T.C.
ERCİYES ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

ÇOKLU OMİK VERİLERİNİN BİRLEŞTİRİLMESİNDE KULLANILAN YAKLAŞIMLARIN VE SINIFLANDIRMA YÖNTEMLERİNİN PERFORMANSININ ARAŞTIRILMASI

Hazırlayan
Funda İPEKTEN

Danışman
Doç. Dr. Gökmen ZARARSIZ

Yüksek Lisans Tezi

Ağustos 2020
KAYSERİ

T.C.
ERCİYES ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

ÇOKLU OMİK VERİLERİNİN BİRLEŞTİRİLMESİNDE KULLANILAN YAKLAŞIMLARIN VE SINIFLANDIRMA YÖNTEMLERİNİN PERFORMANSININ ARAŞTIRILMASI

(Yüksek Lisans Tezi)

Hazırlayan
Funda İPEKTEN

Danışman
Doç. Dr. Gökmen ZARARSIZ

Bu çalışma; Erciyes Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından TYL-2019-9600 kodlu proje ile desteklenmiştir.

Ağustos 2020
KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu tezin kendi çalışmam olduğunu, tüm bilgilerin akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda akademik ve etik kuralların gerektirdiği gibi tüm materyal ve sonuçları tam olarak aktardığımı, başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel kurallara uygun olarak atıfta bulunduğumu ve kaynaklar listesinde gösterdiğimi belirtirim.

Adı-Soyadı: Funda İPEKTEN

YÖNERGEYE UYGUNLUK ONAYI

"Çoklu Omik Verilerinin Birleştirilmesinde Kullanılan Yaklaşımların ve Sınıflandırma Yöntemlerinin Performansının Araştırılması" adlı Yüksek Lisans tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi 'ne uygun olarak hazırlanmıştır.

Tezi Hazırlayan	Danışman
Funda İPEKTEN	Doç. Dr. Gökmen ZARARSIZ

Halk Sağlığı Anabilim Dalı/Biyoistatistik Bilim Dalı Başkanı

Prof. Dr. Ahmet ÖZTÜRK

Doç. Dr. Gökmen ZARARSIZ danışmanlığında **Funda İPEKTEN** tarafından hazırlanan **"Çoklu Omik Verilerinin Birleştirilmesinde Kullanılan Yaklaşımların Ve Sınıflandırma Yöntemlerinin Performansının Araştırılması"** adlı bu çalışma jürimiz tarafından Erciyes Üniversitesi Sağlık Bilimleri Enstitüsü Halk Sağlığı Anabilim Dalı Biyoistatistik Bilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

24/08/2020

JÜRİ:

Danışman :Doç. Dr. Gökmen Zararsız

(Erciyes Üniversitesi Biyoistatistik AD)

Üye :Prof. Dr.Ahmet Öztürk

(Erciyes Üniversitesi Biyoistatistik AD)

Üye :Doç. Dr. Halef Okan Doğan

(Cumhuriyet Üniversitesi Biyokimya Temel Tıp Fak.)

ONAY:

Bu tezin kabulü Enstitü Yönetim Kurulunun …… tarih ve ….. sayılı kararı ile onaylanmıştır.

…./…../…..

Prof. Dr. Bilal AKYÜZ

Enstitü Müdürü

TEŞEKKÜR

Bu tezin oluşmasında emekleri olan ve araştırmalarımın her aşamasında bilgi ve deneyimini esirgemeyerek akademik ortamda olduğu kadar insani ilişkilerde de desteğiyle ilerlememe katkıda bulunan danışmanım Doç. Dr. Gökmen Zararsız'a

Tez sürecimde bilgi ve deneyimlerini esirgemeyerek her anımda yanımda olan Prof. Dr. Ahmet ÖZTÜRK'e

Araştırma sürecimde bana destek olan bilgi, tecrübe ve manevi desteğini esirgemeyen Prof. Dr. M. Betül AYCAN'a

Yine araştırma sürecimde bana destek olan bilgi ve tecrübelerini benimle paylaşan Doç. Dr. Rachel CAVILL'e,

Tez çalışma sürecimde bilgi ve tecrübeleriyle birlikte manevi desteklerini esirgemeyen Dr. Dinçer GÖKSÜLÜK'e, Dr. Gözde ERTÜRK ZARARSIZ'a, Dr. Elif ÇELİK'e, Meltem ÜNLÜSAVURAN'a ve Cem SÖNMEZ'e,

Bu tez çalışmasına maddi destek veren Erciyes Üniversitesi Bilimsel Araştırma Projeleri Birimi'ne (Proje No: TYL-2019-9600),

Hayatımın her anında yanımda olan kız kardeşim olarak gördüğüm Ceren TAŞDEMİR'e ve her zaman manevi desteğiyle yanımda olan Merve Nur AYTAŞ'a ve Emrah KAYA'ya,

Çalışmalarım süresince birçok fedakârlıklar gösterip beni destekleyerek her an yanımda olan kıymetli teyzem Saime TAŞDEMİR ve ailesine

Hayatımın her döneminde desteklerini esirgemeyen bugünlere gelmem de çok önemli katkıları olan sevgi dolu aileme en derin duygularla teşekkürlerimi sunarım.

Funda İPEKTEN

Kayseri, Ağustos 2020

ÇOKLU OMİK VERİLERİNİN BİRLEŞTİRİLMESİNDE KULLANILAN YAKLAŞIMLARIN VE SINIFLANDIRMA YÖNTEMLERİNİN PERFORMANSININ ARAŞTIRILMASI

Funda İPEKTEN

Erciyes Üniversitesi, Sağlık Bilimleri Enstitüsü

Biyoistatistik Anabilim Dalı

Yüksek Lisans Tezi, Ağustos 2020

Danışman: Doç. Dr. Gökmen ZARARSIZ

ÖZET

Yaşam bilimlerinin birçok alanında başarıyla uygulanan omik teknolojilerinin gelişmesiyle birlikte binlerce metabolitin veya genin ekspresyon düzeyleri eş zamanlı olarak ölçülebilmekte ve uygun istatistiksel yöntemler kullanılarak hastalıklara ilişkin tanı ve teşhis yapılabilmektedir. Bu alanda kullanılan birleştirme ve sınıflandırma yöntemlerinin biyolojik sistemde kullanılabilecek modeller ve hastalığa ilişkin tanı ve tedaviler açısından önemli katkılar sağlayabileceği düşünülmektedir. Bu çalışmada çoklu omik veriler üzerinde veri birleştirme ve sınıflandırma yöntemlerinin performanslarının değerlendirilmesi amaçlanmıştır. Ham veriler ile mRMR ve PCA analizi kullanılarak oluşturulan veriler art arda, dönüşüm tabanlı ve model tabanlı birleştirme yöntemleri ile birleştirildi. Art arda birleştirme ve model tabanlı birleştirme yöntemleri kullanılarak birleştirilen veriler MKL, NSC, RF ve SVM sınıflandırma yöntemleriyle, dönüşüm tabanlı birleştirme yöntemiyle birleştirilen verilerde CANetwork, RVM ve Ada-boost RVM sınıflandırma yöntemleriyle hastalığın alt sınıfları tahmin edilmektedir. Gerçek veri setlerinden elde edilen uygulama sonuçları kullanılan veri ve yöntemlerin performanslarının birbirlerine göre üstünlüklerinde farklılık gösterebilmektedir. mRMR değişken seçimi kullanılan verilerin art arda birleştirme yöntemiyle birleştirilmesi ve bu verilerin MKL sınıflandırma yönteminin hastalıklarının alt tiplerinin tahmininde diğer yöntemlere nazaran daha doğru ayırabilmektedir. Ancak daha kapsamlı değerlendirmeler yapılabilmesi için benzetim çalışmaları ile desteklenmesi gerekmektedir.

Anahtar kelimeler: Çoklu omik, birleştirme yöntemleri, sistem biyolojisi, sınıflandırma yöntemleri

PERFORMANCE INVESTIGATION OF APPROACHES AND CLASSIFICATION METHODS USED IN INTEGRATION MULTI-OMICS DATA

Funda İPEKTEN
Erciyes University, Graduate School of Health Sciences
Department of Biostatistics
Master of Thesis, August 2020
Supervisor: Asst. Prof. Dr. Gökmen ZARARSIZ

ABSTRACT

With the development of omics technogoies successfully applied in many fields of life sciences, expression levels of thousands of metabolites or genes can be measured simultaneously and diagnosis of diseases can be made using appropriate statistical methods. It is thought that the combination and classification methods used in this field can make significant contributions in terms of models that can be used in the biological system and diagnosis and treatment of the disease. In this study, it was aimed to evaluate the performance of data integration and classification methods on multi-omic data. The raw data and generated using mRMR and PCA analysis were combined using concatenate integration, transformation based integration, and model based integration methods. Data integration with consecutive and model-based methods, MKL, NSC, RF and SVM methods, transformation based combination method, data integrated with CANetwork, RVM, Ada-boost RVM methods to estimate the subtype of the disease. Application results obtained from real data sets may differ in the superiority of the performances of the data and methods used. The mRMR feature selection can be combined with the concatenation of the data used and this data can be separated more accurately than other methods in the prediction of the subtypes of the diseases of the MKL classification method. However, it should be supported by simulation studies in order to make more comprehensive evaluations.

Key words: Multi-Omics data, Integration methods, Systems biology, Classification methods

İÇİNDEKİLER

KISALTMALAR ve SİMGELER

Ada-boost RVM	: Adaptive-boosting Relevance Vector Machine
ANN	: Artificial Neural Networks - Yapay Sinir Ağları
AUC	: Area Under The Curve - Eğri Altında Kalan Alan
CANetwork	: Composite Association Network - Bileşik İlişki Ağı
FS	: Feature Selection - Değişken Seçimi
GAW	: Genetic Analysis Workshop - Genetik Analiz Çalıştayı
GB	: Gradient Boosting - Gradyan Arttırma
GBM	: Gradient Boosting Machines - Gradyan Artırma Makineleri
GC-MS	: Gas Chromatography-Mass Spectrophotometry - Gaz Kromatografi-Kütle Spektrofotometri
kDa	: Kilodalton
LC-MS	: Liquid Chromatography- Mass Spectrophotometry - Sıvı Kromatografi-Kütle Spektrofotometri
MCC	: Matthew Correlation Coefficient - Matthew Korelasyon Katsayısı
miRNA	: Micro Ribonucleic Acit - Mikro Ribonükleik Asit
MKL	: Multiple Kernel Learning - Çoklu Çekirdek Öğrenimi
mRMR	: Minimum Redundancy-Maximum Relevance - Minimum Artıklık-Maksimum İlişki
NSC	: Nearest Shrunken Centroits - En Yakın Küçültülmüş Merkezler
PCA	: Principal Component Analysis - Temel Bileşen Analizi
PCR	: Principal Componenet Regression Analysis - Temel Bileşen Regresyon Analizi

PLS	: Partial Least Squares - Kısmi En Küçük Kareler
RF	: Random Forest - Rastgele Orman
RFE	: Recursive Future Elimination - Yinelemeli Değişken Seçimi
RNA-Seq	: RNA Sequence - RNA-dizilime
ROC	: Receiver Operating Characteristic - Alıcı İşlem Karakteristiği
RVM	: Relevance Vector Machine - İlişki Vektör Makinesi
SAM	: Significance Analysis Of Microarrays - Mikrodizilerin Anlam Analizi
simpleMKL	: Simple Multiple Kernel Learning - Basit Çoklu Çekirdek Öğrenimi
SVM	: Support Vector Machine - Destek Vektör Makinesi
TCGA	: The Cancer Genom Atlas - Kanser Genom Atlası
XGBoost	: eXtreme Gradient Boosting - Mutlak Gradyan Arttırma

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

1. GİRİŞ ve AMAÇ

Yüksek verimli analitik yaklaşımlar ve bilişim teknolojileri ve yüksek verimli analitik teknikler kullanılarak omik alanında yapılan ilk çalışmalarda bir dizi örneğin hedefli (targeted) ve hedefsiz (untargeted) yaklaşımlarda sadece tek bir omik teknolojisini kullanma eğiliminde olduğu bilinmektedir. İlk çalışmalarda tek bir omik teknolojisinde hedefsiz (untargeted) yaklaşımlar yaygın bir şekilde kullanılsa da kısa bir süre içerisinde tek bir analitik platformun biyolojik sistemler tarafından üretilen metabolitlere genel bir bakış sağlama olasılığının düşük olduğu öne sürülmektedir (Beale ve ark., 2018;Pinu, 2018). Bu nedenle, analitik ve veri işleme sistemlerindeki ilerlemelerle kolaylaştırılan analitik platformların kombinasyonunu kullanmak, yüzlerce metabolitin aynı anda güvenilir şekilde tanımlanması, saptanması ve sayısallaştırılmasında önemli olmaktadır (Pinu, 2016). Çoklu omik platformların kullanımı, sistem biyolojisini, hücresel metabolizmayı, hastalık etyolojisini daha ayrıntılı inceleme fırsatı sunmaktadır. Hastalık etyolojisinin moleküler karmaşıklığının çözümlemelerinde tek bir omik teknolojisini kullanmak yerine çoklu omik teknolojileri kullanarak daha güçlü ve kolay yorumlanabilir sonuçlar elde edilebilmektedir. Ancak tek bir omik verisinde kullanılan yöntemlerden elde edilen bilgiler, çoklu omik verilerinde kullanılan yöntemlerden elde edilen bilgilere göre sonuçların doğruluğu, güvenirliliği ve yorumlanması açısıdan nispeten yetersiz kalabilmektedir. Bu yüzden omik veri türlerinin (metabolomik, proteomik, transkriptomik gibi) artmasıyla çeşitli biyolojik problemlere ve zorluklara cevap verebilecek veri birleştirme yöntemlerinin geliştirilmesine ihtiyaç duyulmasına neden olmaktadır (Pruitt ve ark., 2012). Biyolojik sistemin temelinde yer alan her bir omik (metabolomik, proteomik, transkriptomik gibi) sayesinde hastalıkları etkileyebilecek metabolitlerin veya genlerin neler olabileceği

belirlenebilmektedir. Ancak tek bir omik ile tek bir yolak üzerindeki metabolit veya gen dikkate alınacağından asıl hastalıkla ilişkisi olabilecek diğer metabolitleri değerlendirmek mümkün olmamaktadır (Ramell ve ark., 2018). Ayrıca hastalıkların belirlenmesi açısından önemli olan metabolitlerin birbirleri ile olan etkileşimleri de dikkate alınamamaktadır. Bu problemlerin üstesinden gelebilmek için omik verilerin birleştirilmesi gibi çözümler üretilmektedir. Bu çözümler sayesinde araştırmacılar, tekli omik verilerin birleştirilmesi ile hastalıklara ilişkin tanı ve tedavilerde daha detaylı araştırmalar yapabilme imkânı bulabilmektedir. Aynı zamanda hastalığa ilişkin yanıt değişkenlerini daha iyi tahmin etmeye ve biyolojik düzenleme modelleri arasındaki boşluğu doldurmaya yardımcı olmaktadır. Ancak metabolomik ve diğer omik verilerinde örnek sayısının, değişken sayısından oldukça düşük olması, verilerin heterojen yapıda olmaları, yüksek çıktılı teknoloji kullanımı sonucunda verilerde kayıp (missing) gözlemlerin olması bu alanda kullanılabilecek geleneksel yöntemlerin kullanımını kısıtlamaktadır. Bu problemlerin üstesinden gelebilen ve araştırmalara dayalı biyolojik sistemin klinik olarak anlaşılması ve bu klinik bilgileri istatistiksel olarak çözümlemek için makine öğrenimi yöntemlerinin kullanımına ihtiyaç duyulmaktadır (Hastie ve ark., 2009).

Bu alanda yapılan araştırmalarda, makine öğrenimi yöntemlerinin kullanıldığını ve başarılı şekilde uygulandığını gösteren birçok çalışma bulunmaktadır. Örneğin çoklu çekirdek öğrenim (MKL) sınıflandırma yöntemiyle yumurtalık kanser verisi kullanılarak kanser tanısı konulan bireylerin üç yıldan daha uzun yaşayıp yaşamayacağı tahmin edilmektedir (Wilson ve ark., 2019). MKL yöntemiyle meme kanseri alt tiplerinin sınıflandırılmasında tekli ve çoklu omik veriler kullanılmaktadır (Tao ve ark., 2019). Bağırsak mikrobiyota üzerine yapılan bir çalışmada Rastgele Orman (RF) yöntemiyle hastalık durumunu tahmin etmek için metagenomikler ve metabolomikler entegre edilmektedir (Franzosa ve ark., 2019). Golub ve arkadaşları gen ekspresyon verilerini kullanarak lösemi hastalığının sınıflandırılmasında Destek Vektör Makine (SVM) yöntemini kullandılar (Golub ve ark., 1999). Kolon kanser dokularının sınıflandırmasında Moler ve arkadaşları SVM yöntemini kullandılar (Moler ve ark., 2000). Furey ve arkadaşları yumurtalık kanseri, lösemi ve kolon kanseri olmak üzere üç farklı veri setini kullanarak SVM yöntemini hastalıkların sınıflandırmasında kullandılar (Furey ve ark., 2000). Literatürde verilerin sınıflandırılmasında çok sayıda makine öğrenimi yöntemleri mevcut olduğundan dolayı omik verilerinde bu yöntemlerin

karşılaştırılması gerekmektedir. Bu yöntemlerin birleştirilmiş omik verilerinde kıyaslanması, örneklerin sınıflandırmasını optimize etmek, sistemlerin davranışlarını veya özelliklerini tahmin etmek için kullanılabilecek modelleri üretmek veya sınıflandırma performansı gibi durumlar açısından uygun olan yöntemlerin belirlenmesi gerekmektedir. Verilerin birleştirilmesinde model tabanlı, dönüşüm tabanlı ve artarda birleştirme tabanlı gibi bazı entegrasyon yaklaşımları kullanılmaktadır. Bu yaklaşımlardan elde edilecek bilginin daha doğru kullanılabilmesi açısından her bir yaklaşımın ayrı ayrı değerlendirilmesi gerekmektedir. Literatürde yeni olan bu yaklaşımların omik verilerinde uygulamalarının değerlendirilmesi biyolojik sistemde kullanılabilecek modeller ve hastalığa ilişkin tanı ve tedaviler açısından önemli katkılar sağlayabileceği düşünülmektedir.

Bu tezde çoklu omik verilerinde birleştirme ve sınıflandırma makine öğrenimi yöntemlerini hastalık sınıflarının tahmin edilmesi amacıyla gerçek verilerde uygulayarak en uygun yöntem(ler) kullanılması ve daha doğru sonuçlar elde edilmesi açısından hangi yöntemlerin tercih edilmesi gerektiği yönünde yeni fikirler sunulması amaçlanmaktadır.

2. GENEL BİLGİLER

2.1. Omik Verileri, Kullanım Alanları ve Çeşitleri

Omik veriler genetik, biyoloji, beslenme, ilaçbilim, toksikoloji gibi birçok alanda kullanılan ve başarıyla uygulanmış en yeni teknolojilerinden biridir. Nispeten yeni olmasına rağmen yıllarca birçok çalışmada omik teknolojilerinin faydaları öne sürülmekle birlikte eksik yönlerinin değerlendirilmesi açısından problemlere yönelik fırsatlardan da bahsedilmektedir. "Omik" ekinin moleküler bir terime eklenmesi, incelenen biyolojik sistemin bir bölümünü içeren bir dizi molekülün (genler veya proteinler gibi) daha kapsamlı bir şekilde değerlendirilmesini ifade etmektedir (Hasin ve ark., 2017). Bir dizi molekülden bazılarının özellikleri ayrı başlıklar altında aşağıda bahsedilmektedir.

2.1.1.Metabolomik

Canlı bir organizmanın dokularında ve hücrelerinde da yer alan 1 kDa-1.5 kDa'dan daha küçük herhangi bir metabolitin (aminoasitler, lipitler, vitaminler, ketonlar, organik asitler, polifenoller, karbonhidratlar, nükleik asitler) kütle spektrometri, moleküler spektroskopi veya kromatografik cihazlarla kantitatif olarak ölçülmesi ve tespit edilmesiyle; biyobelirteç keşfine, hücre büyümesi, hücre ölümü, hücresel metabolizma, toksik ajanların (elektro manyetik kirlenme, ilaçlar, endüstriyel gıdalar) fenotip üzerine etkisi ve kanserin erken teşhisi gibi durumlarla ilişkili sinyal moleküllerinin tanımlanmasına yönelik ayrıntılı bir şekilde çalışılmasına imkan sağlamaktadır (Rudaz, 2015).

2.1.2.Proteomik

Proteomik, belirli bir zaman diliminde (farklı gelişim evreleri, çevresel koşullar, yaşlılık gibi süreçler) organizmanın hücre kompartmanlarında ve farklı hücre tiplerinde ifade ettiği farklı proteinlerin miktarını, yapılarını, yerleşimlerini, translasyon sonrası modifikasyonlarını doku ve hücrelerdeki işlevlerini inceleyen dinamik bir yapı olarak

tanımlanmaktadır. Canlı organizmaların yapı taşı olan proteinler sinyalleme ve biyokimyasal yolakların bileşenleri olarak görev yaptıkları için, proteomik çalışmalarında önemli bir yer tutmaktadır. Dolayısı ile proteomik büyüme, gelişme ve çevre ile etkileşimlerin altında yatan moleküler mekanizmaları ortaya çıkarmak için kullanılmaktadır. Ayrıca metabolomik teknolojileri gibi proteomik de hastalık teşhislerinde, ilaç endüstri ve ilaç çalışmaları gibi araştırma alanlarında kullanılmaktadır. Proteomik bu çalışmalarda gen-protein-hastalık ilişkisini dikkate almasından dolayı önemini arttırmakta ve hastalık üzerine etkisi olan aday genlerin ekspresyon sonucunda ortaya çıkan proteinlerin teşhisi ile birlikte klinik uygulamalar için umut vaat eden bir teknoloji olmaktadır (Başaran ve ark., 2010).

2.1.3.Transkriptomik

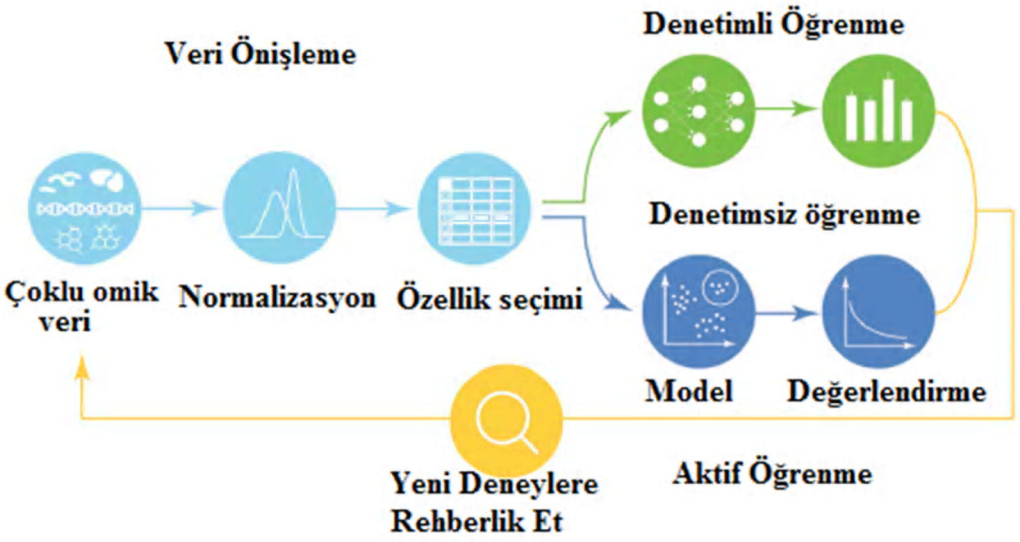
Transkriptomik canlı hücrede veya dokuda genom tarafından üretilen gen transkriptlerini (RNA) eş zamanlı inceleyen bir bilim dalıdır. Transkriptomik teknolojiler, farklı özelliklere sahip dokulardaki genlerin farklı zamanlarda ölçülmesi, genlerin nasıl düzenlendiği ve sistemin biyolojik ayrıntıları hakkında bilgi vermektedir. Bu bilgiler kullanılarak organizmalardaki gen ifade değişimleri hastalıkların anlaşılmasında etkili olmaktadır. Ayrıca farklı özellikteki hücrelerde transkriptomların karşılaştırılabilmesi, hastalıklara özgü tedavinin geliştirilmesinde de öncü olmaktadır (Golub ve ark., 1999).

2.1.4.Genomik

Genomik, organizmadaki genomları (gen topluluklarının kalıtımsal bilgisi) incelemek için kullanılan yöntemler olarak ifade edilmektedir. Genomik, genlerin her birinin ayrı tanımlanmasında, genlerin birbiriyle olan ilişkilerinin ve çevre ile etkileşimlerinin araştırılmasında ve zaman-yer-miktar olarak üretim ve aktivasyonlarının araştırılmasında etkin bir şekilde rol oynamaktadır (Başaran ve ark., 2010). Genomik bilimi ile canlı organizmaların evrimsel olarak benzerlikleri araştırılabilmekte, farklı organizmalara ait kalıtımsal bilgiler karşılaştırılabilmekte ve organizmaların ürettiği proteinlerin türleri, miktarları ve işlevleri hakkında fikir sahibi olunabilmektedir.

2.2. Omik Verilerinde Analizlerin Genel İş Akışı

Yüksek çıktılı teknolojilerden alınan çoklu omik verilerin analizlerinde makine öğrenimi yöntemlerinde yaygın bir şekilde kullanılmaktadır. Öğrenme yöntemleri ile elde edilen çoklu omik verilerine ait modellerin daha genellenebilir, öngörülebilir ve gürültülü değerlere karşı etkisiz hale getirmek için model eğitim ve değerlendirilmesinden önce çoklu omik veriler için normalleştirme ve değişken seçimi yapılmaktadır. Bu verilerin analizinin gerçekleştirilmesinde kullanılan iş akış şeması genel olarak şekil 2.1'deki gibi özetlenmektedir.



Şekil 2.1. Verilerin analiz iş akışı (Kim ve Tagkopoulos, 2018)

Literatürde veri akışında yer alan değişken seçimi aşaması için kullanılan Yinelemeli değişken seçimi, Boruta, Vita, Altmann ve minimum artıklık maksimum ilişki gibi birçok yöntem mevcuttur. Bu yöntemlerden Yinemeli değişken seçimi yöntemi, yüksek boyutlu veri kümesinde iyi bir tahmin modeli kurmak için minimum bir değişken kümesi bulmayı amaçlamaktadır (Diaz ve De Andres, 2016).

Başlangıçta bütün değişkenlerin bulunduğu bir karar ağaç yapısı oluşturmaktadır. Daha sonra değişkenlerin önem sırasına göre sıralanır ve en düşük öneme sahip olan değişkenler kaldırılarak kalan değişkenlerden yeni bir karar ağaç yapısı oluşturulur. Bu adımlar girdi olarak tek bir değişken kalana kadar yinelemeli olarak uygulanır. Her adımda tahmin performansı, model oluşturma için kullanılmayan out-of-bag olarak isimlendirilen örneklere dayalı olarak tahmin edilir. En küçük hataya sahip RF'ye göre değişkenler kümesi seçilir. Bununla birlikte, sıralamayı her adımda yeniden hesaplıyoruz çünkü bu değiştirilmiş algoritmanın ilişkili tahmin ediciler durumunda daha etkili olduğu gösterilmiştir (Gregorutti ve ark., 2013). Yinelemeli değişken seçim yöntemi, transkriptomik (Habermann ve ark., 2009), proteomik (Fusaro ve ark., 2009) ve metabolomik (Dietrich ve ark., 2016) verilerinde başarılı şekilde uygulanmıştır.

Slav mitolojisinde bir orman tanrısının adını alan Boruta yöntemi, sınıflandırma çerçevesinde yer alan tüm ilgili değişkenleri tanımlamak için geliştirilmiştir (Kursa ve Rudnicki, 2010). Bu yaklaşımın temeli, gerçek tahmin edici değişkenlerin önemini, istatistiksel test ve birkaç RF çalışması kullanarak rastgele yapay değişkenlerle karşılaştırmaktır. Her çalışmada, değişkenlerin kopyaları eklenerek gerçek tahmin edici değişkenler iki katına çıkarılmaktadır. Bu yapay değişkenlerin değerleri, gerçek değerlerin gözlemler arasında değiştirilmesiyle üretilir ve bu yüzden sonuçla olan ilişkiyi yok eder. RF, genişletilmiş veri setinde eğitilir ve değişken önem değerleri toplanır ve her bir gerçek değişken için, önemini tüm yapay değişkenlerin en büyük değeri ile karşılaştıran istatistiksel bir test gerçekleştirilir. Önemli ölçüde daha büyük veya daha küçük önem değerlerine sahip değişkenler, sırasıyla önemli veya önemsiz olarak belirlenir. Tüm önemsiz değişkenler ve yapay değişkenler kaldırılır ve önceki adımlar, tüm değişkenler sınıflandırılana veya önceden belirlenmiş sayıda çalıştırma gerçekleştirilene kadar tekrar edilir. Boruta yaklaşımı gen ifade ve mikrobiyom veri analizlerinde kullanılmıştır (Guo ve ark., 2014; Saulnier ve ark., 2011).

Parametrik olmayan yöntemlerden biri olan Altmann ise önemli değişkenlerin puanlarının sıfır dağılımına dayalı olarak *p* değerleri üretir. Permütasyon sayısını azaltmak için Altmann ve arkadaşları normal, lognormal veya Gamma gibi tanımlanmış bir olasılık dağılımını boş önem değerlerinin parametrik olmayan dağılıma uydurarak parametrik *p*-değerlerini kullanmayı önermektedirler. Bu dağılımların parametreleri, en çok olabilirlik yöntemleri kullanılarak tahmin edilir ve *p* değerleri, tahmin edilen

dağılımın altında gerçek önem puanından daha büyük bir önem puanının gözlemlenebilme olasılığı olarak hesaplanır. Altmann yaklaşımı, gen ifade ve mikrobiyom verilerinde kullanılmış ve diğer yöntemlere nazaran daha az tercih edilmektedir (Ji ve ark., 2014; Ning ve Beiko, 2015).

