüksektir. Akciğer kanseri veri seti için GSA yöntemi kullanıldığında, üç farklı yanlış bulgu oranına göre hazırlanan üç veri setinde de tüm performans ölçüleri bakımından en yüksek değerlere sahip sınıflama yöntemin DÖ'dir. Doğruluk değerleri DESeq2 yöntemi ile elde edilen veri setlerinde%85,3 ile %96,4 arasında değişmekte iken, GSA yöntemi ile elde edilen veri setlerinde daha geniş bir aralıkta, %80,0 ile %97,3 değişmektedir. Benzer şekilde EAKA, kesinlik, duyarlılık, F ölçütü ve Kappa değerleri de DESeq2 yönteminde, GSA yöntemine göre daha dar aralıkta değer almaktadır. Akciğer kanseri ile ilgili tüm veri setlerinde performansı en düşük yöntem YSA'dır. DESeq2 yöntemi ve fdr<0,02 filtresi ile oluşturulan veri setin dışında diğer beş veri setinde DÖ en başarılı sınıflama performansına sahiptir. Performans ölçülerine bakıldığında, aynı filtreler dikkate alındığında genellikle DESeq2 yönteminin GSA'ya göre daha yüksek değerler aldığı görülmektedir. Bu veri setin için DESeq2 yönteminin, GSA'ya göre daha tercih edilebilir olduğu söylenebilir. DESeq2 yöntemiyle seçilen gen sayısının, GSA yöntemiyle seçilen gen sayısından yaklaşık beşbin fazla olması nedeniyle bu durum ortaya çıkmış olabilir.

Böbrek kanseri veri seti, dengesiz sınıf dağılımına sahip olduğu için genellikle F ölçütü ve Kappa değerlerinin yorumlanması üzerinde durulmuştur. DESeq2 yöntemi ve fdr<0,01 filtresi seçildiğinde elde edilen veri setinde, çalışmada hesaplanan tüm performans ölçüleri için en başarılı performansı DÖ yöntemi göstermiştir. fdr<0,02 olduğu durumda ise DVM en yüksek değerlere sahiptir ve RO onu izlemektedir. fdr<0,05 filtresi kullanıldığında elde edilen veri setinde en başarılı performans, fdr<0,02'de olduğu gibi DVM'dir. DÖ yönteminin ise ikinci sırada yer aldığı söylenebilir. GSA yöntemi ve fdr<0,01 filtresi seçildiğinde en yüksek Kappa ve F ölçütü değerlerine sahip yöntem DVM'dir. fdr<0,02 iken DÖ tüm performans ölçülerinde en yüksek değerlere sahiptir. fdr<0,05 olduğu durumda en yüksek F ölçütü değerine sahip yöntem DVM iken, DÖ'nün Kappa değeri diğer yöntemlerden daha büyüktür. DESeq2 yöntemi ile oluşturulan veri setlerinde F ölçütü %79,5 ile %93,7

arasında, Kappa değeri %72,6 ile %93,8 arasında değişmektedir. GSA yöntemi kullanılarak oluşturulan veri setlerinde F ölçütü %79,1 ile %95,3 arasında, Kappa değeri %70,4 ile %94,5 arasında değer almaktadır. Akciğer kanseri veri setlerinde olduğu gibi, böbrek kanseri veri setlerinde de performans ölçülerinin en düşük değerleri aldığı sınıflama yöntemi YSA yöntemidir.