Minimum artıklık maksimum ilişki yaklaşımında $D = \{x_{i,k}\}_{nxK}$ veri matrisimizi belirtmektedir. Burada $x_{i,k}$ ifadesi k'ıncı örneğin i'inci değişkenin vektörünü, K örnek sayısını, n değişken sayısını göstermektedir. Veri setimizin -1 ve 1 olmak üzere iki sınıflı olduğu varsayılsın. Nihai değişken kümesi S^* olsun, *mRMR*(Minimum Redundancy Maximum Relevance) eşitliği minimum artıklık maksimum ilişki düzeyine sahip en iyi S^*'yi bulmaktadır.

$$\max_{S^*} \frac{\sum i \in S^*, h \in \{+1,-1\} I(h,i)}{\frac{1}{|S^*|}\sum_{i,j,\in S^*} I(j,i)} \qquad (2.1)$$

Eşitliğin üst kısmı, veri sınıfı ve değişken değeri arasındaki ilişki değerini ifade etmekte ve bu ifadenin olabildiğince büyük olması istenmektedir. Eşitliğin alt kısmı değişkenler arasındaki ortak bilgiyi göstermekte ve bu değerin küçük olması istenmektedir. $I(j,i)$ ise j ve i değişkeni arasındaki ortak bilgiyi göstermektedir. Hesaplama performansını iyileştirmek için mRMR yöntemi nihai S^*'yi oluşturmak için yinelemeli ve eklemeli bir yöntem kullanmaktadır. 2.1'de belirtilen denklem 2.2'deki denklemdeki forma dönüştürülür.

$$\max_{i \in \Omega_{S^*}} \frac{I(+1,i) + I(-1,i)}{\frac{1}{|S^*|}\sum_{j,\in S^*} I(i,j)} \qquad (2.2)$$

Burada Ω genel değişken seti ve ifadesi ile gösterilir $\Omega_S = \Omega - \mathrm{S}$. İki veri seti arasındaki ortak bilginin hesaplanması denklem 2.3'te gösterilmektedir.

$$I(i,j) = \sum_{i,j} p(x_i, y_j) log \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \qquad (2.3)$$

Burada $p(x,y), x$ ve y'nin koşullu olasılığı anlamına gelmektedir. Minimum artıklık maksimum ilişki yöntemi kesikli giriş verilerine ihtiyaç duymakta ve bu nedenle veri kümemize dağıtmak için işlev kullanılmaktadır. Veri seti $D = \{x_{i,j}\}_{nxK}$şeklinde tanımlanır daha sonra bu matrisin kesikli değeri denklem 2.4 olarak tanımlanır.

$$\tilde{x}_{i,j} = \begin{cases} 1 & x_{i,j} \geq \bar{x}_j + \alpha\sigma_{j,} \\ -1 & x_{i,j} \leq \bar{x}_j + \alpha\sigma_{j,} \\ 0 & diğerleri \end{cases} \qquad (2.4)$$

Burada σ_j, j'inci özelliğin standart sapmasını göstermekte, $\bar{x}_j$, j'inci özelliğin ortalama değeri ve α ise 0.5'e ayarlanmış bir parametredir. Sınıflandırma analiz sürecinde 2.4'te yer alan ifade ile oluşturulan değerler kullanılmamaktadır. Değişkenler seçildikten sonra, MKL yöntemi veya diğer yöntemlerle sınıflandırma analizi yapmak için gerçek değerler kullanılmaktadır (Zhang ve ark., 2014). Her bir değişken seçimi omik verilerde kullanılmaktadır. Ancak çoklu omik verilerinde yaygın olarak kullanılması ve iyi performans göstermesi açısından çalışmamızda mRMR yönteminin kullanılması tercih edilmektedir.

Ayrıca değişken seçimi aşamasında boyut indirgeme işlemleri de gerçekleştirilebilmektedir. Bu işlemlerden biri olan temel bileşen analizi, örnekler arasında ortak değişkenleri kullanarak veri setindeki varyansı azaltma da kullanılmakta ve büyük boyutlu veri kümesini daha küçük boyuta indirgemektedir. Bu sayede hastalıkla alakalı olabilecek genlerin daha doğru sınıflandırılmasında ayırıcı bir özellik katmaktadır. Birleştirilmiş veride değişken seçimi ve temel bileşen analizi kullanılmasında önemli bir nokta da her bir veri için kullanılan modellerin sonuçları birbiri ile örtüşmeyebilir. Bu sorunu ortadan kaldırmak için verilerin birleştirilmesi ve birbirleriyle ilişkilendirilebilmesi için önem taşımaktadır (Spicker ve ark., 2007).

Ayrıca veri boyutunu küçültürken gerekenden daha fazla bilgi kaybı yaşanmaması önemli bir durumdur. Veri kümesindeki varyasyon, saklamak istenilen bilgileri temsil etmektedir. Temel bileşen analizi, olabildiğince fazla varyasyonu korurken, verilerin boyutunu azaltmak için kullanılan matematiksel bir teknik olarak da ifade edilebilmektedir (Barshan ve ark., 2010). Bu istatistiksel yöntemin gerçekleştirilmesinde kullanılan basamaklar aşağıda bahsedilmektedir.

1. Veriler standartlaştırılır. Makine öğrenimi ve optimizasyon yöntemlerinin çoğu, tüm değişkenler aynı ölçekte olduğunda daha iyi performans göstermektedir. Standartlaştırma, her bir değişken değerinden, ortalamanın farkının alınması ve elde edilen farkın standart sapmaya bölünmesidir. Böylece ham veriler standart verilere dönüştürülerek, ölçü birimi farklılığı ortadan kaldırılmaktadır. Standartlaştırma da kullanılan eşitlik $x_s^i = \frac{x^i - \mu_x}{\sigma_x}$ şeklindedir. Burada μ_x değişkenlerin ortalamasını, σ_x örneklerin varyansını göstermektedir.

2. Veride yer alan değişkenlerin kovaryans matrisini hesaplanır. Kovaryans iki özelliğin birbirine göre nasıl bir değişim gösterdiğini ölçmektedir. Pozitif ya da negatif yönde değişimin artması veya azalması gibi değişimleri belirtmektedir. İki değişken vektörü x_j ve x_k için aralarındaki kovaryans σ_{jk} aşağıda yer alan formülle hesaplanabilmektedir.

$$\sigma_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_j^i - \mu_j)(x_k^i - \mu_k) \qquad (2.5)$$

Hesaplanan kovaryans matrisi Σ izleyen şekilde verilebilir.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix} \qquad (2.6)$$

Sütunlarda yer alan değişkenler standartlaştırıldığından her birinin ortalaması sıfıra eşit olduğundan kovaryans matrisi $\Sigma = \frac{1}{n-1}X^tX$ eşitliği ile hesaplanabilmektedir.

3. Kovaryans matrisinde öz değer ayrıştırma (eigen decomposition) yapılır. Özvektörler, kovaryans matrisinin temel bileşenlerini yani en büyük varyansı temsil etmektedir. Özdeğerler bu değerlerin büyüklüklerini göstermektedir. Özdeğeri en büyük değer olan özvektör en büyük varyansı temsil etmekte ve bir özvektör $\Sigma v = \lambda v$ eşitliğindeki koşulu sağlamalıdır. Eşitlikteki λ skaler bir özdeğerdir.

4. Özvektörleri, karşılık gelen özdeğerlerinin büyüklüğüne göre azalan sırada sıralanır.

5. Sıralama aşamasından sonra seçilecek en önemli temel bileşenlerin sayısı belirlenir. Boyut indirgeme için alınması gereken temel bileşen sayısı, özdeğerlerin

kümülatif toplamı hesaplanarak elde edilmektedir. Bu kümülatif toplam $\frac{\lambda_j}{\sum_{j=1}^{d} \lambda_j}$ ifadesi ile hesaplanmaktadır. Bu sayı belirlendikten sonra bir sonraki aşamada W matrisi oluşturulmaktadır.

6. W matrisi seçilen en uygun temel bileşen sayısından oluşturulur.

7. Yeni k boyutlu değişken matrisi elde edilir. W matrisini oluşturan temel bileşenler belirlendikten sonra Z skorları $Z = XW$ ifadesi ile bulunur.

Temel bileşen analizinde özdeğer ayrıştırmasının yapılabilmesi için girdi matrisinin kare matris olma zorunluluğu ve hesaplanmasının maliyetli olması gibi bazı zorlukları bulunmaktadır. Kısmen bu nedenlere bağlı olarak temel bileşen sayısının bulunmasında tekil değer ayrıştırması kullanılmaktadır (Meng ve ark., 2016). Özdeğer ayrıştırması $X = USV^*$ eşitliği ile bulunmaktadır. U, sol tekil vektörü, V^* sağ tekil vektörlerin bileşik eşleniği ve S tekil değerler olarak tanımlanmaktadır. Tekil değerler, özdeğer ayrışımından hesaplanan öz değerlerle ilişkilendirilir. Tekil değer ayrıştırmasının bir diğer özelliği, tekil değerlerin büyüklük sıralamasını yapıyor olması nedeniyle yeniden sıralama işlemi yapılmasına gerek duyulmamaktadır. Sağdaki tekil vektörler, özdeğer ayrıştırmasından hesaplanan öz vektörlerle aynı olduğu için $W = V$'dir (Bair ve ark., 2006).

2.3. Makine Öğrenimi Uygulamalarında Omik Verilerinin Kullanım Alanları

Makine öğrenimi yöntemleri sağlık, sosyal gibi birçok bilimde kullanılan verilerin analizinde kullanılmaktadır. Bu yöntemler LC-MS, GC-MS gibi yeni teknolojilerden elde edilen omik verilerinde de kullanılmakta ve kullanım alanı zamanla artış göstermektedir. Bu çeşitli kullanım alanlarına özgü örneklerden aşağıda bahsedilmektedir.

- Çoklu omik verilerde makine öğrenim yöntemleri birçok sektörde kullanılmakta ve geniş çapta etkiler yaratmaktadır. Tıbbi uygulamalarda, terapötik hedeflerin ve biyobelirteçlerin bulunması insan sağlığı için önemli bir yer tutmaktadır (Ahmad ve Fröhlich, 2016).
- Tıbbi uygulamada önemli olan bu konular ile çalışmalar gerçek dünyaya çevrilmektedir (Eagle, BERG, Genomiks). Küresel bir tehdit olarak kabul edilen antibiyotik direnci, klinik izolatların moleküler yapısında etkili antibiyotikleri

seçerek antibiyotik dirençlerinin tahmin edilmesinde makine öğrenim yöntemleri kullanılmaktadır (Davis ve ark., 2016).

- Biyoteknolojik uygulamalarda, belirli bir maddeden elde edilebilecek maksimum verimin alınmasında genetik ve düzenleyici süreçlerin optimizasyonu omik verilerinde eğitilmiş bir tahmin modeli ile sağlanmaktadır (Sweetlove ve ark., 2003).
- Tarımda, stres içerikli genlerin tespit edilmesi mahsulün elde edilmesinde büyük bir önem taşımaktadır. Makine öğrenim yöntemleri bu gen tespitlerinin daha hızlı yapılmasına imkân sunmaktadır (Shaik ve Ramakrishna, 2014; Ma ve ark., 2014).
- Gıda ve beslenme alanında makine öğrenim yöntemleri ile kişiye özgü beslenme ve tedavi geliştirmek için bireyin diyet bilgileri ve omik verileri kullanılmaktadır (Zeevi ve ark., 2015).
- Gıda mühendisliğinde ürünlerin fermantasyon süreçlerinde alınan genom kesit profilleri ile en iyi kalite de fermantasyon ürünü üretmek için de makine öğrenim yöntemleri kullanılmaktadır (Schwan ve Wheals, 2004).

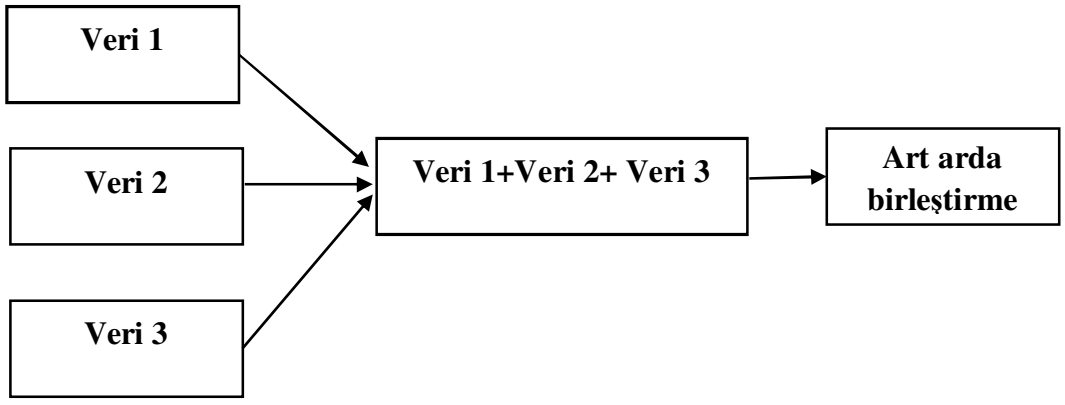
2.4. Omik Verilerinde Birleştirme Yöntemleri

Omik veri türlerinin (metabolomik, proteomik, transkriptomik) artmasıyla çeşitli biyolojik problemlere ve zorluklara cevap verebilecek veri entegrasyon yöntemlerinin geliştirilmesine ihtiyaç duyulmaktadır (Pruitt ve ark., 2012). Veri entegrasyon metodolojisinin temel amacı, herhangi tek bir veri setinden elde edilemeyen bilgileri elde etmek, bütün verileri aynı anda analiz etmek ve bütün verilerine bağlı olarak ortak modeller oluşturmaktır (Leung ve ark., 2016). Nispeten yeni ortaya çıkan entegrasyon yöntemleri, artarda birleştirme-tabanlı, dönüşüm-tabanlı ve model-tabanlı entegrasyon olmak üzere üç farklı başlıkta incelenebilmektedir.

2.4.1. Art arda Birleştirme (Concatenate-Based Integration)

Bütünleşik bir analiz yapmak için her bir omik verilerin ölçümlerini model oluşturmadan önce çok boyutlu matris olarak bir araya getirmektedir. Bu yöntemin avantajı tüm değişkenlerin matriste nasıl birleştirileceğini belirledikten sonra, nicel veya nitel verileri analiz etmek için çeşitli makine öğrenme yöntemlerinin kullanılmasının nispeten daha basit olmasıdır. Veri entegrasyonundaki problem, birleştirilecek olan verilerinin çok farklı boyutlara sahip olması durumunda ortaya çıkmaktadır. Mesela 100 metabolit, 10.000 transkript ile birleştirildiğinde, herhangi bir entegrasyon modelinde

gen verileri hakim olacaktır. Modeldeki bu problemi gidermeye yönelik, her omiğin eşit ağırlığa sahip olması için verilere (omikler) blok ölçeklendirme faktörleri uygulanması gerektiği önerilmektedir. Çok farklı boyutlara sahip olmasının yanı sıra, omik verileri, farklı yapılara, beklenen değerlerden farklı olmasına, gürültünün farklı dağılımlarına ve farklı varyanslara sahip olduğu, farklı teknolojilerden elde edilmektedir (Spicker ve ark., 2008). Sonuç olarak, basit bir entegrasyon verileri arasında bütünleştirici bağlantılar elde etmek kolay değildir Birleştirilen veri seti temel bileşen analizi, kısmi en küçük kareler (PLS) veya başka bir varyans maksimize eden yöntem tarafından işlenmesi yukarıda bahsedilen problemleri minimize etmektedir. Birleştirilmiş yaklaşımların uygulanması basit olmakla birlikte, belirtilen problemlerin sınırlandırılması, diğer yöntemlere nispeten bu yöntemin daha uygun olduğu söylenebilmektedir.

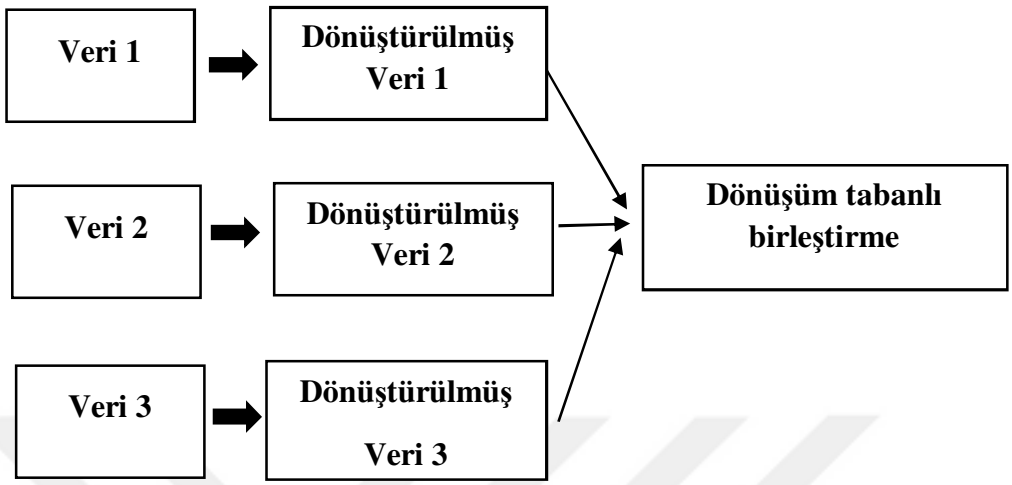


Şekil 2.2. Artarda birleştirme yönteminin iş akış şeması

2.4.2. Dönüşüm-Tabanlı Birleştirme (Transformation-Based Integration)

Dönüşüm tabanlı entegrasyonda, model oluşturmadan önce veri setini bir grafik veya çekirdek matrisi gibi bir ara forma dönüştürdükten sonra çoklu veri setlerini birleştiren bir yaklaşımdır. Bu yaklaşım veri türüne uygun bir haritalanma veya veri dönüşümü gerçekleştirmekte ve veri tipine özgü özellikleri koruyarak bir avantaj sunmaktadır. Diğer bir avantajı, veri tiplerini bağlayan hasta tanımlayıcıları gibi farklı özellikler içerdiği sürece, nicel veya nitel değerler ve sekans verileri dahil olmak üzere birçok veri tipinin entegre edilmesinde kullanılabilmesidir (Ritchie ve ark., 2015). Bu yöntemin

sınırlı yanı, bireysel genin fenotip üzerindeki etkisinin nasıl, ne şekil etkili olduğuyla ilgili detaylı yorumlanamamasıdır.



Şekil 2.3. Dönüşüm-tabanlı birleştirme yönteminin iş akış şeması

2.4.2.1. Dönüşüm-Tabanlı Birleştirme Yöntemlerinde Kullanılan Grafik Ve Çekirdek Tabanlı Yöntemler

İki sınıfa ait verilerin sınıflandırılması için birden fazla verinin entegre edilmesinde grafik ve çekirdek tabanlı yöntemler yaygın olarak kullanılmaktadır. Son zamanlarda, insan meme kanseri hücre dizilerinde ilaç duyarlılığı gibi sürekli bir sonucu tahmin etmek için grafik tabanlı ve çekirdek tabanlı yöntemlere başvurulmaktadır (Costello ve ark., 2014). Grafik veya çekirdek tabanlı yöntemler kullanılarak ham veriler öncelikle veri entegrasyon adımından önce örnekler arasında ilişkiler oluşturmak için eşlenir. Grafik tabanlı yöntemde örnekleri göstermek için düğümler (nodes), örnekler arasındaki ilişkileri de göstermek için kenarlar (edges) kullanılmaktadır. Dönüşüm tabanlı birleştirme yönteminde grafiksel şekillerden yararlanılmaktadır. Tek bir ağ üzerinden grafik tabanlı yöntem için (Zhou ve ark., 2004) n dizine sahip düğümlü bir G ağı olduğunu varsayalım. $Y_1, Y_2 \ldots, Y_p \; ve \; Y_i \in \{-1,1\}$ ve ikili sınıf için p düğümleri bilinen durum olarak belirlenir, kalan $n-p$ etiketsiz düğümler (bilinmeyen durum) olarak belirlenir. Grafik tabanlı öğrenmenin temel amacı etiketsiz düğümlerden faydalanarak düğümlerle ilgili ağ yapısını kullanarak sınıflandırma işlemini gerçekleştirmektir. Simetrik ağırlık matrisi (W) düğümler arasındaki ilişkileri temsil etmektedir. Simetrik ağırlık matrisi ilişkinin derecesini temsil etmekte ve bu değer

sıfırdan büyüktür ($w_{ij} \geq 0$). $w_{ij} = 0$ ise i ve j düğümleri arasında kenar olmadığını belirtmektedir. İki tane varsayımı bulunan algoritmanın çıktı fonksiyonu;

$f = (f_1, f_2, \ldots, f_n)^T$ şeklindedir. F fonksiyonunu bulmak için bu iki varsayım sağlanmalı. Varsayımlardan birincisi f_i değeri, y_i etiketli düğüm ile benzer olmalı, ikinci varsayım ise f_i değeri komşu düğümlerin değerine yakın olmalıdır. Bu varsayımlar sağlandıktan sonra f amaç fonksiyonu $\min_f \sum_{i=1}^{n}(f_i - y_i)^2 + c\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$ denklemi ile elde edilebilir. Denklemde $\sum_{i=1}^{n}(f_i - y_i)^2$ yer alan f_i değeri ile y_i gerçek değeri arasındaki farkın kareler toplamı kayıp fonksiyonun karesine eşittir. Bu ifade ise $\sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$ düzgünlük (smoothness) varsayımına karşılık gelmektedir. Denklemdeki c ise kayıp fonksiyona karşı düzgünlüğün kontrol edilmesinde kullanılan denge (trade-off) parametresi olarak kullanılmaktadır. Bu amaç fonksiyon $\min_f (f - y)^T(f - y) + cf^T Lf$ şeklinde yazılabilir. Burada $y = (y_1, y_2, \ldots, y_n)^T$ ve L, G ağının Laplace matrisi olarak ifade edilmektedir. $L = D - W$, $D = diag(d_i)$ ve $d_i = \sum jw_{ij}$ dir. En uygun çözüm $f = (I + cL)^{-1} y$ eşitliği ile elde edilebilir. Daha sonra etiketlenmemiş düğümleri ortanca kesim değeri ile tahmin etmek mümkün olacaktır. f_i fonksiyon değeri etiketli düğümlerin ortanca fonksiyon değerlerine daha yakın olduğunda $f_i = 1$, aksi takdirde -1 olarak sınıflandırılmaktadır. L matrisinin boyutu arttığında hesaplama süresi uzayabilir ve daha fazla belleğe ihtiyaç duyulabilir. Gerçekte L çok seyrek(sparse) olabilir, bu durumda büyük boyutlu ağlarda grafik tabanlı yöntemin uygulanması mümkün olmaktadır. Grafik tabanlı yöntemlerden biri olan CANetwork algoritmasında simetrik ağırlık matrisi (W_i) ile ilişkili m ağlarının ve kenar kuvvetlerini gösteren W_i elemanlarının hepsinin negatif olmadığını varsayalım. $y = (y_1, y_2, \ldots, y_n)^T$ ağlardaki düğümlerin sınıf etiket vektörünü gösteren ve y_i ikili bir değişken $y_i \in \{-1,1\}$. Hedef ağ (T) y'nin fonksiyonel ilişkileri olarak tanımlanmaktadır. Bu durumda T_{ij} için üç farklı değer bulunmaktadır.

$$T_{ij} = \begin{cases} (n_+/n)^2 & y_i = y_j = -1 \\ (n_-/n)^2 & y_i = y_j = 1 \\ (n_+n_-/n)^2 & y_i \neq y_j \end{cases} \qquad (2.7)$$

Formülde bulunan $n+/n-$ ifadeleri sınıf etiket vektöründe bulunan pozitif/negatiflerin toplam sayısını göstermektedir. Buradaki amaç m ilişkili ağları $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$

ağırlıkları ile entegre etmektir. Buradaki bileşik ağırlık matrisi $\overline{W} = \sum_{i=1}^{m} a_i W_i$ ifadesine eşit olmaktadır. Sezgisel olarak, T hedef ağında pozitif/negatif sınıf etiketli düğüm çiftleri yüksek benzerlik gösterirken, pozitif düğüm ve negatif düğüm çiftleri düşük benzerlik göstermektedir. $\underset{\alpha}{^{mintrace}}$ $((\overline{W} - T)^T (\overline{W} - T))$ ifadesinde T değerleri, bileşik ilişki ağlarının ağırlıklarını etkilemektedir. Bu ifadenin amacı hedef ağ T ile bileşik ağırlık matrisi ($\overline{W}$) arasındaki en küçük kareler hatasını en aza indirgemektir. $trace(AB) = vec(A)^T vec(B)$ fonksiyonunu yeniden yazdığımızda $\underset{\alpha}{^{min}}$ $(\Omega\alpha - \text{vec(T)})^T(\Omega\alpha - \text{vec(T)})$ $\Omega = [vecW_1 \ldots . vecW_m]$ ifadesi elde edilir. Optimal çözüm için α'nın türevini sıfıra eşitleyerek $\alpha = (\Omega^T\Omega)^{-1}(\Omega^T vec(T))$ elde edilebilir. Daha önce bahsedildiği gibi, hedef ağ T yalnızca üç değer alır, yani vec (T), çifte özgü ortak değişkenler olarak kabul edilebilir. Çalışmada üç kategorik değişken pozitif-pozitif, negatif-negatif ve pozitif-negatif olacak şekilde belirlenmektedir (Mostafavi ve ark., 2008). Bileşik ilişki ağı ile elde edilen ağırlıklar negatif olabilir. Bu durumdan kaçınmak için α_i sıfıra ayarlanır. İlişki ağları için ortalama ağırlıklar ($\alpha_i = 1/m$) tüm i'ler için $\alpha_i \leq 0$ olduğunda gerçek ağırlıkların üzerine yazılmaktadır. Uygulamada, α'ya sapma ağırlığı (α_0) eklenerek ve Ω'nın ilk sütunu bir ile doldurulmaktadır. İlişkilendirme ağlarının ağırlık matrisleri entegre edilirken de α_0 atılacaktır. Bileşik ağırlık matrisi ($\overline{W}$) elde edildiğinde, tek bir ağ için grafik tabanlı yöntem kullanılmaktadır. Fonksiyon değerleri, $f = (I + cL)^{-1}$ y formülüyle çözülebilir, ve burada L, ağırlık matrisi ($\overline{W}$) ile ilişkili Laplace matrisidir. Mostafavi ve Morris tarafında yapılan çalışmada olduğu gibi bileşik ilişki ağı için c 1 olarak alınmaktadır (Mostafavi ve Morris 2010). Çekirdek tabanlı yöntemlerden biri olan Uygunluk Vektör Makinesi (RVM), destek vektör makinesi ile aynı işlevsel forma sahip bir makine öğrenme tekniğidir, ancak olasılıklı sonuçlar elde etmek için Bayesci çıkarım kullanmaktadır (Tipping ME., 2001, 2003). Bir dizi girdi örneği verildiğinde $\{x_n\}_{n=1}^{N}$ ifadesine karşılık gelen çıktı $\{y_n\}_{n=1}^{N}$ ve $xn \in Rd \; ve \; yn \in \{-1,1\}$. RVM sınıflandırma modeli, çekirdek fonksiyonunun (k) doğrusal bir kombinasyonu olarak $Y(x;w) = \sum_{i=1}^{N} w_i k(x,x_i) = W^T K$ şeklinde yazılabilir. Burada $W = [w1, w2, \ldots, wN] \; ve \; K = [k(x,x1), k(x,x2), \ldots, k(x,xN)]$ eşittir. Son aşamada ise m örneklerin uygunluk noktaları olarak ayrılacak ve olasılık $P(y_i = 1/W) = \frac{1}{1+e^{-Y(x;w)}}$ sigmoid fonksiyonu ile hesaplanabilir. RVM'nin performansı SVM'ye çok benzerdir ancak RVM'nin sonucu SVM'den daha seyrektir(sparse) ve

çekirdek hesaplama süresi büyük ölçüde azaltılabilir. RVM, sınıf olasılıklarını kestirerek sınıflandırma problemleri için olasılıklı tahmin sağlayabilir. RVM'de, kayıp parametre şartı bulunmamaktadır. RVM'deki çekirdek fonksiyonu, Mercer'in koşul (Cortes ve Vapnik, 1995) kısıtlaması olmadan daha esnektir. K farklı verilerine karşılık gelen Y çıktısı ile ilişkilendirdiğini varsayalım. Burada $Y = (y_1, y_2, \dots, y_n)^T$ ve $y_i \in \{-1, 1\}$. Her bir veri kümesi için ilgili çekirdek matrisi ile ayrı bir RVM modeli oluşturulur. Birden fazla RVM modelinden elde edilen olasılık tahmin sonuçlarının k kümesi olarak $p_1, p_2, \dots, p_k$ verilmekte ve p_1 $nx1$ olan bir vektördür. Nihai sonuç için olasılık $\bar{P} = \frac{P_1+P_2+\cdots+P_k}{k} = (p_1, p_2, \dots., p_n^T)$ ifadesine eşit olmaktadır. $Y_i = 1$ olasılığının p_i olduğuna dikkat edilmelidir. Kesim noktası 0.5 olmalı yani $p_i > 0.5$ olduğunda örnek 1 sınıfına atanmaktadır. p_i ne kadar büyük olursa Y_i'nin 1 olarak sınıflandırılma şansı o kadar yüksek olmaktadır.

Bir diğer çekirdek tabanlı yöntem olan Ada-boost RVM, nihai sonuç performansını iyileştirmek için farklı öğrenici türlerini birleştiren bir makine öğrenimi yöntemidir. Son sınıflandırıcı, birden fazla zayıf öğrenicilerin ağırlıklı toplamından elde edilmektedir. RVM ile birleştirildiğinde şu adımlar takip edilmektedir. Eğitim örneklerinin $\{x_n\}_{n=1}^{N}$ ifadesine karşılık gelen çıktı $\{y_n\}_{n=1}^{N}$ ve $x_n \in R_d \; ve \; y_n \in \{-1,1\}$ olduğunu varsayalım. Eğitim örneklerinin ağırlıkları w_i=1/N olarak gösterilmektedir. İlk olarak t olarak belirtilen eğitim setinden değiştirilmeden seçilen n rastgele örnek üzerinde bir RVM modeli eğitilir, daha sonra verilen $\varepsilon_t = \sum_{i=1}^{N} w_i$ bu formüle göre t'inci yinelemede eğitim örneklerinde yanlış sınıflandırma için ağırlıklı hata hesaplanmaktadır. $\varepsilon_t \geq 0.5$ ise sonraki yinelemeye geçecek, aksi takdirde bu öğrenici RVM t'nin ağırlığına $\alpha_t = \frac{1}{2} ln(\frac{1-\varepsilon t}{\varepsilon t})$ eşit olacak şekilde ayarlanır ve nihai model $RVM_{son} = RVM_{son} + \alpha_t RVM_t$ şeklinde ifade edilir.

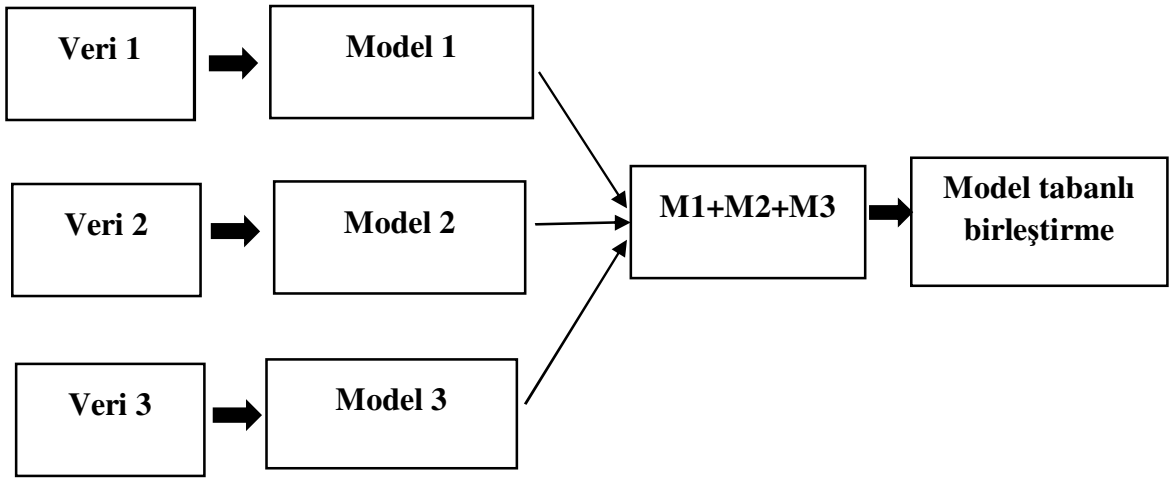
$$w_i = \begin{cases} w_i e^{\alpha t} & ise \; RVM_t(x_i) \neq y_i \\ w_i e^{-\alpha t} & ise \; RVM_t(x_i) = y_i \end{cases} \qquad (2.8)$$

Bir sonraki iterasyona geçmeden önce yeni ağırlıklar(w_i) $\sum_{i=1}^{N} w_i = 1$ formülü ile normalize edilir. T iterasyondan sonra nihai model $RVM_{son} = \sum ja_j RVM_j$ olarak elde edilir. Ve $\varepsilon_j < 0.5$'dir. RVM yönteminin hesaplama maliyeti yüksek olduğundan, RVM için Ada-boost kullanımı büyük ölçekli öğrenme sorununu çözebilir ve hesaplama maliyetini düşürebilir. Bu kullanımda temel yapı, gerçek eğitim setinden birçok küçük

eğitim setini örneklemek ve daha sonra her model daha küçük bir eğitim seti ile eğitilerek hesaplama maliyetini düşürmektir. Yeterli sayıda temel model oluşturulduğunda, tüm eğitim setinin farklı yönlerinin çoğu, nihai birleşik modelde yer alabilmektedir. Ada-boost RVM yöntemini kullanırken uygun bir yeniden örnekleme boyutunu ve maksimum yineleme sayısını belirlemek gerekir. Yeniden örnekleme boyutu ve yineleme sayısı için bir dizi değer, 5 kat çapraz geçerlilik ile değerlendirilir. Uygun sayı için yeniden örnekleme boyutunu ve maksimum yineleme sayısı araştırılır (yeniden örnekleme $\in \{0.2N, 0.4N, 0.6N, 0.8N\}$, yineleme sayısı $\in \{1, 5, 10, 20, 30\}$ buradaki N eğitim seti örneklem büyüklüğüdür).

2.4.3. Model-Tabanlı Birleştirme (Model-Based Integration)

Bu entegrasyonda farklı veri türlerinden eğitim (train) seti oluşturulmakta ve eğitim sırasında oluşturulan çoklu modellerin çıktılarının birleştirilmesi sonucunda bir model üretmektedir (Ritchie ve ark., 2015). Bu yöntemin avantajı belirli hastalık veya genotip ile ilişkili farklı omik yapılar arasındaki etkileşimi dikkate almaktadır. Diğer bir avantajı tek bir omik ile elde edilen tahmin modelinin performansı ile karşılaştırarak birleştirilmiş verinin daha iyi bir tahmin modeli oluşturup oluşturamadığını değerlendirme fırsatı sunmaktır. Son olarak çoklu omik verilerindeki tüm değişkenlerden en iyi modeli elde edebilmekte ve nihai model için dengeli bir doğruluk sağlamaktadır. Kolay yorumlanabilmektedir ancak verilerin bütünsel olarak görünümünü oluşturma açısından kısıtlılığı bulunmaktadır.



Şekil 2.4. Model tabanlı birleştirme yönteminin iş akış şeması. M1: Model 1, M2: Model 2, M3: Model 3

Tablo 2.1. Omik yapılarda makine öğrenimi yöntemlerinin kısa bir özeti

Yöntemler	Özet	Güçlü yönleri	Zayıf yönleri
Model-tabanlı birleştirme	Çoklu omik verileri ile çoklu kestirim modeller üretilir ve bu modeller kullanılarak nihai bir tahmin modeli oluşturulur.	Aynı fenotipe sahip çeşitli hasta gruplarından toplanan çoklu omik verileri birleştirilebilir.	Aşırı uyum (overfitting) problemini önlemek zor olabilir.
Art arda birleştirme	Büyük bir girdi matrisi oluşturulur ve bu matrisle tahmin modeli oluşturulur.	Nicel veya nitel verilerde çeşitli makine öğrenim yöntemlerini uygulamak kolaydır.	Büyük boyutlu verileri tek bir matriste birleştirmek güç olabilir.
Dönüşüm-tabanlı birleştirme	Çoklu omik verileri ilk önce girdi matrisinde birleştirilen ara formlara dönüştürülür ve girdi matrisi ile tahmin modeli oluşturulur.	Nicel veya nitel verileri birleştirmek için hasta tanımlayıcı gibi değişkenler de kullanılabilir.	Ara forma dönüştürmek zor olabilir.

(Lin ve Lane, 2017)

2.5. Omik Verilerinde Kullanılan Bazı Sınıflandırma Yaklaşımları

Makine öğrenimi, birçok farklı verilerin analiz edilmesinde kullanılan yöntemleri içermektedir. Bu yöntemler yeni teknolojilerden elde edilen omik verilerinde de kullanılmaktadır. Bu alanda kullanılan bazı yaklaşımlardan kısaca alt başlıklar halinde bahsedilmektedir.

2.5.1. En Yakın Küçültülmüş Merkezler (Nearest Shrunken Centroids)

NSC, yüksek boyutlu veriler için popüler bir sınıflandırma yöntemidir (Tibshirani ve ark., 2002). Köşegenel doğrusal ayırma analiz yönteminde yapılan bir takım düzenlemeler sonucunda Tibshirani ve arkadaşları tarafından önerilmektedir (Dudoit ve ark., 2002). NSC yönteminde, köşegenel doğrusal ayırma analiz yönteminden farklı olarak küçültme ve değişken seçim işlemleri uygulanmaktadır (Tibshirani ve ark., 2003). NSC yöntemi biyomedikal, biyokimya, biyoteknoloji gibi alanlarda da kullanılmaktadır. Bu yöntemin aşamaları aşağıda bahsedilmektedir.

1. X_{ij} değişken olmak üzere;
   $i = 1, 2, \dots, p$ genleri ve $j = 1, 2, \dots, n$ örnekleri göstermektedir.
2. $1, 2, \dots, K$ olmak üzere sınıf sayısını göstermektedir. C_k, k sınıfında bulunan n_k örneklerini göstermektedir.
3. $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij} \,/\, n_k$ olmak üzere i geni için k sınıfındaki ortalama değeri göstermektedir.
4. $\bar{x}_i = \sum_{j=1}^{n} {x_{ij}}/{n}$ olmak üzere i. genin genel ortalama değerini göstermektedir.

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k(s_i + s_0)} \qquad (2.9)$$

Denklem 2.9'da s_i, i geni için sınıf içi standart sapma değerini ifade etmektedir. Denklem 2.10'da her bir gen için sınıf içi standart sapma ile standartlaştırıldıktan sonra sınıf ortalamalarının değerlerini genel ortalama değerlerine doğru küçültme işlemi (Büzme işlemi olarak da ifade edilmektedir.) gerçekleştirilmektedir. Bu

standartlaştırma, aynı sınıftaki örnekler için kararlı olan gen düzeylerine daha fazla ağırlık vermektedir. Bu standartlaştırma yöntemi doğrusal ayırma analizi gibi yöntemlerin doğasında mevcut olarak kullanılmaktadır.

$$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2 \qquad (2.10)$$

$m_k = \sqrt{1/n_k + 1/n}$ ve denklem 2.9'da $m_k . s_i$ payın tahmini standart hatasına eşit olmaktadır. Denklem 2.9'da payda da yer alan s_0 tesadüfen düşük gen düzeylerinden tesadüfi olarak ortaya çıkan büyük d_{ik} değerlerin olasılığını kontrol altına almak için denkleme dâhil edilmektedir (s_0 pozitif sabit ve tüm genler için aynı değer). s_0, gen kümelerinde s_i'nin ortanca değerine eşit olarak alınmaktadır. Bu strateji benzer bir şekilde SAM metodolijisinde kullanılmaktadır (Tusher ve ark., 2001).

Yani, d_{ik}, i geni için k sınıfı ile genel ortalaması t istatistiği ile karşılaştırır. Denklem 2.9 yeniden düzenlenerek denklem 2.10'da verilmektedir.

$$\bar{x}_{ik} = \bar{x}_i + m_k(s_i + s_0)d_{ik} \qquad (2.11)$$

Her d_{ik} değeri 0'a doğru küçülmekte ve denklem 2.11'de verilen d_{ik} küçültülmüş ortalamaları göstermektedir.

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik} \qquad (2.12)$$

Küçültme işleminde Lasso yönteminden yararlanılmaktadır. Denklem 2.13'de d_{ik} mutlak değerinden delta (Δ) çıkarılarak azaltılır ve mutlak değeri sıfırdan küçük ise sıfıra eşitlenmektedir. Bahsedilen hesaplama 2.13 eşitliğinde gösterilmiştir.

$$d'_{ik} = sign(d_{ik})(|d_{ik}| - \Delta)_+ \qquad (2.13)$$

Denklem 2.13'te yer alan + pozitif kısım anlamına gelmektedir. Yani t sıfırdan büyükse $t_+ = t$ eşitken diğer durumlarda sıfıra eşit olmaktadır. Çoğu $\bar{x}_{ik}$ değerleri gürültülü ve genel ortalamaya $(\bar{x}_i)$ yakın olmasından dolayı Lasso yöntemi genellikle daha güvenilir

tahminler üretmektedir. Bu yöntemde küçültme parametresi (Δ) sıfıra yaklaştıkça sınıf tahmininde elimine edilen genlerin çoğu istenilen özelliği taşımaktadır. Özellikle i geni için tüm k sınıfları için d_{ik} sıfıra küçültülürse i geninin ortalama değeri $\bar{x}_i$ olmakta ve tüm sınıflar için aynı olmaktadır. Bu yüzden i geni en yakın ortalamanın hesaplanmasında katkısı bulunmamaktadır. Çapraz geçerlilik yöntemi ile küçültme parametresini (Δ) kendimiz seçmekteyiz. Örneğin bir örneklem kümesini 10 eşit parçaya bölerek 10 kat çapraz geçerlilik kullandığımızı varsayalım. Örneklerin %90'ı ile model eğitilmekte, %10'luk kısmı ile sınıflar tahmin edilmektedir. Bu durum 10 kez tekrar edilir ve toplam hatayı hesaplamak için 10 parçada ortaya çıkan her hata birlikte eklenir. Bu yöntemin detayları Tibshirani ve arkadaşlarının çalışmasında anlatılmaktadır (Tibshirani ve ark., 2003). Sonuç olarak bu küçültmenin iki avantajı bulunmaktadır. Birincisi gürültülü genlerin etkisini azaltarak sınıflandırma modelinin doğruluğunu arttırmak, ikincisi ise otomatik olarak gen seçimi yapmaktır.

2.5.2. Çoklu Çekirdek Öğrenimi (Multiple Kernel Learning)

MKL yöntemi, destek vektör makinesinin uygulanabilirliğini arttırmak için önceden belirlenmiş çekirdek kümesini kullanan ve algoritmanın bir parçası olarak en iyi doğrusal veya doğrusal olmayan çekirdek birleşimini gerçekleştiren bir yöntem olarak bilinmektedir (Kloft ve ark., 2011). Bu yöntem optimal çekirdek grubundan parametreler seçmekte, çekirdek seçiminden kaynaklanan yanlılığı azaltmakta, farklı kaynaklardan gelen verileri birleştirmekte kullanılmaktadır, ancak bu kullanımlar için farklı çekirdekler kullanılması gerekmektedir. Bu yöntem yeni bir çekirdek üretmek yerine her bir veri kümesi için önceden belirlenmiş çekirdekleri birleştirmek için çoklu çekirdek yöntemleri kullanmaktadır. Çekirdekleri birleştirmede kullanılan eşitlik basit haliyle denklem 2.14'de gösterilmektedir. M çekirdekleri doğrusal olarak tek bir çekirdekte birleştirilmektedir.

$$K\left(x, x' = \sum_{m=1}^{M} d_m K_m(x, x'), \quad d_m \geq 0, \sum_{m=1}^{M} d_m = 1\right) \qquad (2.14)$$

Denklem 2.14'te $K_m(x, x')$ farklı olan çekirdekleri göstermektedir. d_m ise $m.$ çekirdeğin ağırlığını göstermektedir. MKL yöntemi kullanılırken çekirdek ağırlıklarının

birçoğu sıfırdır. Çekirdeklerin çoğunluğunun sıfır olması seyrek (sparse) problem olarak adlandırılmaktadır. Bu durum da eğitim verisinde kullanılan çekirdeklerin modele etkisi bulunmamaktadır. Seyrek olmayan (Non-sparse) (Kloft ve ark., 2008) yöntemler geliştirilmiş, ancak bu yöntemde maliyet ve zaman gibi problemler görülmektedir. Bu ve daha önceki problemlerin çözümüne yönelik alternatif bir yöntem olan ve verimli bir yöntem basit çoklu çekirdek öğrenim (simpleMKL) yöntemi kullanılmaktadır. Diğer yöntemlerden daha iyi performans göstermektedir. Bu durumda ağırlıklandırılmış L2-norm düzeltmesine dayanan simpleMKL yönteminin diğer yöntemlerden daha iyi olduğunu ispatlamaktadır (Kloft ve ark., 2008). Yazar, maliyet ve zaman kaybını en aza indirgemek için destek vektör makine optimizasyonunu ve çekirdek birleştirme sürecini tek bir standart destek vektör makine optimizasyonuna entegre etmektedir. Fonksiyon denklem 2.15'te verilmektedir.

$$min_{f,b,\xi} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_i \xi_i$$

$$s.t.\ y_i(f(x_i)+b) \geq 1-\xi_i \quad \forall_i \qquad (2.15)$$

$$\xi_i \geq 0 \quad \forall_i,$$

Denklem 2.15'te $\|f\|_{\mathcal{H}}$ Hilbert uzayındaki çekirdeği K_m çekirdeği ile ilişkilendirmektedir. Bu eşitlik klasik çekirdek optimizasyon problem ikilisine eşit olmaktadır. Bu da denklem 2.16'da gösterilmektedir.

$$f(x) = \sum_{i=1}^{1} a_i^* K(x, x_i) + b^* \qquad (2.16)$$

Genel çekirdek farklı çekirdeklere ayrılabilmekte $\|f_m\|_{\mathcal{H}}$ ifadesi yerine $\sum_m \|f_m\|_{\mathcal{H}\mathrm{M}}$ ifadesi yazılır ve denklem 2.17'de gösterilmektedir.

$$min_{\{f_m\},b,\xi,d} \frac{1}{2}\sum_M \|f_m\|_{\mathcal{H}\mathrm{M}}^2 + C\sum_i \xi_i$$

$$s.t. \quad y_i \sum_m f_m(x_i) + y_i b \geq 1-\xi_i \quad \forall_i$$

$$\xi_i \geq 0 \quad \forall_i \qquad (2.17)$$

$$\sum_m d_m = 1, d_m \geq 0 \quad \forall_m$$

Yukarıdaki denklem Hilbert uzayındaki önemli çekirdeklerin 2-norm oluşumunda birleştiğini göstermektedir. Denklemin ayrıntıları Rakotomamonjy ve arkadaşlarının çalışmasında verilmektedir (Rakotomamonjy ve ark., 2008). Bu optimizasyon problem denklem 2.17'de matematiksel dışbükey optimizasyon yöntemi ile çözülebilmektedir. simpleMKL yöntemi bağımsız çekirdeklerden çekirdek kombinasyonunu kullanmak yerine destek vektör makine optimizasyon problemini entegre ederek hesaplama maliyetini büyük ölçüde azaltmaktadır. simpleMKL şeması en iyi parametreleri bulabilmek için azaltılmış bir gradyan algoritması kullanmaktadır (Zhang ve ark., 2016).

MKL için denetimli, denetimsiz ve yarı denetimli öğrenim yöntemleri kullanılmaktadır. Denetimli algoritmada kullanılan yaklaşımlar Sabit Kural (Fixed rules), Sezgisel (Heuristic), Optimizasyon (Optimization), Bayesci (Bayesian) ve Arttırıcı (Boosting) yaklaşımlar olmak üzere beş çeşit algoritma bulunmaktadır. Bu yöntemlere ilişkin detaylı bilgiler Gönen ve arkadaşının çalışmasında anlatılmaktadır (Gönen ve Alpaydın, 2011).

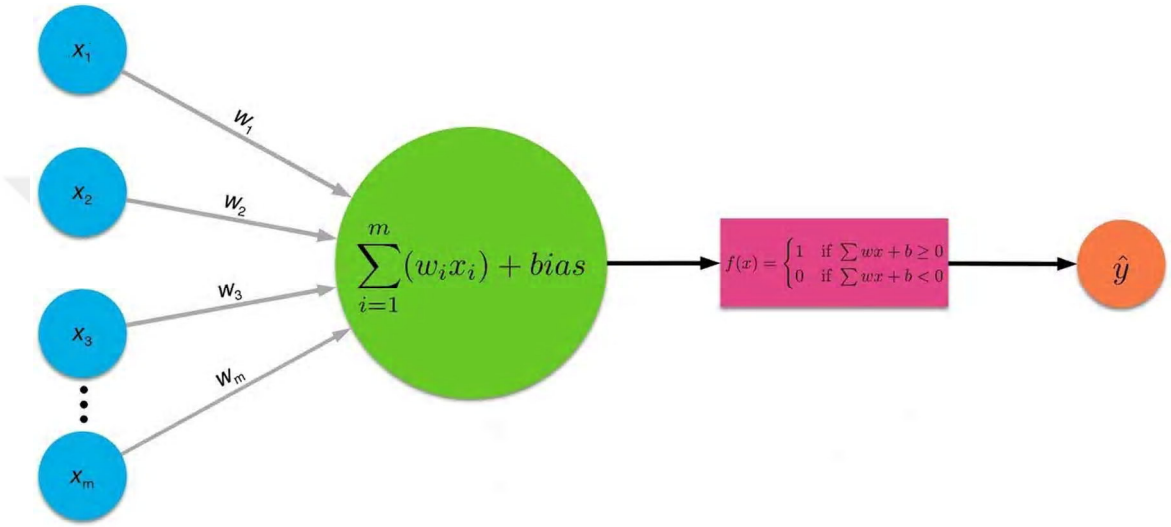
2.5.3. Yapay Sinir ağları (Artificial Neural Networks)

Yapay zeka makinelere düşünme, karar verme, karşılaştırma, analiz yapma gibi bir takım insana özgü özelliklerin entegre edilmesi olarak tanımlanmaktadır. Yapay zeka, ünlü İngiliz Matematikçi Alan Turing'in "Makineler Düşünebilir Mi? sorusu ile ortaya çıkmıştır. Yapay zeka uygulamalarından biri olan yapay sinir ağı, 1943 yılında Warren McCulloch ve Walter Pitts tarafından ilk yapay sinir ağ (nöron) modeli geliştirilmiştir (McCulloch ve Pitts, 1943). Bu çalışmadan yola çıkarak 1950-1960'lı yıllarda bilim adamı Frank Rosenblatt tarafından algılayıcılar (perceptrons) geliştirilmiştir. Algılayıcı(lar) bir veya birden fazla girdileri alır ve sonuç olarak ikili çıktı üretmektedir.

1980'li yıllarda Hopfield'ın çalışmalarında yapay sinir ağının genelleştirilebileceği ve özellikle geleneksel bilgisayar programlama ile çözülmesi zor olan problemlere çözüm üretilebileceği gösterilmektedir. ANN, insan beynindeki biyolojik sinir ağın çalışma şeklini taklit eden yöntem olarak bilinmektedir. Yani bir biyolojik nöronun (işlem elemanı) sinyalleri almak için dendritleri (toplama fonksiyonu), onları işlemek için bir

hücre gövdesi (aktivasyon fonksiyonu) ve diğer nöronlara sinyal göndermek için bir aksonu (eleman çıkışı) olduğu gibi, yapay nöronun bir dizi giriş kanalı, bir işlem aşaması ve yapay nöronlara çıkışı bulunmaktadır (Begley ve ark., 2000).

Yapay nörona ilişkin matematiksel olarak formüle edilmiş hali şekil 4'te gösterilmektedir.



Şekil 2.5. Yapay sinir ağ modeli. (: Hücrelerden gelen girdi bilgileri, : Bağlantı üzerindeki ağırlıklar, : Hücreye gelen net girdiyi hesaplayan fonksiyon, Toplama fonksiyonundan elde edilen net girdiyi bir işlemden geçirerek hücre çıktısını (ŷ) belirleyen aktivasyon fonksiyon.

Geleneksel yöntemlerin aksine, sinir ağları amaçlanan şekilde çalışmak üzere programlanmamakta veya yapılandırılmamaktadır. Tıpkı insan beyni gibi, bir görevi nasıl gerçekleştireceklerini öğrenmek zorundalar. Öğrenme işlemi, nöronlar arasındaki ağırlık vektör değerinin en aza indirgenmesi ile sağlanmaktadır. Ağın öğrenmesi, ağa gösterilen örnekler için doğru çıktıları üretecek ağırlık değerlerini bulmaktır. Ağın doğru ağırlık değerlerine ulaşması örneklerin temsil ettiği durum için genellemeler yapabilme imkânı sunmaktadır. Ağın bu genelleştirme özelliğinin mevcut olması işlemine ağın öğrenmesi denmektedir. Ağın öğrenmesi için geri yayılım, ağırlıkları gerçek çıktı ile istenen çıktı arasındaki farkı en aza indirgemek için tekrar tekrar ayarlama işlemi olarak ifade edilmektedir (Hagan ve ark., 1996).

2.5.3.1. İletim Yönüne Göre Yapay Sinir Ağ Türleri

1. İleri beslemeli ağ, nöronlar girişten çıkışa doğru katman(lar) halinde bulunmaktadır. Her bir katman kendinden sonraki katmanla bağlantılı olacak şekilde yapılanmaktadır. Ağa gelen bilgiler önce girdi katmanına sonra ara katman(lar)a ve son aşama da çıktı katmanına geçerek işlemi sonuçlandırmaktadır. Bu durum doğrusal olmayan statik bir davranış olmaktadır.
2. Geri beslemeli ağ, en az bir hücrenin çıkışı kendisine ya da diğer hücrelere girdi olarak verilmektedir. Geri besleme, bir katmandaki hücreler arasında olabileceği gibi katmanlar arasındaki hücreler arasında da olabilmektedir. Bu durum doğrusal olmayan dinamik bir davranış olmaktadır (Jain ve ark., 1996).

2.5.3.2. Sinir Ağlarının Öğrenme Stratejileri

2.5.3.2.1. Denetimli Öğrenme (Supervised Learning)

Bilinen sonuçlara sahip bir test veri seti (yeterince büyük) mevcut ise kullanıla bilinmektedir. Bu veri setinde öğrenme şöyle gerçekleşmektedir:

- Bir veri kümesi işlenir.
- Elde edilen kestirimler yanıt değişkeni ile karşılaştırılır.
- Ağ ayarlanır ve tekrar edilir.

Öğrenme stratejisi bu sırayla gerçekleştirilmektedir (Hagan M.T., 1996).

2.5.3.2.2. Denetimsiz Öğrenme (Unsupervised Learning)

Hiçbir test verisi hazır bulunamazsa ve istenen davranıştan bir tür maliyet fonksiyonu elde etmek mümkün ise bu yöntem yararlı olmaktadır. Maliyet fonksiyonu, sinir ağına hedefin ne kadar uzakta olduğunu söylemektedir. Ağ, daha sonra gerçek veriler üzerinde çalışırken, parametrelerini anında ayarlamasını yapabilmektedir (Hagan M.T., 1996).

2.5.3.2.3. Güçlendirilmiş öğrenme (Reinforced learning)

Bu öğrenme metodunu örnek ile açıklayacak olursak, bir yarışmada kazanılacak olan ödüle sahip olabilmek için doğru yolu bulmak olarak tanımlana bilinmektedir. Ağ, ödüle giden birinci yolu dener hedefe ulaşılamazsa ikinci yolu dener ve ulaşabilmek için ağın

yapısındaki ağırlıkları doğru yolu bulana kadar sürekli değiştirerek ayarlamalar yaparak öğrenen bir öğrenme stratejisidir (Hagan M.T., 1996). Yani minimum maliyet ile doğru hedefe ulaşmayı amaçlamaktadır.

2.5.3.3. Yapay Sinir Ağlarının Avantajları ve Zayıflıkları

Yapay sinir ağlarında birden fazla hücre bulunmakta ve bu hücreler eşanlı olarak karmaşık işlevleri çalıştırabilmektedir. Bu çalışma sürecinde bilgiler ağın bütününde saklanmasından dolayı hücrelerden biri işlevini kaybetse bile sistem çalışmasına güvenli bir şekilde devam edebilmektedir. Makinelerin öğrenmesini sağlayarak kararlar verebilmesine imkân sunmaktadır. Eğitim esnasında yapay sinir ağlarına verilen örneklerden genellemeler yapabilmekte ve bu genellemeler ile yeni bir örnek hakkında bilgi verebilmektedir. Örüntü ilişkilendirme, örüntü tanıma, tahmin, sınıflandırma gibi işlevleri gerçekleştirebilmektedir. Eksik bilgiyle çalışabilmekle birlikte hata toleransına sahip yapılar olarak birçok avantajı bulunmaktadır. Otomotiv, bankacılık, savunma, elektronik, finans, robotik, sağlık gibi birçok alanda kullanılması da avantajlarındandır (Papik ve ark., 1998; Dougherty, 1995). Paralel işlem yapabilme yetisinden dolayı donanıma bağımlıdır. Ağ yapısı deneme yanılma yoluyla belirlenmektedir, belirli bir kuralı bulunmamaktadır. Ağ eğitiminin tamamlanma süreci ağın örnekler üzerindeki hatasının belirli bir değerin altına düşmesi ile tamamlanmaktadır, dolayısı ile buna ilişkin belirli bir optimum sonuç verebilecek yöntem(ler) bulunmamaktadır. Probleme yönelik üretilen çözümde kullanılan ağın davranışı bilinmediği için ağa olan güvenin azalmasına sebebiyet vermektedir (Tu J.V., 1996).

2.5.4. Rastgele Orman (Random Forest)

RF yöntemi Leo Breiman tarafından 2001 yılında geliştirilmiştir. Kim Ho tarafından geliştirilen RS (Random Subspace) ve 1996 yılında Leo Breiman tarafından geliştirilen Bagging yöntemlerinin birleşiminden ortaya çıkarılmıştır. Amit ve Geman tarafından 1997'de tanımlanan, her düğüm için en iyi ayrımın rassal olarak seçilmesiyle belirlenen bir çalışmadan esinlenmektedir. RF, sınıflandırma ve regresyon problemlerine uygulanabilen esnek ve kullanımı kolay olan popüler bir makine öğrenmesi yöntemlerinden bir tanesidir. RF, sınıflandırma analizini gerçekleştirirken birden fazla

karar ağacı üretmekte (Ağaçlar her bir düğüm, tüm değişkenler arasında en iyi ayrımı kullanılarak bölünmektedir.) ve bu ağaçlar karar ormanını oluşturmaktadır. Bu karar ağaçları ilgilenilen veri kümesinden rassal olarak seçilen alt kümelerdir. Alt kümelerden alınan tahminler ve oylama (voting) yolu ile en iyi çözüm seçilmektedir. RF, birden fazla ağaç üretmek için tekrarlanan bölümleme ve parçalama aşamalarını kullanan Bagging öğrenme yöntemidir (ChemModLab, 2009).

Yöntemin iki önemli parametresi bulunmaktadır. Birincisi en iyi bölümlemeyi belirlemek için her bir düğümde kullanılan değişken sayısı, ikincisi geliştirilecek ağaç sayısı (Pal, 2003). RF görüntü sınıflandırma, değişken seçimi gibi çeşitli amaçlarla kullanılmaktadır. Moleküler biyoloji, bankacılık, borsa, biyomedikal, yazılım geliştirme, sağlık ve astronomi gibi alanlarda da kullanılmaktadır (Gao ve Zhang, 2009; Boulesteix, 2012; Klassen ve ark., 2008).

2.5.4.1. Yöntemin çalışma prensibi

1. Veri setinde bulunan p özelliklerinden rassal olarak k değişken seçilmesi ($k << p$)
2. K değişkenleri arasında düğümün (d) en iyi ayrılma noktasının kullanılarak belirlenmesi
3. En iyi ayrılma noktasına göre daughter düğümlerinin belirlenmesi
4. Düğüm sayısına ulaşana kadar 1-3 adımlarının tekrar edilmesi
5. N ağaç sayısı oluşturmak için n kez 1-4 adımlarını tekrar ederek ormanın oluşturulması

2.5.4.2. Avantajları ve Zayıflıkları

Oluşturulan ağaç sayısına bağlı olarak doğru sonuçlar üreten sağlam bir yöntem olarak bilinmektedir. Aşırı uyum (overfitting) probleminden önemli derecede etkilenmemektedir. Çünkü bütün tahminlerin ortalama değerini kullanmaktadır. Hem regresyon hem de sınıflandırma analizlerinde kullanılmaktadır. Sınıflandırma anında kuralı durdurma ya da budama işlemleri yapılmamaktadır (Breimen, 2001; Archer 2008). Eğitim veri setindeki mevcut değişkenlerden en önemli değişkeni tanımlayabildiği için daha iyi genellemeler yapılabilir ve geçerli tahminlerde bulunulabilir. Model oluştururken değişkenleri önem sırasına göre sıralaması, araştırmacıların kestirimde kullanılan değişkenleri yorumlayabilmeleri için büyük kolaylık sunmaktadır. Tahmin yapıldığında bütün ağaçlar aynı girdi için tahmin üretip

oylamak zorunda olmasından dolayı zaman alıcıdır. Modelin yorumlanması zordur (Belgiu ve Dragut, 2016).