Akciğer kanseri veri setinde, DESeq2 yöntemi kullanıldığında, farklı fdr filtrelerinde en yüksek değeri alan EAKA, doğruluk ve F Ölçütü ölçülerinin birbirlerine yakın olduğu görülmektedir. Her üç filtre içinde en yüksek doğruluk değerleri %96,4'tür. EAKA değerleri %96,4 ile %96,5; F Ölçütü ise %96,4 ile %96,5 arasında değişmektedir. Bu ölçüler bakımından, fdr filtrelerinin arasında fark olmadığı söylenebilir. Kappa değerlerine bakıldığında, en yüksek Kappa değeri %96,4 ile fdr'nin yüzde ikiden küçük olduğu ve DÖ kullanıldığı duruma aittir. GSA yöntemi kullanıldığında ise, EAKA, doğruluk ve F Ölçütü değerlerinin daha geniş bir aralıkta değer aldığı görülmektedir. Hem EAKA, doğruluk, F Ölçütü hem de Kappa değeri bakımından en yüksek değerler DÖ yöntemi ve fdr<0,02 iken alınmıştır. Böbrek kanseri veri setinde, DESeq2 yöntemi uygulandıktan sonra fdr filtresi uygulandığında, en yüksek EAKA, doğruluk ve Kappa değerleri fdr<0,01 filtresi ve DÖ yöntemi uygulandığında elde edilmiştir. GSA yöntemi uygulandığında ise ise en yüksek doğruluk, F Ölçütü ve Kappa değerlerine f<0,01 filtresi ve DVM yöntemi ile sınıflandırma işlemi gerçekleştiğinde ulaşılmıştır. F Ölçütü ve Kappa ölçüleri bakımından, dengeli dağılan akciğer kanseri veri setlerine ait değerlerin, dengesiz dağılan böbrek kanseri veri setlerine göre daha yüksek değerler aldığı görülmektedir. Sınıflardaki gözlem sayılarının dışında, bu durumun bir diğer nedeni de filtrelere göre seçilen genlerin sayısı olabilir. Akciğer veri setinde bulunan genler filtrelendiğinde, elde edilen yeni veri setlerinde gen sayıları 10190 ile 15861 gen arasında değişirken böbrek kanserinde gen sayıları 6072 ile 10873 arasında değişmektedir. Böbrek kanseri veri setlerindeki gen sayılarının daha düşük olmasının da sınıflandırma performanslarını etkilediği düşünülmektedir.

Çalışma sonuçlarına bakıldığında, RNA dizileme verilerinin sınıflandırılmasında DVM'nin başarılı bir performans gösterdiği söylenebilir. Danaee ve ark. (76) TCGA'dan elde ettikleri, iki sınıflı bir veri setini veri madenciliği yöntemleri kullanarak sınıflandırmışlardır. Kullandıkları veri setinde meme kanserine

sahip 1097 bireye ve 113 sağlıklı bireye ait gen ifade seviyeleri bulunmaktadır. Bu çalışmada, sınıflamada kullanılacak en anlamlı genlerin seçilmesi ile veri setlerinde bulunan gen sayılarını azaltmak için farklı yöntemler kullanılmıştır. Bu yöntemlerden biri, boyut azaltma amacıyla sıklıkla kullanılan Yığınlı Otomatik Kodlayıcılardır. Farklı olarak ifade edilmiş genlerin bulunmasında kullanılan bazı özel yöntemler ve temel bileşenler analizi de, aynı veri setine boyut azaltma amacı ile uygulanmıştır. Ardından farklı yöntemlerle seçilen genler kullanılarak YSA ve iki farklı çekirdek fonksiyonu (doğrusal ve RTF) için DVM kullanılarak sınıflama gerçekleştirilmiştir. Çalışmanın sonucunda en başarılı sınıflama performansının Yığınlı Otomatik Kodlayıcılar kullanılarak boyut azalttığında ve RTF çekirdek fonksiyonlu DVM kullanıldığında elde edildiği görülmüştür. Zararsız ve ark. (77) RNA dizileme verileri olan iki sınıflı rahim ağzı kanseri, Alzheimer ve akciğer kanseri veri setleri ile üç sınıflı böbrek kanseri veri setini çalışmalarında kullanmışlardır. Bu dört farklı veri setinin her biri DVM, bagDVM (*bootstrap* yöntemi ile DVM yönteminin birleşimi olan bir yöntem), CART, RO ve olasılıksal doğrusal diskriminant analizleri kullanılarak sınıflandırılmıştır. Performans ölçülerinden biri olan doğruluk bakımından, tüm veri setlerinde en yüksek değere sahip sınıflandırıcılar DVM ve bagDVM olarak bulunmuştur.

Urda ve ark. (63) yaptıkları çalışmada invaziv meme, kolon ve böbrek kanseri olmak üzere üç farklı kanser türüne ilişkin RNA-dizileme verisi kullanmışlardır. Veri setlerinde yer alan bireylere ilişkin hayati durumlar (hayatta ve öldü olarak) sınıf etiketi olarak atanmış ve her veri seti için iki sınıf değeri elde edilmiştir. Bireylerin sınıflara atanmasının modellenmesinde, öncelikle boyut azalma işlemi iki farklı yöntem kullanarak gerçekleştirilmiştir. Bu yöntemlerden ilki LASSO regresyon kullanılarak boyut azaltılmasıdır. Bu yöntem kullanılarak birinci veri setinde bulunan 20021 adet gen 286 gene, ikinci veri setinde bulunan 19467 adet gen 70 gene, üçüncü veri setinde bulunan 20144 adet gen 269 gene düşürülmüştür. Boyut azaltmada kullanılan ikinci yöntem ise farklı ifade edilmiş genlerin, p değeri 0,001'den az olanlar olarak seçilmesidir. Bu yöntemle birinci veri setinde 242, ikinci veri setinde 37, üçüncü veri setinde 202 gen seçilmiştir. Ardından bu veri setlerindeki hayati durumların tahmin edilmesinde LASSO regresyon ve Derin Sinir Ağları yöntemleri kullanılarak EAKA hesaplanmıştır. Elde edilen EAKA değerleri 0,57 ile 0,77 arasında

değişmektedir. LASSO regresyon ve DÖ modellerine ait EAKA değerleri birbirlerine oldukça yakın çıkmıştır. Bu tez çalışmasında, DÖ modellerinin karmaşık yapılar üzerindeki etkisinin görülmesi istenildiği için, gen sayıları daha esnek filtreler kullanılarak belirlenmiştir. Bu nedenle, örnekte verilen çalışmanın aksine gen sayılarının binlerde olması tercih edilmiştir. RNA dizileme verilerinin DÖ ve klasik veri madenciliği yöntemleri kullanarak sınıflandırılmasını konu alan bir çalışmaya alanyazında rastlanmamıştır.

6. SONUÇ VE ÖNERİLER

Bu çalışmada dengeli ve dengesiz sınıf dağılımlarına sahip olan transkriptom veri setlerinde, farklı filtreleme yöntemleri kullanarak klasik veri madenciliği ve DÖ yönteminin performanslarının karşılaştırılması amaçlanmıştır.

- RNA dizileme gibi büyük ölçekli verilerin karmaşık yapısının yakalanmasında hem dengeli hem dengesiz sınıf dağılımlarında ve her filtrede tek gizli katmanlı YSA modelinin, çok katmanlı DÖ yöntemine göre daha düşük bir sınıflama performansına sahip olduğu görülmüştür.
- RO ve DVM yöntemleri, YSA'ya göre daha iyi performans göstermişlerdir.
- Performans ölçülerine bakıldığında, doğrusal ve doğrusal olmayan ilişkileri yakalamada başarılı bir yöntem olduğu belirtilen DVM'nin genellikle en yüksek sonuçları verdiği alanyazında yapılan çalışmalarda görülmüştür. Bu çalışmada da DVM en yüksek ya da ikinci en yüksek değerlere sahiptir.
- Dengeli sınıf dağılımına sahip olan akciğer kanseri veri setlerinde genellikle DÖ yönteminin DVM'den daha başarılı olduğu görülmüş, dengesiz sınıf dağılımlarına sahip böbrek kanseri veri setlerinde ise bazı filtrelerde DVM'nin, bazılarında DÖ'nün daha başarılı olduğu görülmüştür.

Genel olarak bakıldığında, klasik veri madenciliği yöntemlerine göre RNA dizileme gibi karmaşık yapılı verilerin sınıflandırılmasında DÖ yöntemi daha başarılı sonuçlar vermiştir ve kullanılması önerilmektedir.

İleride yapılacak çalışmalarda hem genetik alanında hem de farklı alanlarda, çeşitli DÖ algoritmalarının kullanılması amaçlanmaktadır. Bu çalışmada CNTK yazılım kütüphanesi ile çalışılmıştır. TensorFlow, Keras gibi farklı DÖ yazılım kütüphaneleriyle de çalışılarak, bu platformların hız, esneklik, sınıflama başarısı gibi parametreler yönünden karşılaştırılması amaçlanmaktadır.