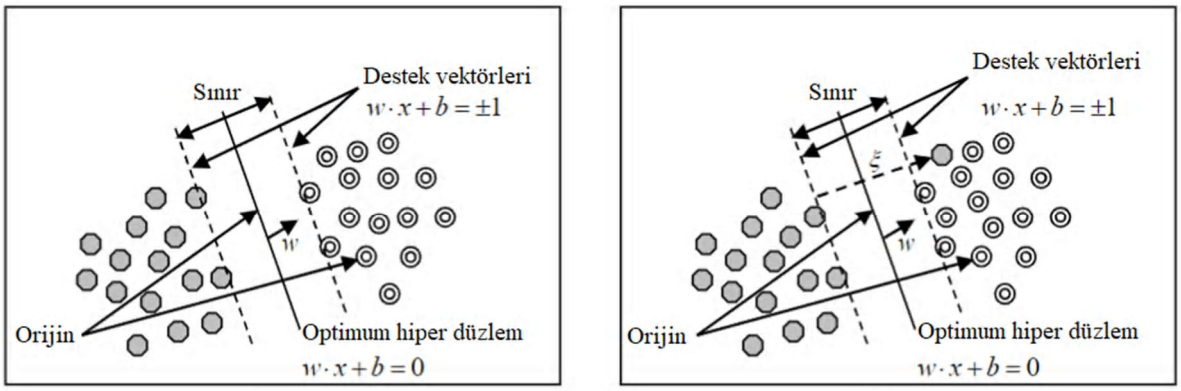
2.5.5. XGBoost (Extreme Gradient Boosting)

XGBoost, GB (Gradient Boosting) özelliğini kullanan karar ağacı tabanlı bir makine öğrenim yöntemidir. Bu yöntem sınıflandırma problemlerinde çapraz entropi (cross-entropy), regresyon problemlerinde hata kareler ortalamasını (mean squared error) kullanarak kayıp fonksiyonu (loss function) optimize etmektedir. Tahminlerde bulunan zayıf bir öğrenici, tamah bir şekilde oluşturulan karar ağacına benzetilmektedir. Kayıp fonksiyonu en aza indirgemek için zayıf öğrenicileri modele dâhil eden eklemeli bir modeldir. Xgboost yöntemi 2014 yılında Chen ve Guestrin tarafından geliştirilmiştir. Hesaplama hızı ve model performansı için tasarlanan bir karar ağaç uygulamasıdır. Yöntemin yapısı eğitim (train) anında CPU çekirdeklerinin tamamını kullanarak ağaç yapılarını paralelleştirmekte, bir küme makinesi kullanarak büyük modelleri eğitmek için dağıtılmış işlem kullanmakta, belleğe sığmayan çok büyük veriler için Out-of-Core hesaplama kullanmakta, donanımdan en iyi şekilde yararlanabilmek için verileri ve yöntemleri saklamaktadır. Bu yöntemde da kullanılan parametrelerden biri olan nround parametresi, sonucun performansını arttırmak için maksimum yenileme sayısını belirlemede kullanılmaktadır. Eta parametresi, öğrenme hızını belirlemede kullanılmaktadır. Gama parametresi aşırı öğrenme problemini kontrol etmek için kullanılmaktadır (Chen ve Guestrin, 2016).

2.5.6. Destek Vektör Makineleri (Support Vector Machines)

İstatistiksel öğrenme teorisine dayanan SVM yöntemi Vapnik ve arkadaşları tarafından geliştirilmiştir. SVM makine öğrenimi yöntemleri arasında yer alan sınıflandırma ve regresyon analizleri için kullanılan güçlü yöntemlerden biridir (Cortes ve Vapnik, 1995). SVM, K boyutlu bir uzayda veri noktalarını geometrik olarak en iyi şekilde sınıflandırabilen bir hiper düzlem (karar sınırı) bulmayı amaçlamaktadır. SVM ile eğitim seti içindeki veriler kullanılarak bir hiper düzlem belirlenir ve bu oluşturulan düzlemin geçerliliği (genelleme yeteneği) test veri seti olarak adlandırılan bağımsız veri setleri kullanılarak doğrulanmaktadır. K boyutlu bir verinin sınıflandırılmasında SVM ile $K-1$ boyutlu bir hiper düzlem geliştirilmektedir. Bir veri kümesine ait girdinin (iki

sınıflı) doğrusal olarak ayrılabildiğini varsayalım. Bu veri kümesini ayırabilen sonsuz sayıda hiper düzlem bulunmaktadır (Sherrod, 2003). Bu hiper düzleme en uygun hiper düzlem ve sınır genişliğini belirleyen (sınırlandıran) noktalar ise destek vektörler olarak bilinmektedir. SVM ile iki sınıflı problemlerin çözümü kullanılarak doğrusal olarak ayrılamayan veya çoklu sınıf problemlerinin çözümü için genelleştirilme yapılabilmektedir (Cortes and Vapnik, 1995). İki sınıflı ve doğrusal olarak ayrılabilen bir problemde SVM'nin eğitimini inceleyecek olursak k sayıda örnekten oluşan eğitim verisinin olduğunu varsayalım. Burada olup boyutlu bir uzayı, ise sınıf etiketlerini belirtmektedir. Hiper düzlemin normali ve eğilim değeri bilinirse örnekler doğrusal olarak ayrılabilir. Buna ilişkin durum şekil 2.6.1'de gösterilmektedir.



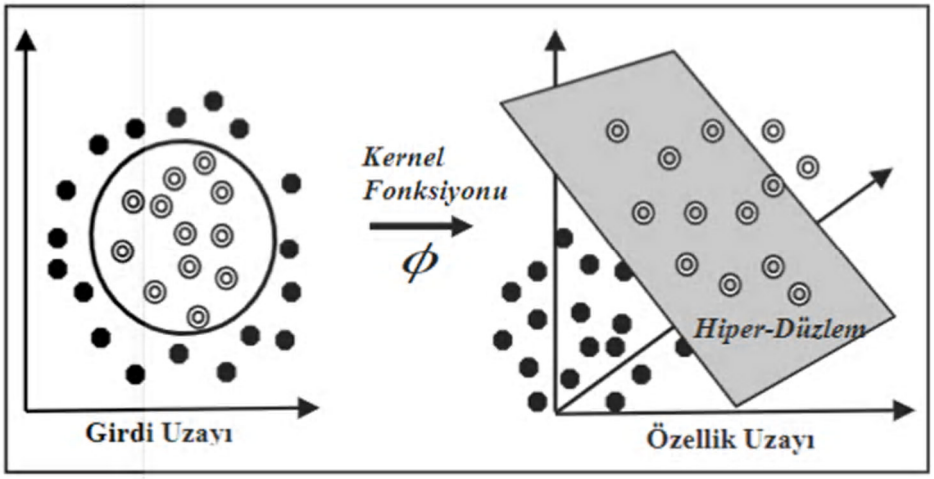
Şekil 2.6. Doğrusal olarak ayrılabilen veriler için hiper düzlemin gösterilmesi. Doğrusal olarak ayrılamayan veriler için hiper düzlemin gösterilmesi (Küçüksille ve Ateş, 2013).

Şekilde görüldüğü gibi optimum hiper düzleme paralel olan iki hiper düzlem bulunmaktadır. Bu hiper düzlemler , şeklinde bir fonksiyon ile tanımlanmaktadır. Dolayısıyla bu düzlemler arasındaki sınır — kadardır (Foody ve Mathur, 2004).

Bu durumda optimum hiper düzlemin belirlenmesi için – sınırlı optimizasyon probleminin çözümünü gerektirmektedir. Bu çözümün sınırları ise

ve şeklinde olur.

Yüksek boyut problemi bulunmayan bu yöntem için seçilen örnek tıbbi tanıma, yüz tanıma, ses analizi ve el yazısı tanıma gibi birçok sınıflandırma probleminde verilerin doğrusal olarak ayrılması mümkün olmamaktadır. Bu durumda Cortes ve Vapnik (1995) sınıf etiketi ile verilen eğitim verilerin optimum hiper düzlem ile aynı tarafta bulunması ile ilgili kısıtlamanın pozitif bir "arttıran yapay değişken" tanımlanması ile yumuşatıldığını ifade etmişlerdir (Şekil 2.6.). Girdi uzayında doğrusal biçimde ayrılamayan veri kümesi daha büyük boyutlu bir uzayda (özellik uzayı) görüntülenmesine imkân sağlamakta ve bu durum şekil 2.7'de gösterilmektedir.



Şekil 2.7. Çekirdek fonksiyon kullanılarak verinin daha yüksek bir boyuta dönüşümü (Küçüksille ve Ateş, 2013)

Destek vektör makineleri çekirdek fonksiyonu yardımıyla doğrusal olmayan dönüşümler uygulayarak verilerin yüksek boyutta doğrusal olarak ayrılabilmesine imkân sağlamaktadır (Cortes ve Vapnik, 1995; Gunn, 1998; Cristianini ve Taylor 2000).

Sonuç olarak; doğrusal olarak ayrılamayan iki sınıflı bir problemde iki sınıfın birbirinden ayrılması ile ilgili karar kuralı
şeklinde yazılabilir (Osuna ve ark.,1997).

Tablo 2.2. SVM'de Kullanılan Bazı Çekirdek Fonksiyonları

KERNEL FONKSİYONU	FORMÜL	AÇIKLAMA
Doğrusal Kernel	$K(x,y) = x * y$	
Polinom Kernel	$K(x,y) = ((x * y) + 1)^d$	d, polinom derecesi
Radyal Tabanlı Kernel	$K(x,y) = e^{-\gamma\|x-x_i\|^2}$	γ, Gauss kernelinin boyutu
Sigmoid Kernel	$K(x,y) = \tanh(b(x * y) + r)$	b, r kernel parametreleri

SVM modelleri dört farklı başlık olarak ele alınabilir. Bunlar sırasıyla doğrusal, polinom, radyal ve sigmoid çekirdektir. Bu fonksiyonlara ilişkin formüller tablo 2.2'de verilmiştir. Doğrusal çekirdek, radyal tabanlı çekirdeğin özel bir durumu olarak düşünebilir. Ayrıca belirli parametreler için sigmoid çekirdek, radyal tabanlı çekirdek gibi bir davranış sergilediğini belirtmişlerdir (Keerthi ve ark., 2003). Radyal tabanlı çekirdek yöntemi, diğer yöntemlere nazaran uzaktan algılama uygulamaları için daha iyi performans sergilediği belirtilmektedir (Keerthi ve ark., 2003; Melgani ve Bruzzone, 2004; Foody ve Mathur 2004; Pal ve Mather 2005).

3. GEREÇ ve YÖNTEM

Bu tez çalışmasında kolon, böbrek ve tiroit olmak üzere üç farklı gerçek veri seti kullanılmıştır. Çalışmada kullanılan verileri tekli veri matrisleri ve birleştirilmiş veri matrisleri olarak iki farklı stratejide değerlendirilmiştir.

Birinci stratejide yer alan verileri üç farklı aşamada değerlendirilmiştir. Bu stratejinin birinci aşamasında kullanılan kolon, böbrek ve tiroit tekli verileri ham olarak NSC, RF ve SVM yöntemleri kullanılarak değerlendirilmiştir. MKL yönteminin kendi içerisinde oluşturduğu tekil matrislerin tersinin hesaplanamamasından dolayı bu aşamada MKL yöntemi kullanılamamıştır. Birinci stratejinin ikinci aşamasında her veriye ayrı ayrı mRMR yöntemi uygulanarak değişken seçimi uygulaması gerçekleştirilerek yeni boyuta sahip yeni veri matrisleri oluşturulmuştur. Bu yeni veri matrisleri MKL, RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. MKL yönteminin kendi içerisinde oluşturduğu tekil matrislerin tersinin hesaplanamamasından dolayı bu aşamada bazı veri matrisleri için MKL yöntemi kullanılamamıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için bu aşamada kullanılmamıştır.

Birinci stratejinin üçüncü aşaması ise ikinci basamakta ayrı ayrı oluşturulan verilerin her birine temel bileşen analizi uygulanıp boyut indirgeme işlemi gerçekleştirilerek yeni boyutlu veriler elde edilmiştir. Yeni boyutlu bu veriler MKL, RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için üçüncü aşamada kullanılmamıştır.

Çalışmamızda önem arz eden ikinci stratejide ise dört farklı aşamada veri entegrasyon yöntemleri kullanılarak verilerin birleştirilmesi uygulaması gerçekleştirilmiştir. Birinci aşamada art arda birleştirme yöntemi, ikinci aşamada dönüşüm tabanlı birleştirme, üçüncü aşamada model-tabanlı birleştirme ve dördüncü aşamada ise mRMR uygulanan verilerden elde edilen değişken sayıları kullanılarak art arda birleştirme uygulaması gerçekleştirilmiştir.

İkinci stratejinin birinci aşamasında art arda birleştirme yöntemi ile birleştirilen veriler üç farklı basamakta değerlendirilmiştir. Birinci basamakta kullanılan kolon, böbrek ve tiroit olmak üzere üç farklı birleştirilmiş veriler ham olarak NSC, RF ve SVM yöntemleri kullanılarak değerlendirilmiştir. Birinci basamakta kullanılan veri matrislerinin boyutlarının çok büyük olması ve MKL yönteminin kendi içerisinde oluşturduğu tekil matrislerin tersinin hesaplanamamasından dolayı bu basamakta MKL yöntemi kullanılamamıştır.

Birinci aşamanın ikinci basamağında birleştirilen her bir veri ayrı ayrı mRMR yöntemi uygulanarak değişken seçimi uygulaması gerçekleştirilerek yeni boyuta sahip yeni veriler oluşturulmuştur. Bu yeni veriler MKL, RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için bu basamakta kullanılmamıştır.

Birinci aşamanın üçüncü basamağında ise ikinci basamakta ayrı ayrı oluşturulan verilerin her birine temel bileşen analizi uygulanıp boyut indirgeme işlemi gerçekleştirilerek yeni veriler elde edilmiştir. Yeni boyutlu bu veriler MKL, RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için üçüncü basamakta kullanılmamıştır.

İkinci stratejinin ikinci aşamasında dönüşüm tabanlı birleştirme yönteminde iki farklı basamak kullanılarak birleştirme işlemi gerçekleştirilmiştir. Bu ikinci aşamanın birinci basamağında veriler çekirdek tabanlı yöntem kullanılarak birleştirilip RVM ve Ada-Boost-RVM yöntemleri kullanılarak sınıflandırılmıştır. İkinci aşamanın ikinci basamağında ise veriler grafik tabanlı yöntem kullanılarak birleştirildikten sonra CANetwork yöntemi kullanılarak sınıflandırılmıştır.

İkinci stratejinin üçüncü aşamasında model tabanlı birleştirme yöntemi kullanılarak birleştirilen verileri üç farklı basamak olarak değerlendirilmiştir. Üçüncü aşamanın birinci basamağında birleştirilen verileri ham olarak NSC, RF ve SVM yöntemleri kullanılarak değerlendirilmiştir. İkinci basamağında birleştirilen her bir veri ayrı ayrı mRMR yöntemi uygulanarak değişken seçimi uygulaması gerçekleştirilerek yeni boyuta sahip yeni veriler oluşturulmuştur. Bu yeni veriler RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için bu basamakta kullanılmamıştır. Üçüncü basamağında ise mRMR değişken seçimi uygulanarak oluşturulan verilerin her birine temel bileşen

analizi uygulanıp boyut indirgeme işlemi gerçekleştirilerek yeni veriler elde edilmiştir. Yeni boyutlu bu veriler RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için üçüncü basamakta kullanılmamıştır.

İkinci stratejinin dördüncü aşaması ise iki farklı basamak olarak değerlendirilmiştir. Birinci basamakta kolon, böbrek ve tiroit tekli verilerinin alt kümelerinin her birine ayrı ayrı mRMR değişken seçim uygulaması gerçekleştirilmiştir. Her bir alt kümede oluşturulan yeni değişkenli veri matrisleri tek bir veri matrisinde birleştirilmiş ve üç farklı birleştirilmiş veri elde edilmiştir. Bu üç farklı veri MKL, RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır.

İkinci basamakta ise birinci basamakta kullanılan verilere temel bileşen analizi uygulanıp boyut indirgeme işlemi gerçekleştirilerek yeni veriler elde edilmiştir. Yeni boyutlu bu veriler MKL, RF ve SVM yöntemleri kullanılarak sınıflandırılmıştır. NSC yöntemi kendi algoritma içerisinde değişken seçimini otomatik olarak gerçekleştirdiği için dördüncü aşamada kullanılmamıştır. Yöntemlerin çalışma süreleri değerlendirilmiştir. Yöntemlere ilişkin değişken sayıları belirlenmiştir. Çalışmada kullanılan gerçek verilere uygulanan sınıflandırma yöntemlerinin performans değerlendirilmesinde ROC eğrisi altında kalan alan, doğruluk oranı, F1 skoru ve Matthew korelasyon katsayı ölçütleri kullanılmıştır. Bütün analizler 25 kez tekrar edilmiştir.

Çalışmada yer alan strateji ve aşamalar için kullanılan veriler tablo 3.1'de verilmektedir.

3.1. Çalışmada kullanılan veriler

Tablo 3.1. Yöntemlerde kullanılan veriler

Veri adı	Veri türü	Değişken sayısı	Örnek sayısı
Kolon kanser			
	miRNA Proteom RNA-Seq	989 8058 13482	95
Böbrek kanser			
	Metilasyon miRNA RNA-Seq	13744 799 20190	89
Tiroit kanser			
	Metilasyon miRNA RNA-Seq	20118 808 19927	496

Bu çalışmada kolon kanseri, böbrek kanseri ve tiroit kanseri olan gerçek veriler kullanılmıştır. Kolon kanser veri kümesinde miRNA, proteom ve RNA-dizileme(RNA-Seq) olarak üç farklı veri türü bulunmaktadır. Böbrek ve tiroit kanser verilerinde metilasyon, miRNA ve RNA-Seq olarak 3 farklı veri bulunmaktadır. Bu verilerde kullanılan örnekler içinden her veri kümesi için aynı fenotipe sahip olan veriler alınmıştır. Bu veriler R yazılımında Reduce ve lapply fonksiyonları kullanılarak belirlenmiştir. Normalize edilmiş bu veriler Vasaikar ve arkadaşlarının oluşturduğu web uygulama kaynağından (LinkedOmics) alınmıştır (Vasaikar ve ark., 2017). LinkedOmics, TCGA projesinden 32 farklı kanser türü için 11.158 hastadan alınan omik verileri ve klinik bilgileri içermektedir. Ayrıca bu uygulama LinkFinder, LinkInterpreter ve LinkCompare olmak üzere üç farklı modül içermektedir. LinkFinder modülü, her bir kanser kohortu için milyarlarca değişken çiftleri arasındaki ilişkileri analiz etme ve görselleştirme ile araştırmacıların ilgilendikleri moleküler ve klinik değişkenler ile diğer değişkenler arasındaki ilişkilerin araştırılmasına imkân sunmaktadır. LinkCompare modülü, çoklu omik verilerinde ve pan kanser analizlerinde yararlı olan LinkFinder modülü tarafından tanımlanan ilişkilendirmelerin kolay bir şekilde karşılaştırılmasında kullanılmaktadır. Son olarak üçüncü modül olan LinkInterpreter ise tanımlanan ilişkileri yolak ve ağ analizler sayesinde biyolojik olarak yorumlanması ve anlaşılmasını sağlamaktadır. Bu uygulamaya http://www.linkedomics.org adresinden ulaşılabilmektedir (Vasaikar ve ark., 2017).

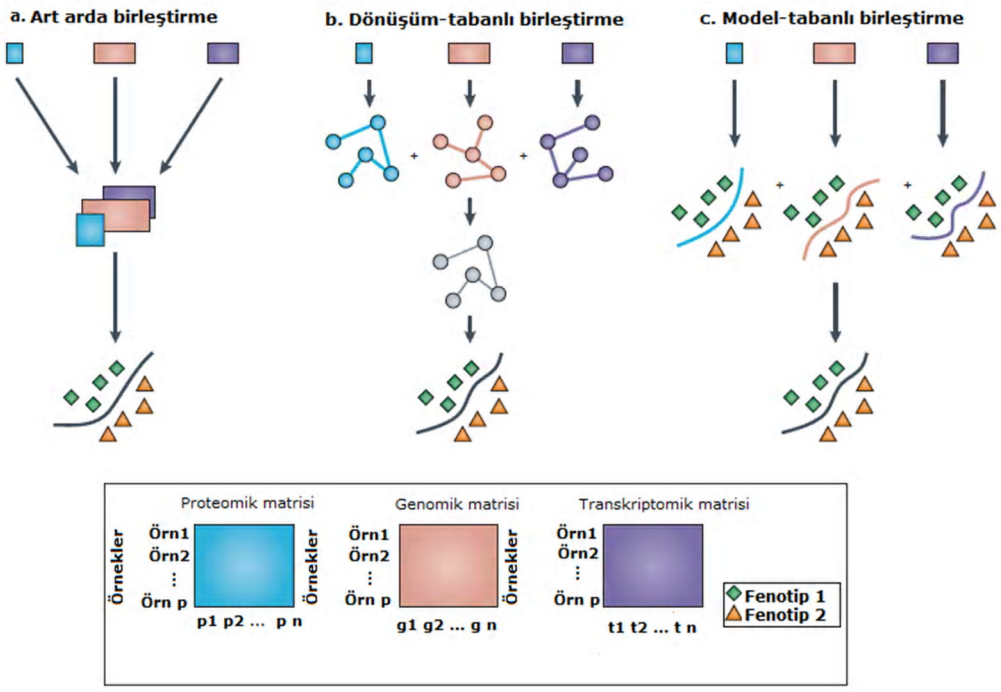
3.2. Verilerin analiz süreci

Satırların örneklerden, sütunların değişkenlerden oluştuğu her omik veri kümesi için kullanılan matris formatı aşağıda verilmektedir.

$$\alpha=\begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ an & \cdots & a_{nm} \end{bmatrix}$$

Öncelikle nxm boyutlu tekli veri matrisleri kullanılarak NSC, RF ve SVM yöntemleri ile sınıflandırılmıştır. Kolon, böbrek ve tiroit verilerinin her bir alt kümesinin boyutunun çok büyük olması NSC, RF ve SVM yöntemleri için problem olmamaktadır. Ancak çoğu zaman her hastalık tipi için ilgili biyobelirteçlerin belirlenmesi çok önemli bir durum olmasından ötürü çok büyük boyutlu verilerle çalışmayı gerektirebilmektedir.

Dolayısıyla öncelikle her bir veri kümesi ayrı ayrı sınıflandırıldıktan sonra her bir alt kümenin birleştirilerek incelenmesi önem arz etmektedir. Örneğin Kolon kanseri verilerinde miRNA, proteom ve RNA-Seq verilerinde bulunan her bir değişken, hastalığın tespiti veya tedavisi, hayati konuda, yaşamın uzaması, maliyet gibi birçok durum üzerine etkili olabilmektedir. Bu yüzden bu verilerde bulunan değişkenlerin her biri üç farklı birleştirme yöntemi kullanılarak birleştirilip değerlendirilmiştir. Bu üç farklı birleştirme yönteminin akışı şekilsel olarak 3.1 nolu şekilde gösterilmektedir.



Şekil 3.1. Birleştirme yöntemlerinin iş akışı (Ritchie ve ark., 2015)

Şekil 3.1.a'da gösterilen art arda birleştirme yöntemi kullanılarak üç farklı veri tek bir matriste birleştirilmektedir. Tek bir veri matrisinde birleştirilen bu verilerin doğrudan birleştirildiği gibi kullanarak sınıflandırma işlemini gerçekleştirmek hatalı olacaktır. Çünkü her bir verinin değişken boyutu farklı olması modelde kullanılacak değişkenlerin eşit ağırlığa sahip olmasını engellemektedir. Bu problemi ortadan kaldırmak için her bir veriye göre blok ölçeklendirme kullanılmıştır (Spicker ve ark., 2007). Blok ölçeklendirme işlemi R yazılımında blockscal fonksiyonu kullanılarak gerçekleştirildi. Blok ölçeklendirme, referans matrisin bloğunun her bir sütununun varyanslarının toplamının karekökü ile ölçeklendirilerek test matrisi elde edildi. Bu test matrisi kullanılarak sınıflandırma analizleri gerçekleştirilmiştir.

Art arda birleştirme yöntemi ile birleştirilen veriler ham olarak değerlendirildikten sonra önem arz eden durumlardan biri olan değişken seçim işlemidir. Ham olarak çalışılan veri matrisinde hastalığın alt tipinin sınıflandırılmasında bazı değişkenlerin etkisi çok büyük olurken bazılarının etkisi hiç etkisi olmayabilir. Dolayısıyla etkisi olmayan değişkenlerin modelde kullanılması gereksiz değişken kalabalığına ve komplekslik gibi problemlerin ortaya çıkmasına neden olabilmektedir. Ayrıca sınıflandırmada modelinde kullanılan değişkenler, hastalık üzerinde etkileri olan omik türlerindeki mevcut değişkenler için detaylı çalışma yapılmak istenebilir. Bu yüzden araştırmacıların yeterli süreye ve maliyete ayıracak güçleri mümkün olmayabilmektedir. Dolayısı ile bu gibi durumları da düşünerek ve hastalığın alt tiplerinin sınıflandırılmasında daha doğru, daha hızlı, daha anlamlı sonuçlar elde edebilmek için değişken seçimi işlemine başvurulmalıdır. Tabi her değişken seçimi yöntemi her veride iyi bir şekilde performans göstermeyebilir. Bu yüzden çalışmanın amacına ve kullanılan verilere uygun olan değişken seçimi yöntemleri kullanılmalı. Biz de literatür taramasına dayalı olarak birleştirilmiş verilerde yaygın olarak kullanılan mRMR değişken seçimi yöntemi kullanmaya karar verdik. Ancak çalışmamızın kapsamında önemli olan bir diğer nokta da bu yöntemle belirlenmesi beklenen değişken sayısıdır.

Tekli veriler, art arda birleştirme ve model tabanlı birleştirme yöntemleriyle elde edilen birleştirilmiş verilerde veri boyutlarının farklı olması ve her veri için kullanılacak olan sınıflandırma modellerinin performans değerlerinin farklılık göstermesi durumu da göz önünde bulundurularak değişken sayılarını 20'den 400'e kadar 20'şer arttırarak her bir değer için sınıflandırma performans değerlerinin pik yaptığı değişken sayıları dikkate alınarak belirlenmiştir. Bu sayılar belirlendikten sonra oluşturulan yeni veriler daha önce bahsettiğimiz stratejiye uygun bir şekilde sınıflandırma analiz aşamaları gerçekleştirilmiştir.

Ham veride olduğu gibi değişken seçimi uygulanan veriden elde edilen yeni veride kullanılan değişken sayıları da fazla olabilmektedir. Bu sayının yüksek olması yine modelin kompleksliğini arttırarak yorumlanmasını da zorlaştırabilmektedir. Her zaman çok sayıda değişken ile çalışmak mümkün olmayabilmektedir. Böyle bir durumda alternatif olarak nasıl bir yol izlenmesi gerektiği yönünde fikir sunabilmek adına değişken seçimi uygulandıktan sonra oluşturulan verilerde temel bileşen analizi uygulanarak fazla boyutlu verilerdeki genel değişkenleri bularak boyut sayısı

azaltılmıştır. Boyut indirgeme aşamasında hastalıkla ilişkisi olmayanların veya ilişkisi yok denecek kadar az olan değişkenler temel alınarak indirgeme işlemi gerçekleştirilmiştir. Temel bileşen analizi R yazılımında prcomp fonksiyonu ile gerçekleştirilmiştir. Çalışmamızın temel amacı olmasa da araştırmacıların çok daha düşük değişkene sahip verilerle çalışmasının gerekliliği gibi durumlarda çözüm olarak hangi yöntemlerin kullanılması gerektiği yönünde fikirler verebilmesi açısından temel bileşen analizi uygulaması da çalışmamıza dâhil edilerek kapsamlı olarak bütün veriler için farklı senaryolarda analizler gerçekleştirilmiştir.

Veriler hazırlandıktan sonra kullanılacak sınıflandırma yöntemleri için kullanılan parametreler ve ayarlamaları belirlenmiştir. Bu parametrelere ilişkin bilgiler bir sonraki paragrafta ifade edilmiştir.

MKL, destek vektör makinesinin uygulanabilirliğini arttırmak için önceden belirlenmiş çekirdek kümesini kullanan ve yöntemin bir parçası olarak en iyi doğrusal veya doğrusal olmayan çekirdek birleşimini gerçekleştiren bir yöntemdir. Bu yöntem optimal çekirdek grubundan parametreler seçmekte, çekirdek seçiminden kaynaklanan yanlılığı azaltmakta, farklı kaynaklardan gelen verileri birleştirmekte kullanılmaktadır, ancak bu kullanımlar için farklı çekirdekler kullanılması gerekmektedir. Bu yöntem yeni bir çekirdek üretmek yerine her bir veri kümesi için önceden belirlenmiş çekirdekleri birleştirmek için çoklu çekirdek yöntemi kullanmaktadır. Çekirdek fonksiyonu kullanan çoklu çekirdek öğrenim yöntemi ve diğer yöntemler gamma, C komplekslik parametresi ve polinom derecesi hiper parametreleri çapraz geçerlilik deneyleri ile optimize edilmiştir.

NSC, her sınıf için ağırlık merkezini hesaplamakta ve Δ parametresini kullanarak ortalamaları sıfıra doğru küçültmektedir. Ayrıca otomatik olarak değişken seçim işlemini gerçekleştirdiği için herhangi bir ayarlama yapılmamıştır.

RF, topluluk öğrenme yöntemi olarak bilinmektedir. Verileri sınıflandırırken birden fazla karar ağacı üreterek sınıflandırma performansını arttırmaya çalışan bir yöntemdir. Üretilen ağaçlar her yaprak düğümü sadece bir sınıfı içerecek şekilde oluşturulmaktadır. Hiperparametre ayarı yapmadan da sonuçlar üretebilmektedir. Esnek kullanımı olan kolay ve güçlü bir makine öğrenimi yöntemidir. m (Her bir düğümde kullanılan değişken sayısı) ve N (Oluşturulacak ağaç sayısı) olmak üzere iki tane parametresi

bulunmaktadır. Ağaç sayısı parametresi eğitim kümesinde yapılan çapraz geçerlilik deneyleri ile optimize edilmiştir.

SVM, sınıflandırma problemlerinde oldukça etkili ve basit bir yöntemdir. Problemlerin çözümüne ilişkin yöntemin uygulanması sürecinde çekirdek fonksiyon seçimi ve C ile γ parametrelerinin optimizasyonu yapılmıştır. Dağılıma ilişkin herhangi bir bilgiye ihtiyaç duymadığı için bu konuda herhangi bir işlem yapılmamıştır.

Çekirdek tabanlı yöntemler olan RVM, Ada-boost RVM yöntemleri için parametre optimizasyonun da yeniden örnekleme boyutu ve maksimum yineleme sayısı için uygun değerler belirlenmiştir.

Grafik tabanlı yöntem CANetwork yönteminde ise bileşik ilişki ağı için c parametresi için optimum değer 1 olarak ayarlanmıştır.