7. KAYNAKLAR

1. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM ve ark. Criteria for The Use of Omics-Based Predictors in Clinical Trials. Nature. 2013;502(7471):317-320.
2. Hasin Y, Seldin M, Lusis A. Multi-omics Approaches to Disease. Genome Biology. 2017:18(1):83.
3. Karahalil B. Overview of Systems Biology and Omics Technologies. Current Medicinal Chemistry. 2016;23(37):4221-4230.
4. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA ve ark. Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences. Briefings in Bioinformatics. 2018;19(2):286-302.
5. Tanman Zıplar Ü, Cansaran Duman D, Türktaş M. Genomic and Transcriptomic Sequencing and Analysis Approaches. Middle Black Sea Journal of Healt Science. 2018;4(1):34-42.
6. Wang Z, Gerstein M, Snyder M. RNA-Seq: A Revolutionary Tool For Trancriptomics. Nat Rev Genet. 2009;10(1):57-63.
7. Ishii N, Ozaki K, Sato H, Mizuno H, Saito S, Takahashi A ve ark. Identification of a Novel Non-coding RNA, MIAT, That Confers Risk of Myocardial Infarction. J Hum Genet. 2006;51:1087–99.
8. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ ve ark. Long Non-coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis. Nature. 2010;464:1071–6.
9. Moran I, Akerman I, van de Bunt M, Xie R, Benazra M, Nammo T ve ark. Human Beta Cell Transcriptome Analysis Uncovers lncRNAs That Are Tissuespecific, Dynamically Regulated, and Abnormally Axpressed in Type 2 Diabetes. Cell Metab. 2012;16:435–48.
10. Knoll M, Lodish HF, Sun L. Long Non-coding RNAs As Regulators of The Endocrine System. Nat Rev Endocrinol. 2015;11:151-160.
11. Malone JH, Oliver B. Microarrays, Deep Sequencing and The True Measure of The Transcriptome. BMC Biology. 2011;9(1):34.
12. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M ve ark. The Transcriptional Landscape of The Yeast Genome Defined by RNA Sequencing. Science. 2008;320(5881):1344-1349.
13. Graves PR, Haystead TAJ. Molecular Biologist's Guide to Proteomics. Microbiol Mol Biol Rev. 2002;66(1):39–63.
14. Robinson P. RNA-seq Quantification and Differential Expression [İnternet]. Erişim adresi: http://www.mi.fu-berlin.de/wiki/pub/ABI/ GenomicsLecture13Materials/ rnaseq2.pdf.

15. Illumina. The Power of Replicates [İnternet]. Erişim adresi: https://www.illumina.com/Documents/products/technotes/technote_power_replicates.pdf.

16. Rapaaport F, Khanin R, Liang Y, Pirun M, Krek, Zumbe Paul ve ark. Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data. Genome Biology. 2013;14(9).

17. Love M, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. Genome Biology. 2014;15(12):550.

18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. Bioinformatics. 2010;26(1):139-140.

19. Anders S, Huber W. Differential Expression Analysis for Sequence Count Data. Genome Biology. 2010;11(10):R106.

20. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. Nucleic Acids Research. 2015;43(7):e47.

21. Dündar F, Skrabanek L, Zumbo P. Introduction to Differential Gene Expression Analysis Using RNA-Seq [İnternet]. 2018. Erişim adresi: http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf.

22. Chen JJ, Roberson PK, Schell MJ. The False Discovery Rate: A Key Concept in Large-Scale Genetic Studies. Cancer Control. 2010;17(1):58-62.

23. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. Bioinform Biol Insights. 2015;9(1):29-46.

24. Walczak JZ, Szabelska A, Handschuh L, Gorczak K, Klamecka K, Figlerowicz M ve ark. The Impact of Normalisation Methods on RNA-Seq Data Analysis. BioMed Research International . 2015.

25. Abbas-Aghababazadeh F, Li Q, Fridley BL. Comparison of Normalization Approaches for Gene Expression Studies Completed with High-Throughput Sequencing. Plos One. 2018;13(10).

26. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics. 2011;12:480.

27. Robinson MD, Oshlack A. A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. Genome Biology. 2010;11:R25.

28. Zwiener I, Frisch B, Binder H. Transforming RNA-Seq Data o Improve the Performance of Prognostic Gene Signatures. Plos One. 2014;9(1):e85150.