Verilerin analizinde veri birleştirme, değişken seçimi ve temel bileşen analizi uygulamaları sonucunda oluşturulan her veri %80 eğitim ve %20 test seti olacak şekilde bölündü. Eğitilen verilerde ayrı ayrı MKL, NSC, RF, SVM, RVM, Ada-boost RVM ve CANetwork sınıflandırma yöntemleri kullanılarak modeller oluşturuldu. Modeller oluşturulmadan önce her modelin kendine özgü parametreleri optimum olacak şekilde ayarlandı. Modeller oluşturulmadan önce modellerin geçerliliğinin değerlendirilmesi için 5 kez yeniden örnekleme yapması ve 5 kez yeniden örneklenen veri için 10 kez çapraz geçerlilik parametreleri ayarlandı. Modellerin performansı test verisi ile değerlendirildi. Değerlendirmeler sonucunda her bir yöntemin sınıflandırma performans değerleri hata matrisi kullanılarak hesaplanmıştır.

3.3. Model Geçerliliği

Sınıflandırma yöntemlerinde model performansının test setinde değerlendirilmesi genelleştirme hatasının yansız bir şekilde tahmin edilmesini sağlamaktadır. Yansız tahmin yapabilmek için model geçerliliğinin sağlanmasında en çok kullanılan yöntemlerden biri çapraz geçerliliktir. 2-parça çapraz geçerlilik, elimizdeki veri setinin bir kısmını eğitim verisi diğer kısmını test verisi olarak iki eşit parçaya ayırır ve bu şekilde bir doğruluk hesaplaması yapmaktadır. Daha sonra test ve eğitim seti olan veriler yer değiştirilerek yeniden bir hesaplama yapılmakta ve bu doğruluk değerlerinin ortalamasına göre modelin doğruluk değeri hesaplanmaktadır. K-parça çapraz geçerlilik ise her verinin bir kez test ve eğitim seti olarak kullanılabileceği geçerlilik

yöntemlerinin genellenmiş halidir. Örneğin veri k eşit parçaya ayrılır ve k parçadan bir tanesi test verisi iken k-1 tanesi eğitim verisi olarak kullanılır ve bu durum k kez tekrar edilir. Bu durumda elde edilen doğruluk değerlerinin ortalaması alınarak yeniden modelin doğruluk değeri hesaplanmaktadır.

3.4. Model Performans Değerlendirme Ölçütleri

Veri hazırlama ve makine öğrenimi modelinin eğitimi önemli bir durum olmakla birlikte model performansının ölçülmesi de aynı şekilde önemli bir kriterdir. Model performansının ölçülmesinde ROC (Receiver Operating Characteristic) eğrisi altında kalan, F1 skoru, duyarlılık, seçicilik, Matthew korelasyon katsayısı (MKK) ve doğruluk gibi bir çok ölçütlerle birlikte yöntemlerin çalışma süreleri ve kullanılan değişken sayıları da dikkate alınarak değerlendirilmektedir. Veri setindeki gerçek hastalık durumu ile sınıflandırma modelimiz ile tahmin edilen hastalık durumu arasındaki sayıların belirlenmesi ile bu ölçütler hesaplanmaktadır. Gerçek durum ile tahmin edilen durumun sayısal gösterimine ilişkin tablo 3.2'de gösterilmektedir.

Tablo 3.2. Sınıflandırma tablosu

		GERÇEK SINIF	
TAHMİN EDİLEN SINIF		**Pozitif**	**Negatif**
	Pozitif	Doğru pozitif (DP)	Yanlış pozitif (YP)
	Negatif	Yanlış negatif (YN)	Doğru negatif (DN)

Çoklu-omik verilerinde kullanılacak olan sınıflandırma yöntemlerinden elde edilen modellerin sınıflandırma performansının değerlendirilmesinde hata matrisine göre F1 skor, MKK, eğri altında kalan alan (EAA) ve doğruluk ölçüleri kullanılarak belirlenmiştir. Doğruluk oranı grup değişkeninin sınıflarının doğru tahmin edilenlerin sayısına karşılık gelen bir ölçümdür. F1 skoru duyarlılık ve seçiciliğin harmonik ortalamasından elde edilen ve genel performans değerlendirmesinde kullanılan önemli ölçütlerden bir tanesidir. Matthews korelasyon katsayısı –1 ile 1 arasında değer alabilen ve modelin ne kadar iyi performans gösterdiğini açıklayan bir başka ölçüm kriteridir. Sınıflandırma modeli için kullanılan verinin grup değişkeni çok farklı boyutlarda olsa bile dengeli bir ölçü olarak kabul edilmektedir. Bu ölçünün 1'e yakınlaşması grup değişkeninin sınıflarının daha doğru bir şekilde ayrıldığını ifade ederken, -1'e yakınlaşması yanlış sınıflandırdığını ifade etmektedir. ROC eğrisi altında kalan alan,

çeşitli eşik değeri için sınıflandırma probleminde kullanılan bir performans ölçüsüdür. Bu ölçüde kullanılan ROC olasılık değeri grup değişkenlerinin sınıflarının ayrılabilirliğini gösteren bir ölçüdür. Eğri altında kalan alan değerin artması sınıfların ayrımının da o kadar iyi olduğunu göstermektedir. Bahsettiğimiz bu ölçülere ilişkin formüller tablo 3.3'te yer almaktadır.

Tablo 3.3. Model performans ölçütleri

Ölçütler	**Formül**
EAA	(DP/(DP+YN)+DN/(DN+YP))/2
Doğruluk oranı	(DP+DN)/(DP+DN+YP+YN)
F1 skor	2*(DP/2*DP+YP+YN)
MKK	(DP*DN)-(YP*YN)/(DP+YP)*(DP+YN)*(YP+DN)*(YN+DN)

Çalışma kapsamında gerçekleştirilen analizler TÜBİTAK ULAKBİM Yüksek Başarımlı ve Grid Hesaplama Merkezinde bulunan TRUBA kaynaklarında yer alan Centos Enterprise 7.3 Linux işletim sistemi üzerinde 3.6.2 versiyonu kurulu olan R yazılımında gerçekleştirildi. TRUBA, Türkiye'deki üniversite bünyesinde çalışan birçok araştırmacıya ve öğrencilere bilimsel çalışmalarında katkıda bulunması üzerine doğrudan hizmet veren bir Ulusal e-Altyapı sistemidir. Daha ayrıntılı bilgi için (www.truba.gov.tr) adresinden ulaşılmaktadır.

4. BULGULAR

Farklı senaryolarla oluşturduğumuz toplam 42 veride kullanılan yöntemlerin performans değerleri ortalama ve standart hata olarak verilmiştir. Sınıflandırma yöntemlerine ilişkin değişken sayıları ortalama ve standart hata olarak, çalışma süreleri ise dakika cinsinden verilmiştir. Analizlere ilişkin değerler tablolarda özetlenmektedir. Ayrıca kolon, böbrek ve tiroit verilerine ait toplu analiz sonuçlarına EK-1, EK-2 ve EK-3'teki Excel dosyalarından erişilebilmektedir.

4.1. Gerçek Veri Analiz Sonuçları

Tablo 4.1. Tekli gerçek verilerde yöntemlerin çalışma süreleri

Veriler	**Yöntemler**			
	MKL	**NSC**	**RF**	**SVM**
Tekli Veriler				
Kolon miRNA	-	0.091	8.521	0.865
Kolon proteom	-	0.359	81.060	1.715
Kolon RNA-Seq	-	0.796	200.640	3.488
FS kolon miRNA	8.972	-	69.960	0.704
FS kolon proteom	0.858	-	52.482	0.404
FS kolon RNA-Seq	9.048	-	71.220	0.736
FS PCA kolon miRNA	14.958	-	96.120	7.670
FS PCA kolon proteom	9.822	-	79.020	7.339
FS PCA kolon RNA-Seq	14.994	-	96.540	7.694
Böbrek metilasyon	-	1.208	280.200	5.064
Böbrek miRNA	-	1.259	284.460	5.149
Böbrek RNA-Seq	-	1.885	414.660	7.125
FS böbrek metilasyon	0.814	-	4.089	0.159
FS böbrek miRNA	0.329	-	2.157	0.110
FS böbrek RNA-Seq	0.162	-	1.125	0.076
FS PCA böbrek metilasyon	9.165	-	72.300	7.201
FS PCA böbrek miRNA	9.088	-	71.820	7.173
FS PCA böbrek RNA-Seq	9.066	-	71.520	7.149
Tiroit metilasyon	-	3.647	3054.24	21.348

Tiroit miRNA	-	3.739	3137.760	21.721
Tiroit RNA-Seq	-	5.506	6130.080	33.079
FS tiroit metilasyon	5.097	-	69.240	0.667
FS tiroit miRNA	2.247	-	59.159	0.493
FS tiroit RNA-Seq	-	-	51.725	0.376
FS PCA tiroit metilasyon	13.684	-	88.620	7.520
FS PCA tiroit miRNA	10.512	-	81.900	7.379
FS PCA tiroit RNA-Seq	9.809	-	78.960	7.316

Çalışma süreleri dakika olarak verilmiştir.

MKL, NSC, RF ve SVM olmak üzere dört farklı yönteme göre veriler, çalışma süreleri açısından değerlendirilmiştir. Kolon verisinin alt kümelerinden miRNA, proteom ve RNA-Seq tekli ham verilerinde maliyeti düşük olan yöntem NSC'dir. Kolon verisinin alt kümelerine mRMR değişken seçimi ve mRMR değişken seçim işleminin ardından temel bileşen analizi uygulanan tekli verilerde gerçekleştirilen sınıflandırma analizinde maliyeti düşük olan yöntemin SVM olduğu görülmüştür.

Böbrek verisine ait alt kümelerinden metilasyon, miRNA ve RNA-Seq tekli ham verilerinde maliyet açısından düşük olan yöntem NSC'dir. Böbrek verisinin alt kümelerine mRMR değişken seçimi ve mRMR değişken seçim işleminin ardından temel bileşen analizi uygulanan tekli verilerde gerçekleştirilen sınıflandırma analizinde en az maliyete sahip yöntemin SVM olduğu görülmüştür.

Tiroit verisine ait alt kümelerinden metilasyon, miRNA ve RNA-Seq tekli ham verilerinde en az maliyete sahip yöntem NSC'dir. Tiroit verisinin alt kümelerine mRMR değişken seçimi ve mRMR değişken seçim işleminin ardından temel bileşen analizi uygulanan tekli verilerde gerçekleştirilen sınıflandırma analizinde en düşük maliyetli olan yöntemin SVM olduğu görülmüştür.

Tablo 4.2. Art arda birleştirilen gerçek verilerde yöntemlerin çalışma süreleri

Veriler	**Yöntemler**			
	MKL	**NSC**	**RF**	**SVM**
Kolon	**-**	2.280	348.960	4.759
Böbrek	**-**	9.582	609.120	15.104
Tiroit	**-**	22.554	6052.32	51.288
FS kolon	12.411	**-**	4.194	0.104
FS böbrek	2.837	**-**	4.854	0.133
FS tiroit	8.894	**-**	17.366	0.350
FS PCA kolon		**-**	82.500	7.406
FS PCA böbrek		**-**	82.860	7.430
FS PCA tiroit		**-**	95.640	7.645
Tekli Verilerde FS Uygulanıp Birleştirilen Veriler				
FS birleştirilmiş kolon	**-**	**-**	19.026	51.335
FS birleştirilmiş böbrek	**-**	**-**	22.239	51.404
FS birleştirilmiş tiroit	**-**	**-**	84.180	51.622
FS PCA birleştirilmiş kolon	15.016	**-**	97.080	7.721
FS PCA birleştirilmiş böbrek	15.059	**-**	97.620	7.747
FS PCA birleştirilmiş tiroit	16.792	**-**	107.340	7.918

Çalışma süreleri dakika olarak verilmiştir.

Art arda birleştirme yöntemi kullanılarak birleştirilen verilerde, mRMR değişken seçimi uygulanan birleştirilmiş verilerde, mRMR değişken seçimi uygulandıktan sonra temel bileşen analizi uygulanan verilerde, tekli verilerde mRMR değişken seçimi uygulandıktan sonra art arda birleştirme yöntemi kullanılarak birleştirilen verilerde ve bu verilere temel bileşen analizi uygulandıktan sonra çalışma süreleri her yöntem için ayrı ayrı değerlendirilmiştir.

Birleştirilmiş kolon, böbrek ve tiroit verilerinde NSC yöntemi en hızlı çalışma süresine sahip olduğu görülmüştür. mRMR değişken seçimi uygulanan ve mRMR değişken seçimi uygulandıktan sonra temel bileşen analiz uygulanan birleştirilmiş kolon, böbrek ve tiroit verilerinde gerçekleştirilen sınıflandırma yöntemlerinden en düşük maliyete sahip yöntem SVM'dir.

Tekli verilerde mRMR değişken seçimi uygulandıktan sonra art arda birleştirme yöntemi kullanılarak birleştirilen kolon ve böbrek verilerinde maliyeti düşük olan yöntem RF iken tiroit verisinde SVM yönteminin düşük maliyetli olduğu gözlenmiştir. Tekli verilerde mRMR değişken seçimi uygulandıktan sonra art arda birleştirme yöntemi kullanılarak birleştirilen kolon, böbrek ve tiroit verilerine uygulanan temel bileşen analiz sonucunda en az maliyetli olan yöntemin SVM olduğu gözlenmiştir.

Tablo 4.3. Dönüşüm tabanlı birleştirilen gerçek verilerde yöntemlerin çalışma süreleri

Veriler	**Yöntemler**		
	RVM	**Ada-boost RVM**	**CANetwork**
Kolon	0.535	2.322	0.548
Böbrek	0.676	0.626	0.949
Tiroit	13.504	27.248	1.683

Çalışma süreleri dakika olarak verilmiştir.

Dönüşüm tabanlı birleştirme yöntemi kullanılarak birleştirilen kolon verisi için en az maliyete sahip yöntem RVM, böbrek verisi için Ada-boost RVM tiroit verisi için CANetwork olduğu görülmüştür.

Tablo 4.4. Model tabanlı birleştirilen gerçek verilerde yöntemlerin çalışma süreleri

Veriler	**Yöntemler**		
	NSC	**RF**	**SVM**
Kolon	0.516	202.020	2.525
Böbrek	0.696	214.200	3.324
Tiroit	2.680	6065.28	24.421
FS kolon	-	2.879	0.150
FS böbrek	-	4.182	0.174
FS tiroit	-	64.020	0.572
FS PCA kolon	-	1.075	0.138
FS PCA böbrek	-	1.047	0.136
FS PCA tiroit	-	15.420	0.308

Çalışma süreleri dakika olarak verilmiştir.

Model tabanlı birleştirme yöntemi kullanılarak birleştirilen kolon, böbrek ve tiroit verileri için en düşük maliyete sahip yöntemin NSC olduğu görülmüştür. Model tabanlı birleştirme yöntemi kullanılarak birleştirilen kolon, böbrek ve tiroit verilerine uygulanan mRMR değişken seçiminden elde edilen verilerde gerçekleştirilen yöntemlerin analiz sonucuna göre en düşük maliyete sahip yöntemin SVM olduğu görülmüştür. Model tabanlı birleştirme yöntemi kullanılarak birleştirilen kolon, böbrek ve tiroit verilerine uygulanan mRMR değişken seçiminin ardından temel bileşen analizi uygulanan verilerde kullanılan yöntemlerde maliyeti en az olan SVM'dir.

Tablo 4.5. Tekli gerçek verilerde yöntemlere göre değişken sayıları

Veriler	**Yöntemler**	
	NSC	**RF-SVM**
Kolon miRNA	4.720(0.308)	889
Kolon proteom	3.880(0.301)	8058
Kolon RNA-Seq	2.040(0.083)	13482
FS kolon miRNA	-	80
FS kolon proteom	-	40
FS kolon RNA-Seq	-	80
FS PCA kolon miRNA	-	25
FS PCA kolon proteom	-	5
FS PCA kolon RNA-Seq	-	15
Böbrek metilasyon	6193.320(288.073)	13744
Böbrek miRNA	451.600(4.625)	799
Böbrek RNA-Seq	13489.840(276.127)	20190
FS böbrek metilasyon	-	280
FS böbrek miRNA	-	120
FS böbrek RNA-Seq	-	100
FS PCA böbrek metilasyon	-	35
FS PCA böbrek miRNA	-	23
FS PCA böbrek RNA-Seq	-	15
Tiroit metilasyon	1198.560(250.341)	20118
Tiroit miRNA	105.040(6.138)	808
Tiroit RNA-Seq	840.400(214.847)	19927
FS tiroit metilasyon	-	60
FS tiroit miRNA	-	40
FS tiroit RNA-Seq	-	340
FS PCA tiroit metilasyon	-	34
FS PCA tiroit miRNA	-	14
FS PCA tiroit RNA-Seq	-	43

Değişken sayıları ortalama(standart hata) olarak verilmiştir.

NSC, RF ve SVM olmak üzere üç farklı yönteme göre veriler, modelde kullanılan değişken sayıları açısından değerlendirilmiştir.

RF ve SVM yöntemlerinin algoritma aşamalarında herhangi bir değişken seçim işlemi bulunmadığından bütün veriler için modelde kullanılan değişken sayıları bu yöntemler için aynı olmaktadır. Bu değişkenlerin sayılarına ilişkin değerler tablo 4.5'te verilmektedir. NSC yönteminin algoritma aşamasında ise değişken seçimi işlemi gerçekleştirildiğinden her veri kümesi için bu yöntem en uygun sayıyı belirleyerek ilgili değişkenleri modele dâhil etmektedir. İlgili sayılar tablo 4.5'te yer almaktadır. NSC yönteminde mevcut olan değişken seçim özelliğinden ötürü, mRMR ve temel bileşen analizleri uygulanan veriler NSC yöntemi tarafından analize dâhil edilmemiştir.

Tablo 4.6. Art arda birleştirilen gerçek verilerde yöntemlere göre değişken sayıları

Veriler	**Yöntemler**	
	NSC	**RF-SVM**
Kolon	787(453.268)	22429
Böbrek	21031.960(922.856)	34733
Tiroit	13669.760(1928.102)	40853
FS kolon	-	400
FS böbrek	-	60
FS tiroit	-	80
FS PCA kolon	-	37
FS PCA böbrek	-	20
FS PCA tiroit	-	43
Tekli Verilerde FS-PCA Uygulanıp Birleştirilen Veriler		
FS birleştirilmiş kolon	-	200
FS birleştirilmiş böbrek	-	500
FS birleştirilmiş tiroit	-	440
FS PCA birleştirilmiş kolon	-	28
FS PCA birleştirilmiş böbrek	-	38
FS PCA birleştirilmiş tiroit	-	66

Değişken sayıları ortalama(standart hata) olarak verilmiştir.

Art arda birleştirilen ham verilerde NSC, RF ve SVM yöntemlerinin kullandığı değişken sayıları NSC yönteminde daha düşük iken diğer yöntemlerde daha yüksek olduğu gözlenmiştir. NSC yönteminde mevcut olan değişken seçim özelliğinden ötürü, mRMR ve temel bileşen analizleri uygulanan veriler NSC yöntemi tarafından analize dâhil edilmezken RF ve SVM yöntemleri için değişken sayıları tablo 4.6'da gösterilmiştir.

Tablo 4.7. Ham kolon verilerinde MKL, NSC, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**Ham veri**			**Art Arda Birleştirilmiş Ham Veri**	**Model tabanlı Birleştirilmiş Ham veri**
	miRNA	Proteom	RNA-Seq	miRNA+Proteom+RNA-Seq	miRNA+Proteom+RNA-Seq
NSC					
AUC	0.537(0.001)	0.542(0.002)	0.537(0.001)	0.532(0.004)	0.534(0.000)
Doğruluk oranı	0.754(0.005)	0.754(0.005)	0.754(0.005)	0.743(0.011)	0.762(0.000)
F1 skoru	0.173(0.003)	0.179(0.004)	0.173(0.003)	0.173(0.005)	0.167(0.000)
MCC skoru	0.140(0.004)	0.144(0.004)	0.140(0.004)	0.128(0.010)	0.138(0.000)
RF					
AUC	0.551(0.004)	0.544(0.004)	0.553(0.004)	0.540(0.005)	0.739(0.011)
Doğruluk oranı	0.762(0.007)	0.751(0.007)	0.761(0.007)	0.754(0.011)	0.877(0.018)
F1 skoru	0.202(0.008)	0.201(0.009)	0.208(0.009)	0.183(0.011)	0.635(0.032)
MCC skoru	0.174(0.010)	0.146(0.011)	0.179(0.011)	0.145(0.015)	0.604(0.060)
SVM					
AUC	0.552(0.003)	0.622(0.004)	0.597(0.004)	0.602(0.007)	0.711(0.027)
Doğruluk oranı	0.732(0.003)	0.778(0.003)	0.756(0.003)	0.774(0.007)	0.855(0.014)
F1 skoru	0.253(0.004)	0.370(0.007)	0.324(0.007)	0.337(0.014)	0.538(0.051)
MCC skoru	0.124(0.006)	0.280(0.008)	0.238(0.008)	0.282(0.017)	0.510(0.052)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

Tablo 4.7'de kolon verilerinin analiz sonucuna göre ham verilerin alt kümeleri dikkate alındığında miRNA verisi için en düşük performansa sahip yöntem NSC iken, proteom ve RNA-Seq verileri için en yüksek performansa sahip yöntem SVM'dir. Art arda birleştirilmiş ham verilerde ise en yüksek performansa sahip yöntem SVM iken model tabanlı birleştirilmiş ham veride RF olduğu gözlenmiştir.

Art arda birleştirilmiş ve model tabanlı birleştirilmiş ham verilerde kullanılan NSC yönteminin sınıflandırma performansları miRNA proteom ve RNA-Seq ham verilerinde kullanılan NSC yönteminin sınıflandırma performanslarından düşük olduğu gözlenmiştir.

Art arda birleştirilmiş ham veride kullanılan RF yönteminin sınıflandırma performansı miRNA proteom ve RNA-Seq ham verilerinde kullanılan RF yönteminin sınıflandırma performanslarından düşük olduğu gözlenmiştir.

Model tabanlı birleştirilmiş ham veride kullanılan RF yönteminin sınıflandırma performansı miRNA proteom ve RNA-Seq ham verilerinde kullanılan RF yönteminin sınıflandırma performanslarından yüksek olduğu gözlenmiştir.

Art arda birleştirilmiş ham veride kullanılan SVM yönteminin sınıflandırma performansı miRNA ve RNA-Seq ham verilerinde kullanılan SVM yönteminin sınıflandırma performanslarından düşük iken proteom ham verisinde kullanılan SVM yönteminin performansından yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilmiş ham veride kullanılan SVM yönteminin sınıflandırma performansı miRNA proteom ve RNA-Seq ham verilerinde kullanılan SVM yönteminin sınıflandırma performanslarından yüksek olduğu gözlenmiştir.

Birleştirilmiş verilerde model tabanlı birleştirilen verilere uygulanan SVM yönteminin performans değeri en yüksek bulunmuştur.

Tablo 4.8. Dönüşüm tabanlı birleştirilmiş kolon verisinde RVM, Ada-boost RVM ve CANetwork yöntemlerinin analiz sonuçları

Ölçütler	**Kolon**
	miRNA +Proteom+ RNA-Seq
RVM	
AUC	0.555(0.016)
Doğruluk oranı	0.768(0.010)
F1 skoru	0.223(0.032)
MCC skoru	0.177(0.038)
Ada-boost RVM	
AUC	0.556(0.015)
Doğruluk oranı	0.773(0.011)
F1 skoru	0.227(0.032)
MCC skoru	0.191(0.039)
CANetwork	
AUC	0.513(0.022)
Doğruluk oranı	0.443(0.013)
F1 skoru	0.308(0.023)
MCC skoru	0.029(0.040)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

Dönüşüm tabanlı birleştirme yöntemi ile birleştirilen kolon verisinin analizinde kullanılan yöntemlerden en düşük performansa sahip yöntem CANetwork iken Ada-boost RVM yöntemi minimal bir farkla RVM'den yüksek olduğu gözlenmiştir.

Tablo 4.9. FS uygulanan Kolon verilerinde MKL, RF ve SVM yöntemlerin analiz sonuçları

Ölçütler	**FS uygulanan veri**			**Art arda birleştirilmiş veride FS uygulanan veri**	**Model tabanlı birleştirilmiş veride FS uygulanan veri**	**Tekli verilerde FS uygulanıp birleştirilen veri**
	miRNA	Proteom	RNA-Seq	miRNA+Proteom+RNA-Seq	miRNA+Proteom+RNA-Seq	miRNA +Proteom+ RNA-Seq
MKL						
AUC	0.812(0.006)	0.956(0.003)	0.516(0.002)	1.000(0.000)	-	0.538(0.004)
Doğruluk oranı	0.903(0.004)	0.979(0.001)	0.617(0.010)	1.000(0.000)	-	0.756(0.015)
F1 skoru	0.733(0.010)	0.947(0.004)	0.204(0.004)	1.000(0.000)	-	0.174(0.009)
MCC skoru	0.728(0.010)	0.940(0.004)	0.062(0.006)	1.000(0.000)	-	0.142(0.011)
RF						
AUC	1.000(0.000)	0.531(0.001)	0.536(0.001)	0.553(0.005)	1.000(0.000)	0.954(0.005)
Doğruluk oranı	1.000(0.000)	0.745(0.003)	0.749(0.003)	0.750(0.008)	1.000(0.000)	0.973(0.003)
F1 skoru	1.000(0.000)	0.169(0.002)	0.182(0.003)	0.233(0.011)	1.000(0.000)	0.929(0.008)
MCC skoru	1.000(0.000)	0.120(0.003)	0.133(0.003)	0.173(0.015)	1.000(0.000)	0.920(0.008)
SVM						
AUC	0.989(0.001)	0.574(0.004)	0.564(0.004)	0.637(0.008)	0.970(0.013)	0.779(0.011)
Doğruluk oranı	0.994(0.001)	0.727(0.004)	0.660(0.003)	0.778(0.006)	0.987(0.006)	0.869(0.007)
F1 skoru	0.988(0.001)	0.316(0.007)	0.277(0.005)	0.402(0.014)	0.964(0.016)	0.655(0.017)
MCC skoru	0.985(0.002)	0.161(0.008)	0.099(0.008)	0.307(0.017)	0.961(0.017)	0.607(0.019)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

mRMR değişken seçimi uygulanan kolon verilerinin analiz sonucuna göre değişken seçimi uygulanan kolon verilerin alt kümeleri dikkate alındığında

mRMR değişken seçimi uygulanan kolon miRNA verisinde kullanılan RF yönteminin performansı MKL ve SVM yöntemlerinin performansından daha yüksek olduğu gözlenmiş ve hatasız bir şekilde sınıflandırma işleminin gerçekleştirildiği görülmüştür.

mRMR değişken seçimi uygulanan kolon proteom verisinde kullanılan MKL yönteminin performansı RF ve SVM yöntemlerinin performansından yüksek olduğu gözlenmiştir.

mRMR değişken seçimi uygulanan kolon RNA-Seq verisinde kullanılan SVM yönteminin yüksek performansa sahip olduğu gözlenmiştir.

Art arda birleştirilen kolon verisinde mRMR değişken seçimi uygulanarak elde edilen kolon verisinde kullanılan MKL yönteminin performansı RF ve SVM yöntemlerinin performanslarından yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilen kolon verisinde mRMR değişken seçimi uygulanarak elde edilen kolon verisinde kullanılan RF yönteminin performansı SVM yönteminin performansından yüksek olduğu gözlenmiştir.

Tekli verilerde mRMR değişken seçimi uygulandıktan sonra birleştirilen veride uygulanan RF yönteminin performansı MKL ve SVM yöntemlerinin performanslarından yüksek olduğu gözlenmiştir.

Art arda birleştirilen veride mRMR değişken seçimi uygulanarak elde edilen veride MKL yönteminin performansı ve model tabanlı birleştirilmiş veride mRMR değişken seçimi uygulanarak elde edilen veride RF yönteminin performansı diğer verilerden yüksek çıktığı gözlenmiştir.

Tablo 4.10. FS-PCA uygulanan Kolon verilerinde MKL, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**FS+PCA uygulanan tekli veri**			**Art arda birleştirilmiş veride FS+PCA uygulanan veri**	**Model tabanlı birleştirilmiş veride FS+PCA uygulanan veri**	**Tekli verilerde FS+PCA uygulanıp birleştirilen veriler**
	miRNA	Proteom	RNA-Seq	miRNA+Proteom+RNA-Seq	miRNA+Proteom+RNA-Seq	miRNA +Proteom+ RNA-Seq
MKL						
AUC	0.800(0.005)	0.687(0.005)	0.585(0.004)	0.535(0.004)	-	0.543(0.007)
Doğruluk oranı	0.890(0.004)	0.848(0.004)	0.783(0.004)	0.749(0.015)	-	0.755(0.016)
F1 skoru	0.704(0.008)	0.504(0.010)	0.283(0.008)	0.169(0.009)	-	0.189(0.017)
MCC skoru	0.679(0.009)	0.519(0.010)	0.264(0.009)	0.129(0.012)	-	0.156(0.020)
RF						
AUC	0.609(0.004)	0.525(0.002)	0.588(0.004)	0.536(0.001)	0.690(0.016)	0.534(0.001)
Doğruluk oranı	0.801(0.004)	0.718(0.003)	0.782(0.004)	0.752(0.003)	0.863(0.009)	0.744(0.003)
F1 skoru	0.353(0.008)	0.196(0.003)	0.302(0.009)	0.172(0.002)	0.531(0.035)	0.176(0.003)
MCC skoru	0.339(0.009)	0.081(0.005)	0.270(0.010)	0.135(0.003)	0.540(0.033)	0.127(0.005)
SVM						
AUC	0.878(0.003)	0.536(0.001)	0.617(0.004)	0.603(0.005)	0.809(0.025)	0.775(0.005)
Doğruluk oranı	0.888(0.002)	0.752(0.003)	0.768(0.003)	0.695(0.004)	0.885(0.011)	0.830(0.003)
F1 skoru	0.762(0.004)	0.173(0.002)	0.358(0.007)	0.356(0.006)	0.688(0.042)	0.609(0.007)
MCC skoru	0.711(0.005)	0.135(0.002)	0.282(0.009)	0.188(0.008)	0.655(0.043)	0.537(0.009)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

mRMR değişken seçimi ve PCA yöntemi uygulanan kolon verilerin alt kümeleri dikkate alındığında miRNA ve RNA-Seq tekli verilerinde kullanılan SVM yönteminin performansları MKL ve RF yöntemlerinin performanslarından yüksek bulunmuştur. Proteom tekli verisinde MKL yönteminin performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek olduğu gözlenmiştir.

Art arda birleştirme yöntemiyle birleştirilen kolon verisinde mRMR değişken seçimi ve PCA yöntemi uygulanarak elde edilen veride kullanılan SVM yönteminin performansı MKL ve RF yöntemlerinin performansından daha yüksek olduğu gözlenmiş ve hatasız bir şekilde sınıflandırma işleminin gerçekleştirildiği görülmüştür.

Model tabanlı birleştirme yöntemiyle birleştirilen kolon verisinde mRMR değişken seçimi ve PCA yöntemi uygulanarak elde edilen veride kullanılan SVM yönteminin performansı RF yönteminin performansından daha yüksek olduğu gözlenmiştir.

Tekli verilerde mRMR değişken seçimi ve PCA uygulandıktan sonra birleştirilen veride uygulanan SVM yönteminin performansı MKL ve RF yöntemlerinin performanslarından yüksek olduğu gözlenmiştir.

Birleştirme yöntemlerinde model tabanlı birleştirilen veriye uygulanan SVM yönteminin performansının en yüksek olduğu gözlenmiştir.

Tablo 4.11. Ham böbrek verilerinde NSC, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**Ham veri**			**Art arda Birleştirilmiş ham Veri**	**Model tabanlı birleştirilmiş ham veri**
	Metilasyon	miRNA	RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq
NSC					
ROC	0.919(0.007)	0.941(0.006)	0.910(0.007)	0.915(0.011)	0.987(0.009)
Doğruluk oranı	0.955(0.004)	0.969(0.003)	0.941(0.004)	0.951(0.006)	0.995(0.003)
F1 skoru	0.972(0.002)	0.981(0.002)	0.962(0.003)	0.969(0.004)	0.997(0.002)
MCC skoru	0.857(0.012)	0.903(0.010)	0.823(0.012)	0.845(0.019)	0.983(0.012)
RF					
ROC	0.883(0.008)	0.909(0.008)	0.929(0.008)	0.890(0.013)	0.947(0.016)
Doğruluk oranı	0.948(0.003)	0.958(0.004)	0.972(0.003)	0.958(0.005)	0.981(0.006)
F1 skoru	0.969(0.002)	0.974(0.002)	0.983(0.002)	0.974(0.003)	0.989(0.003)
MCC skoru	0.817(0.014)	0.863(0.012)	0.906(0.010)	0.857(0.017)	0.932(0.020)
SVM					
ROC	0.922(0.003)	0.971(0.002)	0.926(0.003)	0.927(0.006)	0.960(0.015)
Doğruluk oranı	0.960(0.001)	0.988(0.001)	0.967(0.001)	0.969(0.003)	0.986(0.005)
F1 skoru	0.975(0.001)	0.993(0.001)	0.980(0.001)	0.981(0.002)	0.992(0.003)
MCC skoru	0.873(0.005)	0.962(0.003)	0.888(0.005)	0.900(0.009)	0.949(0.018)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

Böbrek verilerinin analiz sonucuna göre ham verilerin alt kümeleri dikkate alındığında metilasyon ve miRNA ham verilerinde kullanılan SVM yöntemlerinin performans değerleri NSC ve RF yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir. RNA-Seq ham verisinde RF yönteminin performans değeri NSC ve SVM yönteminin performans değerinden yüksek çıktığı gözlenmiştir.

Art arda birleştirilmiş ham veride kullanılan SVM yönteminin sınıflandırma performans değeri NSC ve RF yöntemlerinin sınıflandırma performans değerlerinden yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilmiş ham veride kullanılan NSC yönteminin sınıflandırma performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek olduğu gözlenmiştir.

Birleştirme yöntemlerinde en yüksek performans değeri ise model tabanlı birleştirilen ham veriye uygulanan NSC yöntemi olduğu gözlenmiştir.

Tablo 4.12. Dönüşüm tabanlı birleştirilmiş Böbrek verisinde RVM, Ada-boost RVM ve CANetwork yöntemlerinin analiz sonuçları

Ölçütler	**Böbrek**
	Metilasyon+ miRNA+ RNA-Seq
RVM	
AUC	0.927(0.022)
Doğruluk oranı	0.974(0.008)
F1 skoru	0.985(0.004)
MCC skoru	0.904(0.029)
Ada-boost RVM	
AUC	0.885(0.023)
Doğruluk oranı	0.958(0.008)
F1 skoru	0.975(0.005)
MCC skoru	0.845(0.030)
CANetwork	
AUC	0.746(0.030)
Doğruluk oranı	0.875(0.015)
F1 skoru	0.926(0.009)
MCC skoru	0.537(0.060)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

Dönüşüm tabanlı birleştirme yöntemi ile birleştirilen böbrek verisinin analizinde kullanılan yöntemlerden en yüksek performansa sahip yöntem RVM çıkmış ve Ada-boost RVM yönteminin performans değeri ise CANetwork yönteminin performans değerinden yüksek olduğu gözlenmiştir.

Tablo 4.13. FS uygulanan böbrek verilerinde MKL, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**FS uygulanan veri**			**Art Arda birleştirilmiş veride FS uygulanan Veri**	**Model tabanlı birleştirilmiş veride FS uygulanan veri**	**Tekli verilerde FS uygulanıp birleştirilen veriler**
	Metilasyon	miRNA	RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq
MKL						
AUC	0.916(0.019)	0.913(0.004)	0.802(0.006)	1.000(0.000)	-	0.940(0.011)
Doğruluk oranı	0.965(0.007)	0.965(0.002)	0.911(0.003)	1.000(0.000)	-	0.974(0.005)
F1 skoru	0.978(0.004)	0.978(0.001)	0.947(0.002)	1.000(0.000)	-	0.984(0.026)
MCC skoru	0.884(0.024)	0.885(0.005)	0.703(0.010)	1.000(0.000)	-	0.921(0.012)
RF						
AUC	0.853(0.005)	0.897(0.005)	0.720(0.006)	0.841(0.010)	0.953(0.018)	0.893(0.009)
Doğruluk oranı	0.930(0.002)	0.958(0.002)	0.876(0.003)	0.919(0.005)	0.984(0.006)	0.953(0.004)
F1 skoru	0.957(0.001)	0.975(0.001)	0.927(0.002)	0.950(0.003)	0.990(0.004)	0.971(0.003)
MCC skoru	0.757(0.008)	0.859(0.006)	0.547(0.012)	0.725(0.017)	0.939(0.024)	0.848(0.013)
SVM						
AUC	0.867(0.003)	0.981(0.001)	0.946(0.003)	0.901(0.006)	0.967(0.014)	0.927(0.006)
Doğruluk oranı	0.939(0.001)	0.986(0.001)	0.976(0.001)	0.932(0.004)	0.988(0.005)	0.969(0.003)
F1 skoru	0.963(0.001)	0.991(0.001)	0.985(0.001)	0.957(0.002)	0.993(0.003)	0.981(0.002)
MCC skoru	0.785(0.005)	0.961(0.002)	0.925(0.004)	0.779(0.012)	0.958(0.017)	0.900(0.009)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

mRMR değişken seçim yöntemi kullanılan Böbrek verilerinin analiz sonucuna göre miRNA ve RNA-Seq verilerinde kullanılan SVM yöntemlerinin performans değerleri MKL ve RF yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir. Metilasyon verisinde MKL yönteminin performans değeri RF ve SVM yönteminin performans değerinden yüksek çıktığı gözlenmiştir.

Art arda birleştirilmiş veride kullanılan MKL yönteminin sınıflandırma performans değeri RF ve SVM yöntemlerinin sınıflandırma performans değerlerinden yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilmiş veride kullanılan SVM yönteminin sınıflandırma performans değeri RF yönteminin performans değerinden yüksek olduğu gözlenmiştir.

Tekli verilerde mRMR uygulanıp birleştirilen verilerde MKL yönteminin performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir.

Birleştirme yöntemlerinden en iyi çıkan yöntem ise art arda birleştirilen veride mRMR uygulanarak elde edilen veride kullanılan MKL yöntemidir.

Tablo 4.14. FS-PCA uygulanan Böbrek verilerinde MKL, RF ve SVM yöntemlerin analiz sonuçları

Ölçütler	**FS+PCA uygulanan veri**			**Art arda birleştirilmiş veride FS+PCA uygulanan veri**	**Model tabanlı birleştirilmiş veride FS+PCA uygulanan veri**	**Tekli verilerde FS+PCA uygulanıp birleştirilen veriler**
	Metilasyon	miRNA	RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq
MKL						
AUC	0.916(0.004)	0.909(0.004)	0.752(0.006)	0.926(0.004)	-	0.919(0.004)
Doğruluk oranı	0.965(0.001)	0.960(0.002)	0.892(0.003)	0.967(0.002)	-	0.965(0.002)
F1 skoru	0.978(0.001)	0.975(0.001)	0.937(0.002)	0.980(0.001)	-	0.978(0.001)
MCC skoru	0.884(0.005	0.868(0.005)	0.627(0.010)	0.888(0.006)	-	0.886(0.005)
RF						
AUC	0.845(0.006)	0.898(0.006)	0.734(0.005)	0.853(0.006)	0.933(0.019)	0.807(0.007)
Doğruluk oranı	0.940(0.002)	0.965(0.002)	0.884(0.003)	0.940(0.002)	0.976(0.007)	0.912(0.003)
F1 skoru	0.965(0.001)	0.980(0.001)	0.931(0.002)	0.964(0.001)	0.986(0.004)	0.948(0.002)
MCC skoru	0.775(0.010)	0.857(0.008)	0.584(0.009)	0.788(0.009)	0.914(0.025)	0.683(0.012)
SVM						
AUC	0.856(0.003)	0.952(0.002)	0.834(0.005)	0.829(0.003)	0.950(0.015)	0.911(0.004)
Doğruluk oranı	0.894(0.002)	0.965(0.001)	0.930(0.002)	0.896(0.002)	0.979(0.006)	0.947(0.002)
F1 skoru	0.933(0.001)	0.978(0.001)	0.958(0.001)	0.936(0.001)	0.987(0.003)	0.966(0.001)
MCC skoru	0.680(0.007)	0.887(0.005)	0.760(0.007)	0.662(0.007)	0.928(0.020)	0.831(0.007)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

mRMR değişken seçim yöntemi ve PCA kullanılan Böbrek verilerinin analiz sonucuna göre miRNA ve RNA-Seq verilerinde kullanılan SVM yöntemlerinin performans değerleri MKL ve RF yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir. Metilasyon verisinde MKL yönteminin performans değeri RF ve SVM yönteminin performans değerinden yüksek çıktığı gözlenmiştir.

Art arda birleştirilmiş veride kullanılan MKL yönteminin sınıflandırma performans değeri RF ve SVM yöntemlerinin sınıflandırma performans değerlerinden yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilmiş veride kullanılan SVM yönteminin sınıflandırma performans değeri RF yönteminin performans değerinden yüksek olduğu gözlenmiştir.

Tekli verilerde mRMR uygulanıp birleştirilen verilerde MKL yönteminin performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir.

Birleştirme yöntemlerinden en iyi çıkan yöntem ise model tabanlı birleştirilen veride mRMR ve PCA uygulanarak elde edilen veride kullanılan SVM yöntemidir.

Tablo 4.15. Ham Tiroit verilerinde NSC, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**Ham veri**			**Art Arda Birleştirilmiş Ham Veri**	**Model tabanlı Birleştirilmiş Ham veri**
	Metilasyon	miRNA	RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq
NSC					
AUC	0.615(0.006)	0.702(0.003)	0.695(0.003)	0.680(0.007)	0.704(0.017)
Doğruluk oranı	0.699(0.003)	0.753(0.002)	0.735(0.002)	0.724(0.005)	0.773(0.009)
F1 skoru	0.364(0.016)	0.576(0.005)	0.568(0.004)	0.535(0.014)	0.546(0.041)
MCC skoru	0.232(0.010)	0.407(0.006)	0.384(0.005)	0.353(0.013)	0.416(0.033)
RF					
AUC	0.610(0.017)	0.636(0.009)	0.611(0.007)	0.631(0.013)	0.669(0.010)
Doğruluk oranı	0.717(0.012)	0.746(0.008)	0.717(0.009)	0.737(0.027)	0.782(0.003)
F1 skoru	0.392(0.043)	0.452(0.017)	0.413(0.018)	0.444(0.025)	0.514(0.020)
MCC skoru	0.290(0.031)	0.330(0.017)	0.289(0.017)	0.354(0.033)	0.415(0.012)
SVM					
AUC	0.637(0.002)	0.595(0.002)	0.647(0.002)	0.660(0.006)	0.665(0.008)
Doğruluk oranı	0.737(0.001)	0.665(0.001)	0.733(0.001)	0.747(0.003)	0.778(0.005)
F1 skoru	0.459(0.003)	0.423(0.002)	0.485(0.003)	0.500(0.002)	0.502(0.017)
MCC skoru	0.315(0.004)	0.192(0.003)	0.326(0.004)	0.355(0.009)	0.400(0.017)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

Tiroit verilerinin analiz sonucuna göre ham verilerin alt kümeleri dikkate alındığında miRNA ve RNA-Seq ham verilerinde kullanılan NSC yöntemlerinin performans değerleri RF ve SVM yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir. Metilasyon ham verisinde SVM yönteminin performans değeri NSC ve RF yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir.

Art arda birleştirilmiş ham veride kullanılan NSC yönteminin sınıflandırma performans değeri RF ve SVM yöntemlerinin sınıflandırma performans değerlerinden yüksek olduğu gözlenmiştir. Model tabanlı birleştirilmiş ham veride kullanılan NSC yönteminin sınıflandırma performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek olduğu gözlenmiştir.

Birleştirme yöntemlerinden en yüksek performansa sahip yöntem ise model tabanlı birleştirilen ham verilere uygulanan NSC yöntemidir.

Tablo 4.16. Dönüşüm tabanlı birleştirilmiş Tiroit verisinde RVM, Ada-boost RVM ve CANetwork yöntemlerinin analiz sonuçları

Ölçütler	**Tiroit**
	Metilasyon+ miRNA+ RNA-Seq
RVM	
AUC	0.675(0.010)
Doğruluk oranı	0.776(0.008)
F1 skoru	0.528(0.017)
MCC skoru	0.407(0.023)
Ada-boost RVM	
AUC	0.669(0.008)
Doğruluk oranı	0.775(0.007)
F1 skoru	0.515(0.013)
MCC skoru	0.402(0.018)
CANetwork	
AUC	0.694(0.012)
Doğruluk oranı	0.761(0.008)
F1 skoru	0.558(0.019)
MCC skoru	0.399(0.023)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

Dönüşüm tabanlı birleştirme yöntemi ile birleştirilen tiroit verisinin analizinde kullanılan yöntemlerden en yüksek performansa sahip yöntem CANetwork çıkmış ve RVM yönteminin performans değeri ise Ada-boost RVM yönteminin performans değerinden yüksek olduğu gözlenmiştir.

Tablo 4.17. FS uygulanan Tiroit verilerinde MKL, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**FS uygulanan veri**			**Art Arda Birleştirilmiş Veride FS uygulanan Veri**	**Model tabanlı Birleştirilmiş Veride FS uygulanan veri**	**Tekli verilerde FS uygulanıp birleştirilen veri**
	Metilasyon	miRNA	RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq
MKL						
AUC	1.000(0.000)	0.982(0.001)	-	1.000(0.000)	-	0.940(0.011)
Doğruluk oranı	1.000(0.000)	0.989(0.000)	-	1.000(0.000)	-	0.974(0.005)
F1 skoru	1.000(0.000)	0.981(0.001)	-	1.000(0.000)	-	0.984(0.026)
MCC skoru	1.000(0.000)	0.974(0.001)	-	1.000(0.000)	-	0.921(0.012)
RF						
AUC	0.585(0.001)	0.603(0.002)	0.603(0.002)	0.605(0.003)	0.636(0.010)	0.893(0.009)
Doğruluk oranı	0.722(0.002)	0.734(0.002)	0.729(0.001)	0.732(0.003)	0.771(0.005)	0.953(0.004)
F1 skoru	0.334(0.003)	0.373(0.004)	0.377(0.004)	0.380(0.007)	0.431(0.026)	0.971(0.003)
MCC skoru	0.241(0.003)	0.277(0.004)	0.271(0.003)	0.278(0.007)	0.366(0.020)	0.848(0.013)
SVM						
AUC	0.600(0.002)	0.669(0.002)	0.593(0.002)	0.581(0.003)	0.669(0.012)	0.927(0.006)
Doğruluk oranı	0.714(0.002)	0.763(0.002)	0.654(0.002)	0.684(0.003)	0.786(0.007)	0.969(0.003)
F1 skoru	0.397(0.003)	0.514(0.003)	0.429(0.003)	0.378(0.006)	0.504(0.023)	0.981(0.002)
MCC skoru	0.246(0.004)	0.392(0.003)	0.186(0.003)	0.182(0.007)	0.419(0.023)	0.900(0.009)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

mRMR değişken seçim yöntemi kullanılan Tiroit verilerinin analiz sonucuna göre metilasyon ve miRNA verilerinde kullanılan MKL yöntemlerinin performans değerleri RF ve SVM yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir. RNA-Seq verisinde MKL yönteminin performans değeri RF ve SVM yönteminin performans değerinden yüksek çıktığı gözlenmiştir.

Art arda birleştirilmiş veride kullanılan RF yönteminin sınıflandırma performans değeri SVM yönteminin sınıflandırma performans değerinden yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilmiş veride kullanılan SVM yönteminin sınıflandırma performans değeri RF yönteminin performans değerinden yüksek olduğu gözlenmiştir.

Tekli verilerde mRMR uygulanıp birleştirilen verilerde MKL yönteminin performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir.

Birleştirme yöntemlerinden en iyi çıkan yöntem ise art arda birleştirilen veride mRMR uygulanarak elde edilen veride kullanılan MKL yöntemidir.

Tablo 4.18. FS-PCA uygulanan Tiroit verilerinde MKL, RF ve SVM yöntemlerinin analiz sonuçları

Ölçütler	**FS+PCA uygulanan veri**			**Art Arda Birleştirilmiş Veride FS+PCA uygulanan veri**	**Model tabanlı birleştirilmiş veride FS+PCA uygulanan veri**	**Tekli verilerde FS+PCA uygulanan veri**
	Metilasyon	miRNA	RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq	Metilasyon+ miRNA+ RNA-Seq
MKL						
AUC	0.531(0.003)	0.913(0.005)	0.497(0.000)	0.500(0.001)	-	0.919(0.004)
Doğruluk oranı	0.407(0.009)	0.934(0.003)	0.385(0.007)	0.387(0.007)	-	0.965(0.002)
F1 skoru	0.468(0.004)	0.864(0.007)	0.376(0.007)	0.385(0.007)	-	0.978(0.001)
MCC skoru	0.067(0.010)	0.844(0.007)	-0.025(0.002)	-0.007(0.003)	-	0.886(0.005)
RF						
AUC	0.554(0.001)	0.612(0.002)	0.616(0.002)	0.549(0.002)	0.639(0.011)	0.807(0.007)
Doğruluk oranı	0.716(0.002)	0.739(0.002)	0.730(0.001)	0.711(0.002)	0.769(0.006)	0.912(0.003)
F1 skoru	0.226(0.004)	0.395(0.004)	0.414(0.004)	0.210(0.005)	0.440(0.027)	0.948(0.002)
MCC skoru	0.189(0.004)	0.296(0.004)	0.285(0.004)	0.166(0.004)	0.361(0.022)	0.683(0.012)
SVM						
AUC	0.580(0.002)	0.628(0.001)	0.633(0.002)	0.582(0.001)	0.661(0.012)	0.911(0.004)
Doğruluk oranı	0.722(0.001)	0.742(0.001)	0.734(0.001)	0.712(0.001)	0.779(0.006)	0.947(0.002)
F1 skoru	0.321(0.004)	0.434(0.003)	0.455(0.003)	0.337(0.003)	0.486(0.026)	0.966(0.001)
MCC skoru	0.230(0.004)	0.318(0.003)	0.308(0.004)	0.215(0.004)	0.397(0.024)	0.831(0.007)

Yöntemlerin performans değerleri ortalama(standart hata) olarak verilmiştir. AUC: Area Under The Curve.

mRMR değişken seçim yöntemi ve PCA kullanılanTiroit verilerinin analiz sonucuna göre metilasyon ve RNA-Seq verilerinde kullanılan SVM yöntemlerinin performans değerleri MKL ve RF yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir. miRNA verisinde kullanılan MKL yönteminin performans değeri RF ve SVM yönteminin performans değerlerinden yüksek çıktığı gözlenmiştir.

Art arda birleştirilmiş veride kullanılan SVM yönteminin sınıflandırma performans değeri MKL ve RF yöntemlerinin sınıflandırma performans değerlerinden yüksek olduğu gözlenmiştir.

Model tabanlı birleştirilmiş veride kullanılan SVM yönteminin sınıflandırma performans değeri RF yönteminin performans değerinden yüksek olduğu gözlenmiştir.

Tekli verilerde mRMR uygulanıp birleştirilen verilerde MKL yönteminin performans değeri RF ve SVM yöntemlerinin performans değerlerinden yüksek çıktığı gözlenmiştir.

Birleştirme yöntemlerinde en iyi çıkan yöntem ise tekli verilere mRMR değişken seçimi uygulanıp birleştirilen veriye PCA uygulanarak elde edilen veride kullanılan MKL yöntemidir.

5. TARTIŞMA ve SONUÇ

Bu tez çalışmasında çoklu omik verilerin birleştirilmesinde kullanılan yöntemler ve sınıflandırma analizlerinde kullanılan istatistiksel yöntemler üç farklı veri setinde değerlendirilmiştir. Bu üç farklı veri kümesi kapsamlı olacak şekilde çoklu omik verilerinin alt kümelerinde bulunan tekli verilerde, mRMR değişken seçimi uygulanan ve değişken seçiminden sonra temel bileşen analiz işlemi uygulanan verilerde, art arda birleştirilmiş verilerde, model tabanlı birleştirilmiş verilerde MKL, NSC, RF ve SVM sınıflandırma yöntemleri, dönüşüm tabanlı birleştirilmiş verilerde ise RVM, Ada-boost RVM ve CANetwork sınıflandırma yöntemlerinin performansları değerlendirilmiştir.

Yapılan literatür taramasında omik verilerinde üç farklı birleştirme yöntemlerinin performanslarının karşılaştırılmasının mevcut olduğu uygulama bildiğimiz kadarıyla bulunmamaktadır. Bu yüzden yaptığımız araştırmalar ve uygulamalarda öncelikli olarak omik verilerin tek başına ve bütünsel olarak farklı stratejilerde değerlendirilmesi sonucunda nasıl bir yol izlenmeli konusunda fikir sunmaya çalıştık. Tekli ve birleştirilmiş verilerde zaman ve maliyet açısından kullanılan yöntemler hakkında bilgiler sunmaya çalıştık.

Bulgularımız doğrultusunda seçilen birleştirme yöntemlerinin tercih sebebi verilere özgü özellikleri koruması, çok sayıda omik türlerini ve farklı ölçeklerdeki verileri entegre edebilmesidir. Veri analizlerinde kullanılan NSC, RF ve SVM sınıflandırma algoritmalarının tercih edilme sebebi ise genel olarak yüksek boyut probleminin olmaması ile birlikte çoklu omik verilerinde güçlü sonuçlar verebilmesi ve sorunsuz çalışmasıdır. MKL yönteminin tercih edilme nedeni ise değişken seçimine bağlı olarak verinin boyutunun indirgenmesiyle kullanılacak olan yeni düşük boyutlu verilerde iyi performans göstermesidir. Literatür taraması ve yaptığımız uygulamalar sonucunda bulduğumuz sonuçları aşağıda sırasıyla belirttik.

Chierici ve arkadaşlarının yaptığı çalışmada art arda birleştirme yöntemi kullanılarak RF ve SVM yöntemleriyle birleştirilmiş verilerin ve tekli verilerin performanslarını

araştırmışlar ve uygulamalarını gerçekleştirmişlerdir. Her bir stratejide RF ve SVM yöntemleri için benzer sonuçlar bulmuşlar. Bu sonuçlar doğrultusunda östrojen reseptörü durumunu, meme invaziv karsinom alt tiplerini ve renal berrak hücreli karsinomların sağkalımını tahmin etmek için kullandıkları sınıflandırma yöntemlerinin performanslarının birleştirilmiş verilerde iyi olduğunu göstermişlerdir. Miyeloid lösemi durumunu tahmin etmek için kullandıkları sınıflandırma yöntemlerinin performanslarının tekli verilerde iyi olduğunu göstermişlerdir (Chierici ve ark., 2020).

Bizde çalışmamızda literatürle uyumlu olan sonucumuzun SVM yönteminde art arda birleştirilmiş kolon verisinin performansının miRNA ve RNA-Seq tekli kolon verilerinin performansından yüksek, NSC yönteminde art arda birleştirilmiş böbrek verisinin performansının RNA-Seq tekli böbrek verisinin performansından yüksek, RF yönteminde art arda birleştirilmiş böbrek verisinin performansının metilasyon tekli böbrek verisinin performansından yüksek, SVM yönteminde metilasyon ve RNA-Seq tekli böbrek verilerinin performansından yüksek, NSC yönteminde art arda birleştirilmiş tiroit verisinin performansının metilasyon tekli tiroit verisinin performansından yüksek, RF yönteminde art arda birleştirilmiş tiroit verisinin performansının metilasyon ve RNA-Seq tekli tiroit verilerinin performansından yüksek, SVM yönteminde ise art arda birleştirilmiş tiroit verisinin performansının metilasyon, miRNA ve RNA-Seq verilerinin performansından yüksek olduğunu gözledik.

Kolon, böbrek ve tiroit verilerinden bazıları için tekli verilerde uygulanılan yöntemin performansının yüksek olduğunu görmüş olduk. Bunlardan kolon ve böbrek tekli verilerinden NSC, RF ve SVM yöntemlerinin performansının art arda birleştirilmiş kolon ve böbrek verilerinin performansından yüksek çıktığını, tiroit tekli verilerinde NSC ve RF yöntemlerinin performansının art arda birleştirilmiş tiroit verisinin performansından yüksek olduğunu gözledik.

Ma ve arkadaşları çoklu omik verilerini incelemişlerdir. Meme kanseri tanısının tahmini için verileri art arda birleştirme yöntemi kullanarak birleştirmişlerdir. Birleştirilmiş verilerdeki analizlerin daha bilgilendirici olduğunu belirtmişlerdir. Ancak inceledikleri bir diğer durum ise verilerin birleştirme stratejilerinde kullanılan yöntemlerin uygulanış sırası. Ma ve arkadaşları, verileri ham olarak birleştirip modeli eğitmişler, verileri ham olarak birleştirip değişken seçim yöntemini uyguladıktan sonra modeli eğitmişler, verileri birleştirmeden önce değişken seçim yöntemini uygulamışlar ve daha sonra verileri birleştirip modeli eğitmişlerdir. Sonuç olarak kullandıkları bu stratejiler için

birbirlerine göre üstünlüklerinin olmadığını benzer performans sergilediklerini belirtmişlerdir (Ma ve ark., 2016).

Bizim çalışmamızda art arda birleştirilen kolon, böbrek ve tiroit kanser verilerinin değişken seçiminin önce veya sonra uygulanmasında RF ve SVM yöntemlerinin performans değerlerinin birbirine yakın çıktığını birbirine göre üstünlükleri açısından belirgin bir fark olmadığını gözlemledik. Art arda birleştirilen verilerin ve art arda birleştirilen verilerde değişken seçimi uygulanan verilerin sınıflandırma performansları tekli verilerin sınıflandırma performanslarından yüksek veya düşük olabileceği yönünde sonuçlar elde edilmiştir.

Wilson ve arkadaşlarının yaptığı çalışmada, yumurtalık kanser verisi kullanılarak kanser tanısı konulan bireylerin üç yıldan daha uzun yaşayıp yaşayamacağı tahmin edilmektedir. Bu çalışmada MKL yöntemi ile klinik ve miRNA verilerinin entegre edilebildiği gösterilmektedir. Sonuç olarak miRNA verilerinin kullanılması klinik bilgilerle benzer tahmin doğruluğuna sahip olduğunu ancak her iki verinin de kullanılmasının önemli ölçüde daha yüksek doğruluk sağladığını ifade etmişlerdir (Wilson ve ark., 2019).

Başka bir çalışmada yapılan araştırmalar sonucunda Zhang ve arkadaşlarının Glioblastoma multiforme tanısını tahmin etmek için yaptığı çalışmada mRMR değişken seçim yöntemini kullanarak 71 tane gen ifade, 50 tane gen metilasyon, 3 tane miRNA, 4 tane gen kopyası olmak üzere toplam 130 değişkenden oluşan bir veri kümesini elde etmişler ve entegre bu veri kümesini çoklu çekirdek öğrenme yöntemi ile sınıflandırmayı önermişlerdir. Çalışmalarının sonucunda birleştirilen veride çoklu çekirdek öğrenme yönteminin diğer istatistiksel yöntemlere kıyasla GBM tanısının doğruluğunu arttırabileceğini göstermişlerdir (Zhang ve ark., 2016).

Biz de kendi çalışmamızda kolon ve böbrek verilerinde mRMR değişken seçimi kullanarak hem kolon verisi hem de böbrek verisi için metilasyon, miRNA ve RNA-Seq verilerini entegre ederek MKL yöntemi ile sınıflandırdık ve literatürle uyumlu olarak bizde birleştirilen veride çoklu çekirdek öğrenme yönteminin diğer istatistiksel yöntemlere kıyasla kolon ve böbrek kanserlerinin tanısının doğruluğunu arttırabileceği yönünde sonuçlar elde etmiş olduk. Bir diğer verimiz olan tiroit verisinde ise bu durum miRNA verisi için literatürle uyumlu bulunmuşken, metilasyon verisi için entegre edilen verinin sonucu ile aynı olduğunu ve literatürden farklı olduğunu bulduk. Bulduğumuz sonuçlardan yola çıkarak entegre veriler için uyarlanmış olan çoklu

çekirdek öğrenimi yönteminin hastalıkların sınıflandırılmasında performansı arttırdığını belirtmekle birlikte yeni teknolojilerden elde edilen verilerin çok gürültülü yapılara sahip olması ve bu verilerin normalize edilmesi, filtrelenmesi gibi ön işlem aşamalarından geçtiğini hatırlatarak bu aşamalardan birinde kaynaklı olabilecek problemlerden dolayı yöntemin performansının her zaman arttırıcı etki göstermeyebileceğini düşünerek yapılacak çalışmalarda kullanılacak veriler ve yöntemler için çok çeşitli verilerde pilot çalışmalar yaparak ön sonuçlarının değerlendirilmesinde fayda olabileceğini ifade etmekteyiz.