29. Love M. Rlog [İnternet]. Erişim adresi: https://www.rdocumentation.org/packages/ DESeq2/versions/1.12.3/topics/rlog.

30. Anders S, Huber W. Differential Expression Analysis for Sequence Count Data. Genome Biology. 2010;11:R106.

31. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression [version 1; peer review: 2 approved]. F1000Research. 2015;4:1070.
32. Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
33. Amit Y, Geman D. Shape Quantization and Recognition with Randomized Treees. Neural Computation. 1997;9:1545-1588.
34. Pal M. Random Forest Classifier for Remote Sensing Classification. International Journal of Remote Sensing. 2005;26(2):217-222.
35. Antipov EA, Pokryshevskaya EB. Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and A Cart-Based Approach for Model Diagnostics. Expert Systems with Applications. 2012;39(2):1772-1778.
36. Gupta B, Rawat A, Jain A, Arora A, Dhami N. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications. 2017;163(8):15-19.
37. Deng H, Runger G, Tuv E. Bias of Importance Measures for Multi-values Attributes and Solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). 2011.
38. James G, Witten D, Hastie T, Tibshiranı R. Support Vector Machines. Casella G, Fienberg S, Olkin I, editors. An Introduction to Statistical Learning with Applications in R. New York: Springer; 2013.
39. Lin HT, Lin CJ. A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods. Neural Computation. 2003.
40. Hsu CW, Chang CC, Lin CJ. A Practical Guide to Support Vector Classification. BJU International. 2008;101(1):1396-1400.
41. Tomar D. A Survey on Data Mining Approaches for Healthcare. International Journal of Bio-Science and Bio-Technology. 2013;5(5):241-266.
42. Mijwel M. Artificial Neural Networks Advantages and Disadvantages [İnternet]. 2018. Erişim adresi: https://www.researchgate.net/publication/323665827_Artificial_Neural_Networks_Advantages_and_Disadvantages.
43. Cilimkovic M. Neural Networks and Back Propagation Algorithm [İnternet]. 2010. Erişim adresi: http://dataminingmasters.com/uploads/studentProjects/NeuralNetworks.pdf.

44. Sharma V, Rai S, Dev A. A Comprehensive Study of Artificial Neural Networks. International Journal of Advanced Research in Computer Science and Software Engineering. 2012;2(10):278-284.
45. Elaraby NM, Elmogy M, Barakat S. Deep Learning: Effective Tool for Big Data Analytics. International Journal of Computer Science Engineering. 2016;5:254-262.
46. McCulloch WS, Pitts WH. A Logical Calculus of the Idead Immanent in Nervous Activity. Bulletin of Mathematical Biophysics. 1943;5:115-133.
47. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*. *1958;65*(6):386-408.
48. Fukushima K. A Self-Organizing Neural Network Model for A Mechanism of Pattern Recognition Unaffacted by Shift in Position. Biological Cybernetics. 1980;36(4):193-202.
49. Hinton GE, Osindero S, Teh YW. A Fast Learning Algorithm for Deep Belief Nets. Neural Computation. 2006;18(7):1527-2554.
50. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science. 2006;313:504-507.
51. Mohanty SP, Hughes DP, Salathe M. Using Deep Learning for Image-Based Plant Disease Detection. Frontiers in Plant Science. 2016;7.
52. Kaynak: Min S, Lee B, Yoon S. Deep Learning in Bioinformatics. Briefing in Bioinformatics. 2017;18(5):851-869.
53. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B ve ark. Deep Learning for Health Informatics. IEEE Journal of Biomedical and Health Informatics. 2017; 21(1): 4-21.
54. Suk HI, Shen D. Deep Learning-Based Feature Representation for AD/MCI Classification. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2013.
55. Hua KL, Hsu CH, Hidayati SC, Cheng WH, Chen YJ. Computer-Aided Classification of Lung Nodules on Computed Tomography Images via Deel Learning Technique. Onco Targets Ther. 2015;8:2015-2022.

56. Havaei M, Davy A, Warde-Farley D, Biard A, Courvilee A, Bengio Y ve ark. Brain Tumor Segmentation with Deep Neural Networks. Medical Image Analysis. 2017;35:18-31.
57. Cao Y, Liu C, Liu B, Brunette MJ, Zhang N, Sun T ve ark. Improving Tuberculosis Diagnostics Using Deep Learning and Mobile Health Technologies among Resource-Poor and Marginalized Communities. IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies. 2016.
58. Asgari E, Mofrad MR. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PLoS One. 2015;10(11):e0141287.
59. Ramsundar B, Kearnes S, Webster D, Konerding D, Pande V. Massively Multitask Networks for Drug Discovery. International Conference on Machine Learning. 2015.
60. Ibrahim R, Yousri NA, Ismail MA, El-Makky N. Multi-Level Gene/MiRNA Feature Selection Using Deep Belief Nets and Active Learning. 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2014.
61. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. Clin Cancer Res. 2018;24(6):1248-1259.
62. Fakoor R, Nazi A. Using Deep Learning to Enhance Cancer Diagnosis and Classification. Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare. 2013.
63. Urda D, Montes-Torres J, Moreno F, Franco L, Jerez JM. Deep Learning to Analyze RNA-Seq Gene Expression Data. IWANN. 2017:50-59.
64. Sokolova M, Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing and Management. 2009;45:427-437.
65. Arie BD. Comparison of Classification Accuracy Using Cohen's Weighted Kappa. Expert Systems with Applications. 2008;34:825-832.
66. Alpar R. Spor, Sağlık ve Eğitim Bilimlerinden Örneklerle Uygulamalı İstatistik ve Geçerlik-Güvenirlik. 4. Baskı. Ankara: Detay Yayıncılık; 2016.
67. Tharwat A. Classification assessment methods: a detailed tutorial. 2018.

68. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning. 2001;45(2):171-186.
69. Gene-Specific Analysis [İnternet]. Erişim adresi: https://documentation.partek.com/ display/FLOWDOC/Gene-specific+Analysis.
70. The Microsoft Cognitive Toolkit [İnternet]. 2017. Erişim adresi: https://docs.microsoft.com/en-us/cognitive-toolkit/.
71. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.
72. Probst P, Wright MN, Boulesteix AL. Hyperparameters and Tuning Strategies for Random Forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018:e1301.
73. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [İnternet]. 2015. Erişim adresi: https://CRAN.R-project.org/package=e1071.
74. Mantovani RG, Rossi ALD, Vanschoren J, Bischl B, Carvalho ACPLF. To Tune or Not To Tune: Recommending When to Adjust SVM Hyper-Parameters via Meta-Learning. International Joint Conference on Neural Networks. 2015.
75. RBF SVM Parameters [İnternet]. Erişim adresi: https://scikit-learn.org/stable/ auto_examples /svm/plot_rbf_parameters.html.
76. Danaee P, Ghaeini R, Hendric DA. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. Pac Symp Biocomput. 2016;22:219-229.
77. Zararsız G, Goksuluk D, Korkmaz S, Eldem V, Zararsız GE, Duru IP ve ark. A Comprehensive Simulation Study on Classification of RNA-Seq data. PLoS One. 2017;12(8):e0182507.

8. EKLER

EK-1: Tez Çalışması Orijinallik Raporu

Transkriptom Veri Seti Üzerinde Derin Öğrenme Yöntemi İle Klasik Veri Madenciliği Yöntemlerinin Sınıflama Performanslarının Karşılaştırılması

ORİJİNALLİK RAPORU

%4	%3	%1	%3
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	%1
2	medium.com İnternet Kaynağı	%1
3	Submitted to Eskisehir Osmangazi University Öğrenci Ödevi	<%1
4	faculty.washington.edu İnternet Kaynağı	<%1
5	Submitted to Afyon Kocatepe University Öğrenci Ödevi	<%1
6	Submitted to Gumushane University Öğrenci Ödevi	<%1
7	Submitted to Mugla University Öğrenci Ödevi	<%1
8	ceb-institute.org İnternet Kaynağı	<%1

EK-2: Dijital Makbuz

turnitin

Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Merve Kaşıkcı
Ödev başlığı:	Transkriptom Veri Seti Üzerinde De...
Gönderi Başlığı:	Transkriptom Veri Seti Üzerinde De...
Dosya adı:	Tez_-_Merve_Kasikci_27.08.2019_...
Dosya boyutu:	1.74M
Sayfa sayısı:	70
Kelime sayısı:	13,689
Karakter sayısı:	92,426
Gönderim Tarihi:	27-Ağu-2019 07:02PM (UTC+0300)
Gönderim Numarası:	1164042465

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

TRANSKRİPTOM VERİ SETİ ÜZERİNDE DERİN ÖĞRENME YÖNTEMİ İLE KLASİK VERİ MADENCİLİĞİ YÖNTEMLERİNİN SINIFLAMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Merve KAŞIKCI

Biyoistatistik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2019

9. ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı **:** Merve Kaşıkcı

Doğum Yeri ve Tarihi : Ankara, 11.07.1994

İletişim Bilgileri **:** Hacettepe Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı, 06100 Sıhhiye Ankara, 0312 305 14 67.

Eğitim

Yüksek Lisans **:** Hacettepe Üniversitesi Biyoistatistik Anabilim Dalı (2016-)

Lisans **:** Hacettepe Üniversitesi İstatistik Bölümü (2012-2016)

Poster Sunumu

Kaşıkcı M, Coşgun E, Karabulut E. Comparison of classification performances on transcriptomics dataset. 40th Annual Conference of the International Society for Clinical Biostatistics. 14-18 Temmuz 2019, Leuven.