Dönüşüm tabanlı birleştirme yöntemi ile ilgili yapılan bir çalışmada GAW 19, ovaryum kanseri ve göğüs kanseri veri setleri kullanılarak grafik ve çekirdek tabanlı yöntemlerle sınıflandırma işlemini gerçekleştirmişlerdir. Yaptıkları çalışma sonucunda çekirdek tabanlı yöntemin grafik tabanlı yönteme nazaran daha iyi performans gösterdiğini ifade etmişlerdir. Ayrıca çekirdek tabanlı yöntemin hesaplanma süresinin grafik tabanlı yöntemin hesaplanma süresinden daha uzun olduğunu da belirtmişlerdir. Çekirdek tabanlı entegrasyon için örnekler arasındaki doğrusal olmayan ilişkileri tespit etmede iyi performans gösterdiğini ifade etmişlerdir. Küçük örneklem boyutu için olasılıklı tahmin sonucu almak isteyen çalışmacılar için RVM yönteminin tercih edilmesini belirtmişlerdir. Daha büyük örneklem boyutu (maksimum 300 örnek) için Ada-boost RVM kullanılması gerektiğini söylemişlerdir (Yan ve ark., 2017).

Biz de çalışmamızda dönüşüm tabanlı birleştirme yöntemlerinden grafik ve çekirdek tabanlı yöntemleri kolon, böbrek ve tiroit kanser verilerine uyguladık. Kolon ve böbrek verileri için çekirdek tabanlı birleştirme yönteminin performansı grafik tabanlı birleştirme yönteminin performansından daha iyi sonuç verdiğini ve literatürle benzer sonuçlar bulduk. Tiroit verisinde ise minimal farkla grafik tabanlı yöntemin performansını yüksek bulduk. Hesaplama süresi açısından tiroit verilerinde çekirdek tabanlı yöntemlerin daha uzun sürdüğü görülmüştür. Kolon verisinde Ada-boost RVM yönteminin hesaplama süresi, böbrek verisinde ise CANetwork yönteminin hesaplama süresinin diğer yöntemlerden uzun sürdüğünü bulduk. Çalışma sürelerinde bazı veriler için uygulanan yöntemlerin çalışma sürelerinde literatürle benzer olmayan sonuçlara rastladık. Buradan çalışmalarda kullanılan veri boyutlarına, çalışmaların analizlerinin gerçekleştirildiği bilgisayar veya iş istasyonlarının performans açısından farklı olmasına bağlı olarak farklı süre sonuçlarının elde edilebileceğini de göstermiş olduk.

Model tabanlı birleştirilmiş ham kolon verisinde RF ve SVM yöntemlerinin performansını miRNA, proteom ve RNA-Seq tekli kolon ham omik verilerinin performansından yüksek çıktığına rastladık. NSC yöntemi için miRNA, proteom ve RNA-Seq tekli ham kolon verilerinin performansını model tabanlı birleştirilmiş yöntemin performansından düşük olabileceği doğrultusunda bir sonuca rastladık. Değişken seçimi uygulanan model tabanlı birleştirilmiş kolon verisinde RF yönteminin performansı miRNA tekli kolon verisinin performansı ile aynı, proteom ve RNA-Seq tekli kolon verilerinin performansından yüksek çıktığını gözlemledik. Değişken seçimi uygulanan model tabanlı birleştirilmiş kolon verisinde SVM yönteminin performansı miRNA tekli kolon verisinin performansından düşük, proteom ve RNA-Seq tekli kolon verilerinin performansından yüksek olabileceği yönünde sonuçlara rastladık. Değişken seçimi ve temel bileşen analizi uygulanan model tabanlı birleştirilmiş kolon verisinde RF yönteminin performansı miRNA, proteom ve RNA-Seq verilerinin performansından yüksek, SVM yönteminin performansı proteom ve RNA-Seq tekli kolon verilerinin performansından yüksek, miRNA tekli kolon verisinin performansından düşük çıkabileceği görülnüştür.

Model tabanlı birleştirilmiş ham böbrek verisinde NSC ve RF yöntemlerinin performansını metilasyon, miRNA ve RNA-Seq tekli böbrek ham verilerinin performansından yüksek bulduk. Metilasyon ve RNA-Seq tekli ham böbrek verilerinde SVM yönteminin performansı model tabanlı birleştirilmiş yöntemin performansından düşük, miRNA tekli ham böbrek verisinin performansını model tabanlı birleştirilmiş yöntemin performansından yüksek çıkabileceği görülmüştür. Değişken seçimi uygulanan model tabanlı birleştirilmiş böbrek verisinde RF yönteminin performansı metilasyon, miRNA ve RNA-Seq tekli böbrek verilerinin performansından yüksek bulduk. Değişken seçimi uygulanan model tabanlı birleştirilmiş böbrek verisinde SVM ve yönteminin performansı metilasyon, ve RNA-Seq tekli böbrek verilerinin performansından yüksek, miRNA tekli böbrek verisinin performansından düşük olabileceğine rastladık. Değişken seçimi ve temel bileşen analizi uygulanan model tabanlı birleştirilmiş böbrek verisinde RF yönteminin performansı metilasyon, miRNA ve RNA-Seq verilerinin performansından yüksek, SVM yönteminin performansı metilasyon ve RNA-Seq tekli böbrek verilerinin performansından yüksek, miRNA tekli böbrek verisinin performansından düşük olabileceği yönünde sonuçlar elde edilmiştir.

Model tabanlı birleştirilmiş ham tiroit verisinde NSC, RF ve SVM yöntemlerinin performansını metilasyon, miRNA ve RNA-Seq tekli böbrek ham verilerinin performansından yüksek olabileceği yönünde sonuca rastladık. Değişken seçimi uygulanan model tabanlı birleştirilmiş tiroit verisinde RF yönteminin performansı metilasyon, miRNA ve RNA-Seq tekli tiroit verilerinin performansından yüksek olabileceği yönünde bir sonuç bulduk. Değişken seçimi uygulanan model tabanlı birleştirilmiş tiroit verisinde SVM yönteminin performansı metilasyon, ve RNA-Seq tekli tiroit verilerinin performansından yüksek, miRNA tekli tiroit verisinin performansını aynı olabileceğini görmüş olduk. Değişken seçimi ve temel bileşen analizi uygulanan model tabanlı birleştirilmiş tiroit verisinde RF ve SVM yöntemlerinin performansı metilasyon, miRNA ve RNA-Seq verilerinin performansından yüksek olabileceği görülmüştür.

Model tabanlı birleştirme yönteminde genel olarak birleştirilen verilerin hastalığın tahmininde tekli verilere göre daha yüksek doğrulukla gerçekleştirilebileceği söylenebilmektedir. Ayrıca model tabanlı birleştirilmiş kolon ve böbrek verileri için uygulanan değişken seçimi sınıflandırma tahmininin doğruluğunu arttırıcı etki yaratabileceği yönünde izlenimler görülmektedir. Ensemble öğrenme yöntemi, genel sınıflandırma doğruluğunu iyileştirmek için birden fazla öğrenme yöntemlerini birleştirmede etkili bir teknik olduğu yönünde görüşler bulunmaktadır (Dietterich, 2000). Yang ve arkadaşları bu yöntemin gelecekte kullanılmasıyla ilgili öngörüler sunmaktadırlar (Yang ve ark., 2010).

Willingale ve arkadaşları proteomik verilerin sınıflandırılmasında Destek Vektör Makine, Genetik Algoritma, Yapay Sinir Ağ ve RF gibi yöntemleri kullanmışlardır. Eğitim seti, kalp yetmezliği bulunan 100 birey ve 100 sağlıklı bireylerden oluşmaktadır. Test seti olarak 32 kalp yetmezliği olanlar ve 20 sağlıklı bireylerden oluşmaktadır. Kullandıkları yöntemlerin sınıflandırma doğruluğunun yöntemler için iyi olduğunu söylemişlerdir (Willingale ve ark., 2006).

Purohit ve Rocke proteomik verilerinde hasta ve kontrol gruplarında boyut indirgeme için temel bileşen analizini kullanmışlar ve verilerin görsel sınıflandırılmasında hiyerarşik kümeleme analizini kullanmışlardır. Ayrıca sınıflandırmada lojistik regresyon ve ayırma analiz yöntemleri ile PCR (Principal Components Regression) ve PLS yöntemlerinin kombinasyonlarını araştırmışlar ve PLS tabanlı sınıflandırma

yönteminin daha yüksek performansla sonuçlandığını göstermişlerdir (Purohit ve Rocke, 2003).

Tao ve arkadaşları tarafından yapılan çalışmada, tekli ve çoklu omik veri setlerini kullanarak meme kanseri alt tiplerinin sınıflandırılması amaçlanmaktadır. Kullandıkları MKL yönteminin sonucunda çoklu verilerin performanslarının tekli omik verilerinin performansına kıyasla doğruluğunun daha yüksek olduğunu göstermişlerdir (Tao ve ark., 2019).

Biz de kendi çalışmamızda tekli ve çoklu omik verilerinde farklı stratejileri değerlendirdik. Yapılan değerlendirmeler sonucunda tekli ham kolon verilerinde proteom verisi kullanılarak en yüksek doğruluğa sahip SVM yöntemi ile %62 performansla ayrım yapmanın mümkün olabileceği söylenebilmektedir.

Tekli ham böbrek verilerinde miRNA verisi kullanılarak en yüksek doğruluğa sahip SVM yöntemi ile %97 performansla ayrım yapmanın mümkün olabileceği söylenebilmektedir. Tekli ham tiroit verilerinde RNA-Seq verisi kullanılarak en yüksek doğruluğa sahip NSC yöntemi ile %70 performansla ayrım yapmanın mümkün olabileceği söylenebilmektedir. Tekli ham kolon verilerinde miRNA verisine uygulanan değişken seçimi ile hastalığın RF yöntemiyle sınıflandırma doğruluğunun hatasız bir şekilde sağlandığı görülmüştür. RF yönteminde kullanılan mtry, ntree ve nodesize gibi parametreler için uygun ayarlamalar yapıldığında yüksek performansa sahip model elde edilebilmektedir. Ayrıca RF yönteminin hem bu verilerde hem de diğer verilerde performanslarının iyi çıkması, kullandığı Bagging yöntemine bağlanabilmektedir. Çünkü oluşturulan model, düşük varyansa ve düşük yanlılığa sahiptir (Uriarte RD ve Andres SA, 2006).

Tekli ham böbrek verilerinde miRNA verisine uygulanan değişken seçimi ile hastalığın SVM yöntemiyle sınıflandırma doğruluğunun %98'e yükseldiğini gözlemledik.

Tekli ham kolon ve böbrek verilerine uygulanan değişken seçimi ve temel bileşen analizi uygulanan verilerde ham verilere göre performansı yüksek bulunurken, değişken seçimi uygulanan verilerin performansından düşük çıkabileceğini gözlemledik. Buradan bazı veriler için değişken seçimi yapılan sonuçların dikkate alınacağı, bazıları içinse değişken seçimi ve temel bileşen analizi uygulanan verilerin sonuçlarının dikkate alınması gerektiği konusunda analizlerin performanslarına bakarak bir fikir yürütebileceği yönünde bilgi sunmaya çalıştık. Ancak bu durum için genelleme yapamamaktayız. Genelleme yapabilmek için çok sayıda gerçek verilerde ve benzetim

verilerinde kapsamlı uygulamaların gerçekleştirilmesine ihtiyaç vardır. Ayrıca diğer değişken seçim ve değişken çıkarım yöntemlerinin de bu verilerde araştırılması genelleme sonuçlarının güvenilirliği konusunda güçlendirici nitelikte olmasına katkı sağlayabileceğini de göstermek adına çeşitli senaryo ve yöntemlerle kapsamlı çalışmalar gerçekleştirilmelidir.

Çalışmamızda uygulanan bir diğer nokta sınıflandırma yöntemlerinin kolon, böbrek ve tiroit verileri için uygulanan stratejiler doğrultusunda çalışma süreleri ve değişken sayılarıydı. NSC yöntemi kendi algoritmasının içinde uyguladığı değişken seçimi ile daha az değişkenle daha hızlı performans göstermektedir. RF ve SVM yöntemlerinde ise verilere değişken seçimi uygulandıktan sonra ham veriye göre daha düşük hızda çalıştığını görmüş olduk. Ayrıca değişken seçiminden sonra temel bileşen analizi uygulanan verilerde değişken sayılarının az olmasına rağmen çalışma sürelerinin değişken seçimi yapılan verilere göre daha uzun çalışabileceğini görmüş olduk.

Tekli ham verilerde zaman ve maliyet açısından en yüksek değere sahip olan yöntemler sırasıyla RF, SVM ve NSC, özellik seçimi ve temel bileşen analiz uygulanan tekli verilerde RF, MKL SVM, art arda ve model tabanlı birleştirilen ham verilerde RF, SVM ve NSC, birleştirilen değişken seçimi ve temel bileşen analizi uygulanan verilerde RF, MKL, SVM, model tabanlı birleştirilen değişken seçimi ve temel bileşen analizi uygulanan verilerde RF, SVM'dir.

Dönüşüm tabanlı birleştirilen verilerde kullanılan RVM, Ada-boost RVM ve CANetwork yöntemlerin zaman ve maliyet açısından art arda birleştirme ve model tabanlı birleştirme yöntemleri kullanılarak birleştirilen verilerde kullanılan yöntemlerden daha az zaman ve maliyetli olabileceğini göstermiş olduk. Sağlık alanında veya diğer alanlarda yapılan çalışmalarda araştırmacılar minimum maliyet minimum zaman ve maksimum doğrulukta bir modelle çalışmayı tercih etmek isteyebilirler. Bu doğrultuda araştırmacıların kullanmış oldukları sınıflandırma yöntemlerinin performansları aynı veya yakın çıktığı durumlarda zaman ve maliyet açısından değerlendirerek hangi yöntemleri kullanmaları gerektiği yönünde fikirler verebilmesi açısından elde ettiğimiz sonuçları sunmaya çalıştık.

Literatür taramamız ve yaptığımız çalışma doğrultusunda birleştirme yöntemlerinin uygulanmasında elde edilen performanslar veriden veriye, uygulanan strateji veya kullanılan sınıflandırma yöntemleri açısından farklılık gösterebileceğine rastladık. Ancak yapılan çalışmalarda çoklu omik veri entegrasyonunun altında yatan hipotez

çeşitli omik verilerinin tamamlayıcı bilgi sağlayabileceği (Cantini ve ark., 2020), bazı durumlarda gereksiz olmasına rağmen (Chai ve ark., 2019) tekli omik yapılarına göre hastalıkların altında yatan mekanizmaların daha iyi anlaşılması için daha kapsamlı bir anlayış sağlayabildiğini araştırmacılar ifade etmişlerdir (KArczewski ve ark., 2018). İfade ettikleri bu varsayım, kardiyovasküler hastalık (Mimila ve ark., 2019), diyabet (Prelot ve ark., 2018), karaciğer hastalığı (Chierico ve ark., 2017) veya mitokondriyal hastalıklar (Khan ve ark., 2020) gibi çeşitli hastalıklar üzerine yapılan çok sayıda çalışma ile doğrulanmıştır ve ayrıca yapısal olarak hastalık ne kadar karmaşıksa entegrasyonun o kadar avantajlı olduğunu düşündürmektedir (Tarazona ve ark., 2018). Birden fazla nedenin ve ilişkili olayların bir arada ortaya çıkması, hücre transformasyonuna bağlı olarak tümör oluşumu ve kanser gelişiminin iyi bilinen bir özelliği olduğundan, birden fazla kaynaktan üretilen verilerin entegrasyonu bu nedenle kanserin ayırt edici özelliklerinin belirlenmesi için özellikle yararlı olabileceği belirtilmiştir (Chakraborty ve ark., 2018; Cantafio ME ve ark.,2018;Sathyanarayanan ve ark., 2019; Liu ve ark., 2019). İfade edilen bu durumlar doğrultusunda verilerdeki değişkenleri baz alan birleştirme türü olan art arda birleştirme yönteminin kullanılmasının önem taşıyabileceğini ifade edebilmekteyiz. Bu bilgiler ışığında dönüşüm tabanlı birleştirme yönteminin bireysel genin fenotip üzerine etkisinin nasıl olacağı konusunda yetersiz kalabileceği göz önünde bulundurularak art arda birleştirme ve model tabanlı birleştirme yöntemlerinin bu açıdan kurtarıcı yöntemler olabileceği söylenebilmektedir. Veriler de kullanılan yöntemlerinin performanslarının iyileştirilmesi konusunda birleştirme yöntemlerine ek olarak değişken seçim yöntemlerinin veya boyut indirgeme yöntemlerinin kullanılmasında fayda olabileceği de belirtilmektedir.

Model tabanlı birleştirilen mRMR değişken seçimi uygulanan kolon ve böbrek verilerinde RF yönteminin çalışma süreleri, art arda birleştirilen mRMR değişken seçimi uygulanan kolon ve böbrek verilerinin çalışma sürelerinden daha hızlı çalıştığını ve daha iyi performans gösterdiğini bulduk. Model tabanlı birleştirilen mRMR değişken seçimi uygulanan tiroit verisinde ise performans açısından iyiyken çalışma süresi açısından daha yavaş çalıştığını bulduk.

Art arda birleştirilen mRMR değişken seçimi uygulanan kolon verisinde MKL yönteminin performansı RF yöntemine göre daha yüksek performans göstermekte ancak çalışma süresi açısından az bir farkla yavaş çalışmaktadır.

Art arda birleştirilen mRMR değişken seçimi uygulanan böbrek ve tiroit verilerinde kullanılan MKL sınıflandırma yönteminin performans değerinin RF yöntemine göre daha iyi olduğunu ve çalışma süresi açısından daha hızlı olduğunu bulduk.

Sonuç olarak sadece hastalığın doğru sınıflandırılması gerçekleştirilmek isteniyorsa model tabanlı birleştirilen veriler ve bu verilere değişken seçimi yöntemi uyguladıktan sonra elde edilen veriler kullanılabilir. Çalışma süresi ve değişken sayısını dikkate alarak hastalığın doğru sınıflandırılması gerçekleştirilmek isteniyorsa, en iyi performans gösteren ve nispeten diğer yöntemlerden daha hızlı olan MKL yöntemi, art arda birleştirilen verilere değişken seçim yöntemi uyguladıktan sonra kullanılabilir.

Genel olarak araştırmamız kapsamında sadece bizim uyguladığımız detaylı stratejilerle kesin yorumlar yapılamamaktadır. Kesin ifadeler kullanabilmek ve genelleme yapabilmek için daha fazla gerçek veri ve benzetim verileri ile daha kapsamlı denemelerin gerçekleştirilmesi önerilmektedir.

6. KAYNAKLAR

Ahmad A, Fröhlich H. Integrating Heterogeneous omics Data via Statistical Inference and Learning Techniques, Genomics and Computational Biology. 2016; 2(1): e32.

Archer KJ. Emprical characterization of random forest variable ımportance measure, computational statistical data analysis, CSDA, 2008, 52(4): 2249-2260.

Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Am Stat Assoc. 2006;101(473):119-137.

Barshan E, Ghodsi A, Azimifar Z, Zolghadri Jahromi M. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. Pattern Recognit. 2011;44(7):1357-1371.

Başaran E, Aras S, Cansaran-Duman D, Genomik, Proteomik, Metabolomik Kavramlarina Genel Bakış Ve Uygulama Alanları, Türk Hij ve Den Biyol Derg. 2010; 67(2):85-95.

Beale DJ, Pinu FR, Kouremenos KA, et al. Review of recent developments in GC-MS approaches to metabolomics-based research. Metabolomics. 2018;14(11):152.

Begley RJ, Riege M, Rosenblum J, Tseng D. Adding intelligence to medical devices. Medical Device & Diagnostic Industry Magazine 2000;22(3):150-173.

Belgiu M, Drăgu L. Random forest in remote sensing: A review of applications and future directions. ISPRS J Photogramm Remote Sens. 2016;114:24-31.

Blagus R, Lusa L. Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14(64):1-13.

Boulesteix AL, De Bin R, Jiang X, Fuchs M. IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. Comput Math Methods Med. 2017;2017(1):1-15.

Boulesteix AL, Janitza,S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2012; 2(6): 493–507.

Breiman L. Random Forests, Mach learn. 2001; 45(1): 5-32.

Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D. Benchmarking joint multi-omics dimensionality reduction approaches for cancer study, 2020.

Chai H, Zhou X, Cui Z, et al. Integrating multi-omics data with deep learning for predicting cancer prognosis, 2019:1-7.

Chakraborty S, Hosen MI, Ahmed M, Shekhar HU. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. BioMed Res Int. 2018; 2018 (1):1-15.

Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.

Chierici M, Bussola N, Marcolini A, Francescatto M, Alessandro Z, Lucia T, Claudio A, Giuseppe J, Cesare F. Integrative Network Fusion: a multi-omics approach in molecular profiling. Front. Oncol. 2020:1-14

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.

Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, Mpindi JP, Kallioniemi O, Honkela A, Aittokallio T, Wennerberg K, Community, NCID, Collins JJ, Gallahan D, Singer D, Saez-rodriguez J, Kaski S, Gray JW, Stolovitzky G. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32(12):1202-2014.

Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, England, 2000.

Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R, Overbeek R, Santerre J, Shukla M, Wattam AR, Will R, Xia F, Stevens R. Antimicrobial Resistance Prediction in PATRIC and RAST, Sci. Rep. 2016; 6(27930).

Del Chierico F, Nobili V, Vernocchi P, Russo A, De Stefanis C, Gnani D, Furlanello C, Zandonà A, Paci P, Capuani G, Dallapiccola B, Miccheli A, Alisi A, Putignani L. Gut microbiota profiling of pediatric nonalcoholic fatty liver

disease and obese patients unveiled by an integrated meta-omics-based approach. Hepatology. 2017; 65(2):451–464.

Díaz-uriarte R, Andrés SA De. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7(3):1-13.

Dietrich S, Floegel A, Weikert C, Pischon T, Boeing H, Drogan D. Metabolomics Identification of Serum Metabolites Associated With Incident Hypertension in the European Prospective Investigation into Cancer and Nutrition – Potsdam Study. Hypertension. 2016;68(2):471-477.

Dietterich TG. Ensemble methods in machine learning. Springer. 2000; 1–15.

Dougherty M. A review of neural networks applied to transport. Transp Res Part C. 1995;3(4):247-260.

Du KL, Swamy MNS. Neural networks in a softcomputing framework. Neural Networks a Softcomputing Framew. Published online 2006:1-566.

Dudoit S, Fridlyand J, Speed TP: Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc. 2002; 97(457):77–87. (8 kaynak klasöründeki 8.kaynak)

Foody GM, Mathur A. A relative evaluation of multiclass image classification by support vector machines. IEEE Trans Geosci Remote Sens. 2004;42(6):1335-1343.

Franzosa EA, Sirota-Madi A, Avila-Pacheco et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol. 2019; 4: 293–305

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000; 16(10): 906–914.

Fusaro VA, Mani D, Mesirov JP, et al. Prediction of highresponding peptides for targeted protein assays by mass-spectrometry. Nat Biotechnol 2009;27:190–8.

Gallo Cantafio ME, Grillone K, Caracciolo D, Scionti F, Arbitrio M, Barbieri V, et al. From Single Level Analysis to Multi-Omics Integrative Approaches: A Powerful Strategy towards the Precision Oncology. High-Throughput. 2018; 7(33):1-20.

Gao D, Zhang YX, Zhao YH. Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics. 2009; 9(2): 14–39.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286: 531–537.

Gönen M, Alpaydın E. Multiple Kernel Learning Algorithms. Journal of Machine Learning Research. 2011; (12):2211-2268.

Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2013; 27: 659–678.

Gunn SR. "Support Vector Machines for Classification and Regression", Technical Report, University of Southampton, 1998: 1-66.

Guo P, Luo Y, Mai G, et al. Gene expression profile based classification models of psoriasis. Genomics 2014;103(1):48–55.

Habermann JK, Doering J, Hautaniemi S, et al. The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. Int J Cancer. 2009;124 (7):1552–1564.

Hagan MT, Demuth HB, Beale M. Neural Network, Design (2.baskı), Boston, PWS, 1996:2-23, 81-82, 520-549, 578-580.

Hasin Y, Seldin M, and Lusis A. Multi-omics approaches to disease. Genome Biology, 2017; 18(1):83.

Hastie T, Tibshirani R, Friedman J, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2nd Edition, Springer, New York, USA, 2009.

Jack VT. Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. J Clin Epidemid. 1996;49(11):1225-1231.

Jain A. K., Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer. 1996;29(3):31-44.

Ji S, Fakhry A, Deng H. Integrative analysis of the connectivity and gene expression atlases in the mouse brain. Neuroimage. 2014;84:245–53.

Karczewski K, Snyder M. Integrative omics for health and disease. Nat Rev Genet. 2018; 19(5):299–310.

Keerthi SS, Lin C. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. Neural Comput. 2003;15(7):1667-1689.

Khan S, Ince-Dunn G, Suomalainen A, Elo LL. Integrative omics approaches provide biological and clinical insights: examples from mitochondrial diseases. J Clin Invest. 2020; 130(1): 20–28.

Kim M, Tagkopoulos I. Data integration and predictive modeling methods for multi-omics datasets. Mol Omi. 2018;14(1):8-25.

Klassen M, Cummings M, Saldana G. Investigation of random forest performance with cancer microarray data. Published 2008;1-6

Kloft M, Brefeld U, Laskov P, Sonnenburg S. Non-sparse multiple kernel learning. J. Mach. Learn. Res.2008; 1-4.

Kloft M, Brefeld U, Sonnenburg S, Zien A. Lp-norm multiple kernel learning. J. Mach. Learn. Res. 2011; 12: 953–997.

Kursa MB, Rudnicki WR. Feature Selection with Boruta Package. J Stat Softw. 2010;36(11):1-13.

Küçüksille E.U., Ateş N., Spam e-mail Filtering Using Support Vector Machine. TBBMD. 2016. 6.

Leon-mimila P, Wang J, Huertas-vazquez A. Relevance of Multi-Omics Studies in Cardiovascular Diseases. Front Cardiovasc. Med. 2019;6(91):1-13.

Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. Proc IEEE. 2016;104(1):176–197.

Lin E, Lane HY. Machine learning and systems genomics approaches for multi-omics data. Biomark Res. 2017; 5:2.

Liu X, Wang L, Zhu X, et al. Absent Multiple Kernel Learning Algorithms. IEEE Trans Pattern Anal. Mach. Intell. 2020;42(6):1303-1316.

Liu SH, Shen PC, Chen CY, Hsu AN, Cho YC, Lai YL, Chen FH, Li CY, Wang SC, Chen M, Chung IF, Cheng WC. DriverDBv3: a multi-omics database for cancer driver gene research. Nucleic Acids Res. 2020; 48(D1):D863–D870.

Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants, Trends Plant Sci.2014; 19(12): 798–808.

Marco-Ramell A, Palau-Rodriguez M, Alay A, Tulipani S, Urpi-Sarda M, Sanchez-Pla A, Andres Lacueva C. Evaluation And Comparison Of Bioinformatic Tools For The Enrichment Analysis Of Metabolomics Data. BMC Bioinformatics, 2018;19(1):1-11.

Mariot A, Sgoifo S, Sauli M. I gozzi endotoracici: contributo casistico-clinico (20 casi). Friuli Med. 1964;19(6).

McCulloch WS, Pitts WA. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biol. 1943; (5): 115-133.

Melgani F, Bruzzone L. Classification of Hyperspectral Remote Sensing Images With Support Vector Machines. IEEE Trans Geosci Remote Sens. 2004;42(8):1778-1790.

Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 2016;17(4):628-641.

Moler E, Chow M, Mian I. Analysis of molecular profile data using generative and discriminative methods. Physiol Genomics. 2000; 4(2): 109–126.

Mostafavi S, Morris Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics. 2010; 26(14):1759–65.

Ning J, Beiko RG. Phylogenetic approaches to microbial community classification. Microbiome 2015;3:1

Osuna E, Freund R, Girosi F. Support Vector Machines: Training and Applications. 1997; 1602(144).

Pal M. Random Forest For Land Cover Classification, IEEE Geoscience and Remote Sensing Symposium. 2003;6:3510–3512.

Pal M, Mather PM. Support vector machines for classification in remote sensing. Int J Remote Sens. 2005; 26(5):1007-1011.

Papik K, Molnar B, Schaefer R, Dombovari Z, Tulassay Z, Feher J. Application of neural networks in medicine - A review. Med Sci. Monit. 1998; 4(3): 538-54.

Perez-Riverol Y, Kuhn M, Vizcaíno JA, Hitz MP, Audain E. Accurate and fast feature selection workflow for high-dimensional omics data. PLoS One. 2017;12(12):1-14.

Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, Van Steen K. Integration Analysis of Three Omics Data Using Penalized Regression Methods: An Application to Bladder Cancer. PLoS Genet. 2015;11(12): 709.

Pinu FR. Metabolomics: Applications to Food Safety and Quality Research. In Microbial Metabolomics: Publishing: Cham, Switzerland, 2016; 225–259.

Pinu RF. Grape and Wine Metabolomics to Develop New Insights Using Untargeted and Targeted Approaches. Fermentation 2018; 4(4): 92.

Prélot L, Draisma H, Anasanti M, Balkhiyarova Z, Wielscher M, Yengo L, et al. Machine Learning in Multi-Omics Data to Assess Longitudinal Predictors of Glycaemic Trait Levels. Genetic Epidemiology 2018; 42(7): 1-21.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012; 40: D130–D135.

Purohit P.V, Rocke D.M, Discriminant models for high□throughput proteomics mass spectrometer data. Proteomics. 2003;3(9):1699-703.

Rakotomamonjy A, Bach FR, Canu S, Grandvalet Y. “SimpleMKL,” JMLR. 2008; 9: 2491–2521.

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015; 16(2):85–97.

Rudaz, S. (Ed.). 2015. "Identification and data-processing methods in metabolomics". London, UK: Future Science Ltd, ISBN (online): 978-1-910420-28-7.

Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Brief Bioinform Advance Access. 2019; 1-17.

Saulnier DM, Riehle K, Mistretta T, et al. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology. 2011;141(5):1782-1791.

Schwan RF, Wheals AE. The microbiology of cocoa fermentation and its role in chocolate quality, Crit. Rev. Food Sci. Nutr. 2004; 44(4): 205–221.

Shaik R, Ramakrishna W. Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice, Plant Physiol. 2014; 164(1): 481–495.

Sherrod PH. Classification and Regression Trees and Support Vector Machines for Predictive Modeling and Forecasting. 2003.

Spicker JS, Brunak S, Frederiksen KS, Toft H. Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. Toxicol Sci. 2008;102(2):444-454.

Sweetlove LJ, Last RL, Fernie AR. Predictive metabolic engineering: a goal for systems biology, Plant Physiol. 2003; 132(2):420–425.

Tarazona S, Balzano-Nogueira L, Conesa A. Multiomics Data Integration in Time Series Experiments (Elsevier), Compr Anal Chem. 2018; 82(18): 505- 532.

Tao M., Song T, Du W, Han S, Zuo C, Li Y, Wang Y, Yang Z. Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data. Genes. 2019; 10(3): 200.

Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc NatAcad Sci USA. 2002; 99(10):6567–6572.

Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNAmicroarrays. Stat Sci. 2003; 18:104–117

Tipping ME, Faul AC. Fast marginal likelihood maximisation for sparse Bayesian models. AISTATS; 2003. 1-13.

Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1(3):211–44.

Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996; 49(11):1225-1231.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001; 98 *(9):* 5116-5121.

Uriarte RD, Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7(3).

Vasaikar S, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. 2017;46(D1):D956-D963.

Wani N, Raza K. Integrative approaches to reconstruct regulatory networks from multi-omics data: A review of state-of-the-art methods. Comput Biol Chem. 2019;83:1-20.

Willingale R, Jones D.J.L, Lamb J.H, Quinn P, Farmer P.B, Ng L.L. Searching for biomarkers of heart failure in the mass spectra of blood plasma. Proteomics. 2006;6:5903-5914.

Wilson CM, Li K, Yu X, Kuan PF, Wang X. Multiple-kernel learning for genomic data mining and prediction. BMC Bioinformatics. 2019; 20(1):426.

Yan KK, Zhao H, Pang H. A comparison of graph- and kernel-based – omics data integration algorithms for classifying complex traits. BMC Bioinformatics. 2017;18(539):1-13.

Yang P, Yang YH, Zhou B, Zomaya AY. A review of ensemble methods in bioinformatics. 2010; 5: 296-308.

Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, Suez J, Mahdi JA, Matot E, Malka G, Kosower N, Rein M, Zilberman-Schapira G, Dohnalová L, Pevsner-Fischer M, Bikovsky R, Halpern Z, Elinav E, Segal E. Personalized Nutrition by Prediction of Glycemic Responses, Cell. 2015; 163(5): 1079–1094.

Zhang Y, Li A, Peng C, Wang M. Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning. IEEE/ACM Trans Comput Biol Bioinforma. 2016;13(5):825-835.

Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. Adv Neural Inf Proces Syst. 2004;16(16):321–8.

EK-1. Kolon Veri Analiz Sonucu

Ölçütler	Ham veri			Art Arda Birleştirilmiş Ham Veri	Model tabanlı Birleştirilmiş Ham veri	FS uygulanan veri			Art Arda Birleştirilmiş Veride FS uygulanan Veri	Model tabanlı Birleştirilmiş Veride FS uygulanan veri	FS+PCA uygulanan veri			Art Arda Birleştirilmiş Veride FS+PCA uygulanan veri	Model tabanlı birleştirilmiş veride FS+PCA uygulanan veri	Ölçütler	Kolon	Ölçütler	FS uygulanan veri	FS+PCA uygulanan veri
	miRNA	Proteom	RNAseq	miRNA+Prcteom+RNAseq	miRNA+Proteom+RNAseq	miRNA	Proteom	RNAseq	miRNA+Proteom+RNAseq	miRNA+Proteom+RNAseq	miRNA	Proteom	RNAseq	miRNA+Proteom+RNAseq	miRNA+Proteom+RNAseq		miRNA+Proteom+RNAseq		miRNA+Proteom+RNAseq	miRNA+Proteom+RNAseq
MKL																**RVM**		**MKL**		
AUC	-	-	-	-	-	0.812 (0.006)	0.956 (0.003)	0.516 (0.002)	1.000 (0.000)	-	0.800 (0.005)	0.687 (0.005)	0.585 (0.004)	0.535 (0.004)	-	AUC	0.555 (0.016)	AUC	0.538 (0.004)	0.543 (0.007)
Doğruluk oranı	-	-	-	-	-	0.903 (0.004)	0.979 (0.001)	0.617 (0.010)	1.000 (0.000)	-	0.890 (0.004)	0.848 (0.004)	0.783 (0.004)	0.749 (0.015)	-	Doğruluk oranı	0.768 (0.010)	Doğruluk oranı	0.756 (0.015)	0.755 (0.016)
F1 skoru	-	-	-	-	-	0.733 (0.010)	0.947 (0.004)	0.204 (0.004)	1.000 (0.000)	-	0.704 (0.008)	0.504 (0.010)	0.283 (0.008)	0.169 (0.009)	-	F1 skoru	0.223 (0.032)	F1 skoru	0.174 (0.009)	0.189 (0.017)
MCC skoru	-	-	-	-	-	0.728 (0.010)	0.940 (0.004)	0.062 (0.006)	1.000 (0.000)	-	0.679 (0.009)	0.519 (0.010)	0.264 (0.009)	0.129 (0.012)	-	MCC skoru	0.177 (0.038)	MCC skoru	0.142 (0.011)	0.156 (0.020)
NSC																**Ada-boost RVM**		**RF**		
AUC	0.537 (0.001)	0.542 (0.002)	0.537 (0.001)	0.532 (0.004)	0.534 (0.000)	-	-	-	-	-	-	-	-	-	-	AUC	0.556 (0.015)	AUC	0.954 (0.005)	0.534 (0.001)
Doğruluk oranı	0.754 (0.005)	0.754 (0.005)	0.754 (0.005)	0.743 (0.011)	0.762 (0.000)	-	-	-	-	-	-	-	-	-	-	Doğruluk oranı	0.773 (0.011)	Doğruluk oranı	0.973 (0.003)	0.744 (0.003)
F1 skoru	0.173 (0.003)	0.179 (0.004)	0.173 (0.003)	0.173 (0.005)	0.167 (0.000)	-	-	-	-	-	-	-	-	-	-	F1 skoru	0.227 (0.032)	F1 skoru	0.929 (0.008)	0.176 (0.003)
MCC skoru	0.140 (0.004)	0.144 (0.004)	0.140 (0.004)	0.128 (0.010)	0.138 (0.000)	-	-	-	-	-	-	-	-	-	-	MCC skoru	0.191 (0.039)	MCC skoru	0.920 (0.008)	0.127 (0.005)
RF																**CANetwork**		**SVM**		
AUC	0.551 (0.004)	0.544 (0.004)	0.553 (0.004)	0.540 (0.005)	0.739 (0.011)	1.000 (0.000)	0.531 (0.001)	0.536 (0.001)	0.553 (0.005)	1.000 (0.000)	0.609 (0.004)	0.525 (0.002)	0.588 (0.004)	0.536 (0.001)	0.690 (0.016)	AUC	0.513 (0.022)	AUC	0.779 (0.011)	0.775 (0.005)
Doğruluk oranı	0.762 (0.007)	0.751 (0.007)	0.761 (0.007)	0.754 (0.011)	0.877 (0.018)	1.000 (0.000)	0.745 (0.003)	0.749 (0.003)	0.750 (0.008)	1.000 (0.000)	0.801 (0.004)	0.718 (0.003)	0.782 (0.004)	0.752 (0.003)	0.863 (0.009)	Doğruluk oranı	0.443 (0.013)	Doğruluk oranı	0.869 (0.007)	0.830 (0.003)
F1 skoru	0.202 (0.008)	0.201 (0.009)	0.208 (0.009)	0.183 (0.011)	0.635 (0.032)	1.000 (0.000)	0.169 (0.002)	0.182 (0.003)	0.233 (0.011)	1.000 (0.000)	0.353 (0.008)	0.196 (0.003)	0.302 (0.009)	0.172 (0.002)	0.531 (0.035)	F1 skoru	0.308 (0.023)	F1 skoru	0.655 (0.017)	0.609 (0.007)
MCC skoru	0.174 (0.010)	0.146 (0.011)	0.179 (0.011)	0.145 (0.015)	0.604 (0.060)	1.000 (0.000)	0.120 (0.003)	0.133 (0.003)	0.173 (0.015)	1.000 (0.000)	0.339 (0.009)	0.081 (0.005)	0.270 (0.010)	0.135 (0.003)	0.540 (0.033)	MCC skoru	0.029 (0.040)	MCC skoru	0.607 (0.019)	0.537 (0.009)
SVM																				
AUC	0.552 (0.003)	0.622 (0.004)	0.597 (0.004)	0.602 (0.007)	0.711 (0.027)	0.989 (0.001)	0.574 (0.004)	0.564 (0.004)	0.637 (0.008)	0.970 (0.013)	0.878 (0.003)	0.536 (0.001)	0.617 (0.004)	0.603 (0.005)	0.809 (0.025)					
Doğruluk oranı	0.732 (0.003)	0.778 (0.003)	0.756 (0.003)	0.774 (0.007)	0.855 (0.014)	0.994 (0.001)	0.727 (0.004)	0.660 (0.003)	0.778 (0.006)	0.987 (0.006)	0.888 (0.002)	0.752 (0.003)	0.768 (0.003)	0.695 (0.004)	0.885 (0.011)					
F1 skoru	0.253 (0.004)	0.370 (0.007)	0.324 (0.007)	0.337 (0.014)	0.538 (0.051)	0.988 (0.001)	0.316 (0.007)	0.277 (0.005)	0.402 (0.014)	0.964 (0.016)	0.762 (0.004)	0.173 (0.002)	0.358 (0.007)	0.356 (0.006)	0.688 (0.042)					
MCC skoru	0.124 (0.006)	0.280 (0.008)	0.238 (0.008)	0.282 (0.017)	0.510 (0.052)	0.985 (0.002)	0.161 (0.008)	0.099 (0.008)	0.307 (0.017)	0.961 (0.017)	0.711 (0.005)	0.135 (0.002)	0.282 (0.009)	0.188 (0.008)	0.655 (0.043)					

Ölçütler	Ham veri			Art Arda Birleştirilmiş Ham Veri	Model tabanlı Birleştirilmiş Ham veri	FS uygulanan veri			Art Arda Birleştirilmiş Veride FS uygulanan Veri	Model tabanlı Birleştirilmiş Veride FS uygulanan veri	FS+PCA uygulanan veri			Art Arda Birleştirilmiş Veride FS+PCA uygulanan veri	Model tabanlı birleştirilmiş veride FS+PCA uygulanan veri
	Metilasyon	miRNA	RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon	miRNA	RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon	miRNA	RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq
MKL															
AUC	-	-	-	-	-	0.916 (0.019)	0.913 (0.004)	0.802 (0.006)	1.000 (0.000)	-	0.916 (0.004)	0.909 (0.004)	0.752 (0.006)	0.926 (0.004)	-
Doğruluk oranı	-	-	-	-	-	0.965 (0.007)	0.965 (0.002)	0.911 (0.003)	1.000 (0.000)	-	0.965 (0.001)	0.960 (0.002)	0.892 (0.003)	0.967 (0.002)	-
F1 skoru	-	-	-	-	-	0.978 (0.004)	0.978 (0.001)	0.947 (0.002)	1.000 (0.000)	-	0.978 (0.001)	0.975 (0.001)	0.937 (0.002)	0.980 (0.001)	-
MCC skoru	-	-	-	-	-	0.884 (0.024)	0.885 (0.005)	0.703 (0.010)	1.000 (0.000)	-	0.884 (0.005	0.868 (0.005)	0.627 (0.010)	0.888 (0.006)	-
NSC															
AUC	0.919 (0.007)	0.941 (0.006)	0.910 (0.007)	0.915 (0.011)	0.987 (0.009)	-	-	-	-	-	-	-	-	-	-
Doğruluk oranı	0.955 (0.004)	0.969 (0.003)	0.941 (0.004)	0.951 (0.006)	0.995 (0.003)	-	-	-	-	-	-	-	-	-	-
F1 skoru	0.972 (0.002)	0.981 (0.002)	0.962 (0.003)	0.969 (0.004)	0.997 (0.002)	-	-	-	-	-	-	-	-	-	-
MCC skoru	0.857 (0.012)	0.903 (0.010)	0.823 (0.012)	0.845 (0.019)	0.983 (0.012)	-	-	-	-	-	-	-	-	-	-
RF															
AUC	0.883 (0.008)	0.909 (0.008)	0.929 (0.008)	0.890 (0.013)	0.947 (0.016)	0.853 (0.005)	0.897 (0.005)	0.720 (0.006)	0.841 (0.010)	0.953 (0.018)	0.845 (0.006)	0.898 (0.006)	0.734 (0.005)	0.853 (0.006)	0.933 (0.019)
Doğruluk oranı	0.948 (0.003)	0.958 (0.004)	0.972 (0.003)	0.958 (0.005)	0.981 (0.006)	0.930 (0.002)	0.958 (0.002)	0.876 (0.003)	0.919 (0.005)	0.984 (0.006)	0.940 (0.002)	0.965 (0.002)	0.884 (0.003)	0.940 (0.002)	0.976 (0.007)
F1 skoru	0.969 (0.002)	0.974 (0.002)	0.983 (0.002)	0.974 (0.003)	0.989 (0.003)	0.957 (0.001)	0.975 (0.001)	0.927 (0.002)	0.950 (0.003)	0.990 (0.004)	0.965 (0.001)	0.980 (0.001)	0.931 (0.002)	0.964 (0.001)	0.986 (0.004)
MCC skoru	0.817 (0.014)	0.863 (0.012)	0.906 (0.010)	0.857 (0.017)	0.932 (0.020)	0.757 (0.008)	0.859 (0.006)	0.547 (0.012)	0.725 (0.017)	0.939 (0.024)	0.775 (0.010)	0.857 (0.008)	0.584 (0.009)	0.788 (0.009)	0.914 (0.025)
SVM															
AUC	0.922 (0.003)	0.971 (0.002)	0.926 (0.003)	0.927 (0.006)	0.960 (0.015)	0.867 (0.003)	0.981 (0.001)	0.946 (0.003)	0.901 (0.006)	0.967 (0.014)	0.856 (0.003)	0.952 (0.002)	0.834 (0.005)	0.829 (0.003)	0.950 (0.015)
Doğruluk oranı	0.960 (0.001)	0.988 (0.001)	0.967 (0.001)	0.969 (0.003)	0.986 (0.005)	0.939 (0.001)	0.986 (0.001)	0.976 (0.001)	0.932 (0.004)	0.988 (0.005)	0.894 (0.002)	0.965 (0.001)	0.930 (0.002)	0.896 (0.002)	0.979 (0.006)
F1 skoru	0.975 (0.001)	0.993 (0.001)	0.980 (0.001)	0.981 (0.002)	0.992 (0.003)	0.963 (0.001)	0.991 (0.001)	0.985 (0.001)	0.957 (0.002)	0.993 (0.003)	0.933 (0.001)	0.978 (0.001)	0.958 (0.001)	0.936 (0.001)	0.987 (0.003)
MCC skoru	0.873 (0.005)	0.962 (0.003)	0.888 (0.005)	0.900 (0.009)	0.949 (0.018)	0.785 (0.005)	0.961 (0.002)	0.925 (0.004)	0.779 (0.012)	0.958 (0.017)	0.680 (0.007)	0.887 (0.005)	0.760 (0.007)	0.662 (0.007)	0.928 (0.020)

Ölçütler	Böbrek	Ölçütler	FS uygulanan veri	FS+PCA uygulanan veri
	Metilasyon+ miRNA+ RNAseq		Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq
RVM		**MKL**		
AUC	0.927 (0.022)	AUC	0.940 (0.011)	0.919 (0.004)
Doğruluk oranı	0.974 (0.008)	Doğruluk oranı	0.974 (0.005)	0.965 (0.002)
F1 skoru	0.985 (0.004)	F1 skoru	0.984 (0.026)	0.978 (0.001)
MCC skoru	0.904 (0.029)	MCC skoru	0.921 (0.012)	0.886 (0.005)
Ada-boost RVM		**RF**		
AUC	0.885 (0.023)	AUC	0.893 (0.009)	0.807 (0.007)
Doğruluk oranı	0.958 (0.008)	Doğruluk oranı	0.953 (0.004)	0.912 (0.003)
F1 skoru	0.975 (0.005)	F1 skoru	0.971 (0.003)	0.948 (0.002)
MCC skoru	0.845 (0.030)	MCC skoru	0.848 (0.013)	0.683 (0.012)
CANetwork		**SVM**		
AUC	0.746 (0.030)	AUC	0.927 (0.006)	0.911 (0.004)
Doğruluk oranı	0.875 (0.015)	Doğruluk oranı	0.969 (0.003)	0.947 (0.002)
F1 skoru	0.926(0.009)	F1 skoru	0.981 (0.002)	0.966 (0.001)
MCC skoru	0.537 (0.060)	MCC skoru	0.900 (0.009)	0.831 (0.007)

EK-3. Tiroit Veri Analiz Sonucu

Ölçütler	Ham veri			Art Arda Birleştirilmiş Ham Veri	Model tabanlı Birleştirilmiş Ham veri	FS uygulanan veri			Art Arda Birleştirilmiş Veride FS uygulanan Veri	Model tabanlı Birleştirilmiş Veride FS uygulanan veri	FS+PCA uygulanan veri			Art Arda Birleştirilmiş Veride FS+PCA uygulanan veri	Model tabanlı birleştirilmiş veride FS+PCA uygulanan veri	Ölçütler	Tiroit	Ölçütler	FS uygulanan veri	FS+PCA uygulanan veri
	Metilasyon	miRNA	RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon	miRNA	RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon	miRNA	RNAseq	Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq		Metilasyon+ miRNA+ RNAseq		Metilasyon+ miRNA+ RNAseq	Metilasyon+ miRNA+ RNAseq
MKL																**RVM**		**MKL**		
AUC	-	-	-	-	-	1.000 (0.000)	0.982 (0.001)	-	1.000 (0.000)	-	0.531 (0.003)	0.913 (0.005)	0.497 (0.000)	0.500 (0.001)	-	AUC	0.675 (0.010)	AUC	0.505 (0.005)	0.497 (0.000)
Doğruluk oranı	-	-	-	-	-	1.000 (0.000)	0.989 (0.000)	-	1.000 (0.000)	-	0.407 (0.009)	0.934 (0.003)	0.385 (0.007)	0.387 (0.007)	-	Doğruluk oranı	0.776 (0.008)	Doğruluk oranı	0.399 (0.026)	0.385 (0.007)
F1 skoru	-	-	-	-	-	1.000 (0.000)	0.981 (0.001)	-	1.000 (0.000)	-	0.468 (0.004)	0.864 (0.007)	0.376 (0.007)	0.385 (0.007)	-	F1 skoru	0.528 (0.017)	F1 skoru	0.379 (0.026)	0.376 (0.007)
MCC skoru	-	-	-	-	-	1.000 (0.000)	0.974 (0.001)	-	1.000 (0.000)	-	0.067 (0.010)	0.844 (0.007)	-0.025 (0.002)	-0.007 (0.003)	-	MCC skoru	0.407 (0.023)	MCC skoru	-0.009 (0.012)	-0.025 (0.002)
NSC																**Ada-boost RVM**		**RF**		
AUC	0.615 (0.006)	0.702 (0.003)	0.695 (0.003)	0.680 (0.007)	0.704 (0.017)	-	-	-	-	-	-	-	-	-	-	AUC	0.669 (0.008)	AUC	0.614 (0.003)	0.628 (0.002)
Doğruluk oranı	0.699 (0.003)	0.753 (0.002)	0.735 (0.002)	0.724 (0.005)	0.773 (0.009)	-	-	-	-	-	-	-	-	-	-	Doğruluk oranı	0.775 (0.007)	Doğruluk oranı	0.736 (0.003)	0.735 (0.001)
F1 skoru	0.364 (0.016)	0.576 (0.005)	0.568 (0.004)	0.535 (0.014)	0.546 (0.041)	-	-	-	-	-	-	-	-	-	-	F1 skoru	0.515 (0.013)	F1 skoru	0.403 (0.007)	0.440 (0.004)
MCC skoru	0.232 (0.010)	0.407 (0.006)	0.384 (0.005)	0.353 (0.013)	0.416 (0.033)	-	-	-	-	-	-	-	-	-	-	MCC skoru	0.402 (0.018)	MCC skoru	0.293 (0.007)	0.300 (0.004)
RF																**CANetwork**		**SVM**		
AUC	0.610 (0.017)	0.636 (0.009)	0.611 (0.007)	0.631 (0.013)	0.669 (0.010)	0.585 (0.001)	0.603 (0.002)	0.603 (0.002)	0.605 (0.003)	0.636 (0.010)	0.554 (0.001)	0.612 (0.002)	0.616 (0.002)	0.549 (0.002)	0.639 (0.011)	AUC	0.694 (0.012)	AUC	0.616 (0.003)	0.627 (0.002)
Doğruluk oranı	0.717 (0.012)	0.746 (0.008)	0.717 (0.009)	0.737 (0.027)	0.782 (0.003)	0.722 (0.002)	0.734 (0.002)	0.729 (0.001)	0.732 (0.003)	0.771 (0.005)	0.716 (0.002)	0.739 (0.002)	0.730 (0.001)	0.711 (0.002)	0.769 (0.006)	Doğruluk oranı	0.761 (0.008)	Doğruluk oranı	0.676 (0.003)	0.722 (0.002)
F1 skoru	0.392 (0.043)	0.452 (0.017)	0.413 (0.018)	0.444 (0.025)	0.514 (0.020)	0.334 (0.003)	0.373 (0.004)	0.377 (0.004)	0.380 (0.007)	0.431 (0.026)	0.226 (0.004)	0.395 (0.004)	0.414 (0.004)	0.210 (0.005)	0.440 (0.027)	F1 skoru	0.558 (0.019)	F1 skoru	0.456 (0.004)	0.450 (0.003)
MCC skoru	0.290 (0.031)	0.330 (0.017)	0.289 (0.017)	0.354 (0.033)	0.415 (0.012)	0.241 (0.003)	0.277 (0.004)	0.271 (0.003)	0.278 (0.007)	0.366 (0.020)	0.189 (0.004)	0.296 (0.004)	0.285 (0.004)	0.166 (0.004)	0.361 (0.022)	MCC skoru	0.399 (0.023)	MCC skoru	0.233 (0.005)	0.285 (0.004)
SVM																				
AUC	0.637 (0.002)	0.595 (0.002)	0.647 (0.002)	0.660 (0.006)	0.665 (0.008)	0.600 (0.002)	0.669 (0.002)	0.593 (0.002)	0.581 (0.003)	0.669 (0.012)	0.580 (0.002)	0.628 (0.001)	0.633 (0.002)	0.582 (0.001)	0.661 (0.012)					
Doğruluk oranı	0.737 (0.001)	0.665 (0.001)	0.733 (0.001)	0.747 (0.003)	0.778 (0.005)	0.714 (0.002)	0.763 (0.002)	0.654 (0.002)	0.684 (0.003)	0.786 (0.007)	0.722 (0.001)	0.742 (0.001)	0.734 (0.001)	0.712 (0.001)	0.779 (0.006)					
F1 skoru	0.459 (0.003)	0.423 (0.002)	0.485 (0.003)	0.500 (0.002)	0.502 (0.017)	0.397 (0.003)	0.514 (0.003)	0.429 (0.003)	0.378 (0.006)	0.504 (0.023)	0.321 (0.004)	0.434 (0.003)	0.455 (0.003)	0.337 (0.003)	0.486 (0.026)					
MCC skoru	0.315 (0.004)	0.192 (0.003)	0.326 (0.004)	0.355 (0.009)	0.400 (0.017)	0.246 (0.004)	0.392 (0.003)	0.186 (0.003)	0.182 (0.007)	0.419 (0.023)	0.230 (0.004)	0.318 (0.003)	0.308 (0.004)	0.215 (0.004)	0.397 (0.024)					

EK-4 Turnitin Raporu

ÇOKLU OMİK VERİLERİNİN BİRLEŞTİRİLMESİNDE KULLANILAN YAKLAŞIMLARIN VE SINIFLANDIRMA YÖNTEMLERİNİN PERFORMANSININ ARAŞTIRILMASI

ORIJINALLIK RAPORU

%4	%4	%1	%
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	uzalcbs.org İnternet Kaynağı	%2
2	acikerisim.selcuk.edu.tr:8080 İnternet Kaynağı	<%1
3	www.yumpu.com İnternet Kaynağı	<%1
4	jmlr.csail.mit.edu İnternet Kaynağı	<%1
5	BAŞARAN, Esin, ARAS, Sümer and CANSARAN-DUMAN, Demet. "Genomik, proteomik, metabolomik kavramlarına genel bakış ve uygulama alanları", Refik Saydam Hıfzıssıhha Merkezi Başkanlığı, 2010. Yayın	<%1
6	www.researchgate.net İnternet Kaynağı	<%1

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı: Funda İPEKTEN
Uyruğu: Türkiye (TC)
Doğum Tarihi ve Yeri: 15 Aralık 1991, Malatya
Medeni Durumu: Bekâr
Tel : +90 539 926 28 44
email: fundaipekten@gmail.com

EĞİTİM BİLGİLERİ

Derece	Kurum	Mezuniyet Tarihi
Yüksek Lisans	EÜ Sağlık Bilimler Enstitüsü	2020
Lisans	FÜ Fen Fakültesi, Elazığ	2015

ARAŞTIRMA ALANLARI

Makine Öğrenmesi, Omik Veri Türleri, Veri Madenciliği, Derin Öğrenme, Yapay Zekâ, Omik Veri Entegrasyon Türleri

YETKİNLİKLER

Programlar ve Yazılımlar: R, Python, Weka, SQL, IBM SPSS, Minitab, Stata, SigmaPlot, MedCalc, NCSS, PASS, G*Power

Yabancı Dil: İngilizce

YAYINLANAN MAKALELER

1. Semerci E., Durukan P., Yıldırım S., Baykan N., Yakar Ş., İpekten F. "Gastrointestinal Sistem Kanamalı Hastalarda Şok İndeksi ve Hematokrit Düzeylerinin Mortalite Üzerine Etkisi", Akademik Gastroenteroloji Dergisi Sayı 2, ss 85-89, (2018)

2. Günal S.Y., Saçmacı H., Saçmacı Ş., Mirza M., İpekten F. " Effect Of Heavy Metals And Sialic Acid İn Multiple Sclerosis", International Research Journal Of Public And Environmental Health, 7, pp 114-120, (2018)

3. Yel S., Dursun S., Pınarbaşı A.S., Günay N., Özdemir S., Şahin N., Akgün H., İpekten F., Poyrazoğlu H., Düşünsel R., " Patient Outcomes of Henoch–Schönlein Purpura Nephritis According to the New Semiquantitative Classification", Fetal and Pediatric Pathology, ISSN: 1551-3815, DOI: 10.1080/15513815.2019.1658245, (2019)
4. Yel S., Pınarbaşı A.S., Sağıroğlu P., Atalay M.A., İpekten F., Dursun İ. "Anlamlı Bakteriüride Antibiyotik Direnç Durumu: Çocuk Nefroloji Merkezinde Son Durum Nedir?", Klimik Dergisi, Sayı 1, ss 77-81, (2020)

SÖZLÜ BİLDİRİLER

1. İpekten F., Zararsız G., Ünlüsavuran M., Ertürk Zararsız G., Korkmaz S., Göksülük D., et al., "Metabolomik Biyobelirteçlerinin Tespitinde ANOVA-PCA Yaklaşımı", XIX. ve II. Uluslararası Ulusal Biyoistatistik Kongresi, ANTALYA, TÜRKİYE, 25-28 Ekim 2017, pp.-78

2. Ünlüsavuran M., Zararsız G., İpekten F., Ertürk Zararsız G., Korkmaz S., Göksülük D., et al., "Metabolomik Verilerin Sınıflandırılmasında Kısmi En Küçük Kareler Ayırma Analizi Yaklaşımı", XIX. ve II. Uluslararası Ulusal Biyoistatistik Kongresi, ANTALYA, TÜRKİYE, 25-28 Ekim 2017, pp.-77

3. Ünlüsavuran M., Ertürk Zararsız G., Zararsız G., İpekten F., Korkmaz S., Göksülük D., et al., "H2O Otomatik Makine Öğrenme Algoritmasının Metabolomik Verilerinde Performansının Araştırılması", XX. ve III. Uluslararası Ulusal Biyoistatistik Kongresi, GAZİANTEP, TÜRKİYE, 26-29 Ekim 2018, S36

4. Öztürk A., İpekten F., Çiçek B., Zararsız G., Ertürk Zararsız G., Ünlüsavuran M., et al. "Factors Affecting Susceptibility Of Unıversity Students For Depressive Disorders", IV Jubilee Congress Of General Medicine, Plovdiv, Bulgaria 22-24 November 2018, pp.-10

5. Öztürk A., Çavuşoğlu M., Borlu A., Çiçek B., Ünalan D., Zararsız G., et al. "Pubis To Sole Growth Reference Charts For Turkish Children Aged 0-84 Months", IV Jubılee Congress Of General Medıcıne, Plovdiv, Bulgaria 22-24 November 2018, pp.-19

6. Cankurtaran F., Soyuer F., Gültekin M., Menevşe Ö., İpekten F., "Parkinson Hastalarinda Hastalik Evresine Bağli Donma, Non-Motor Semptomlar Ve Yaşam Kalitesinin İncelenmesi" II.Uluslararası Erciyes Bilimsel Araştırmalar Kongresi, 27-29 Eylül 2019 KAYSERİ, ss.-153-155

PROJELER

BAP PROJESİ, TSA-2019, Müzisyenlerde Müziği Kafadan Çalarken Ve Notadan Okurken Beyinde Olan Değişikliklerin Fonksiyonel Mr Görüntüleri İle Analiz Edilmesi

KATILDIĞI KURSLAR

2016

1. IBASPM Ve Mricloud Ile Beyin Parselasyonu Oluşturarak Beyin İçindeki Yapıların Hacimlerinin Hesaplanması Ve Difüzyon Tensör Görüntüleri Ile Traktografi Yapılarak Beyin İçindeki Liflerin Oluşturulması Ve Üç Boyutlu Gösterilmesi (Erciyes Üniversitesi/Kayseri)

2017

1. Uygulamalı İstatistiksel Veri Analizi (Erciyes Üniversitesi/Kayseri)
2. R ile Kestirime Yönelik İstatistiksel Modellerin Oluşturulması (Biyoistatistik Kongresi/Antalya)

2018

3. Nitel Araştırma Eğitimi (Erciyes Üniversitesi/Kayseri)
4. Count Data Modelling And Analysis With Applications In Medicine (Ege Üniversitesi/İzmir)
5. SPM12 ile fMRI Data Analizi ve Voksel Based Morformetri (Erciyes Üniversitesi/Kayseri)
6. R İle Veri Analizine Giriş
7. Deep Learning ile Biyolojik Veri Analizi Nasıl Yapılır? (Biyoistatistik Kongresi/Gaziantep)
8. Proje Sunum Çalıştayı (Erciyes Üniversitesi/Kayseri)
9. Yapay Zeka ve Bulut Bilişim Konferansı (Abdullah Gül Üniversitesi/Kayseri)

2019

1. Evrimsel Genombilim Uygulamalı Eğitim (Ege Üniversitesi/İzmir)
2. Python İle Derin Öğrenme Eğitimi (Erciyes Üniversitesi/Kayseri)
3. SQL İle Veritabanı Yönetimi ve Sorguları Eğitimi (Erciyes Üniversitesi/Kayseri)
4. EBSCOhost Databases and Services Training (Erciyes Üniversitesi/Kayseri)