TC.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

KİŞİSELLEŞTİRİLMİŞ SAĞKALIM TAHMİNİ İÇİN GENİŞ ÇAPLI KANSER VERİSİNİN YAPAY ÖĞRENME VE ÇOKLU-OMİK BAZLI ANALİZİ

Ayşe Nur ÇORUH

Biyoinformatik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2022

TC.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

KİŞİSELLEŞTİRİLMİŞ SAĞKALIM TAHMİNİ İÇİN GENİŞ ÇAPLI KANSER VERİSİNİN YAPAY ÖĞRENME VE ÇOKLU-OMİK BAZLI ANALİZİ

Ayşe Nur ÇORUH

Biyoinformatik Programı
YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI
Doç. Dr. Tunca DOĞAN

ANKARA
2022

HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
KİŞİSELLEŞTİRİLMİŞ SAĞKALIM TAHMİNİ İÇİN GENİŞ ÇAPLI KANSER VERİSİNİN YAPAY ÖĞRENME VE ÇOKLU-OMİK BAZLI ANALİZİ

Öğrenci: Ayşe Nur ÇORUH

Danışman: Doç. Dr. Tunca DOĞAN

Bu tez çalışması 07.09.2022 tarihinde jürimiz tarafından Biyoinformatik Programında yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	*Doç. Dr. Yeşim AYDIN SON* *(ODTÜ)*	*imza*
Tez Danışmanı:	Doç. Dr. Tunca DOĞAN *(Hacettepe Üniversitesi)*	*imza*
Üye:	*Doç. Dr. Ceren SUCULARLI* *(Hacettepe Üniversitesi)*	*imza*

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

04 Ekim 2022

Prof. Dr. Müge YEMİŞÇİ ÖZKAN
Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

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

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

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

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

03 /10/2022

Ayşe Nur ÇORUH

i

---

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

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

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

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

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

ETİK BEYAN

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

Ayşe Nur ÇORUH

TEŞEKKÜR

Yüksek lisans eğitimime başladığım günden beri, bilgi ve tecrübelerini aktararak biyoinformatik alanında gelişimime katkıda bulunan, öğrencilerine paylaşımcı, özgür ve gelişmelerini teşvik eden bir çalışma ortamı sağlayan, çalışma konumun belirlenmesinde ve çalışma sürecinde eleştirileri ile yol gösteren, yürüttüğü önemli projelerde çalışma imkânı veren, tez çalışmam boyunca desteğini esirgemeyen değerli danışman hocam Doç. Dr. Tunca DOĞAN'a,

Tezimi inceleyip yorumlarını paylaşan değerli hocalarım Doç. Dr. Ceren SUCULARLI ve Doç. Dr. Yeşim AYDIN SON'a,

Çalışma boyunca gerek derslerde gerek sorularım olduğunda, bilgi birikimlerinden yararlandığım ve bu alanda ufkumu açarak gelişmeme katkı sağlamış Biyoinformatik Anabilim Dalının tüm değerli öğretim üyelerine,

Ben kendime inanmadığım zamanlar bile bana inan, her zaman benimle olan anneme,

Her zaman bana destek veren kardeşlerime ve aileme,

Çok teşekkür ederim.

Bu tez çalışması, "TÜBİTAK - BİDEB 2210-C - Yurt İçi Öncelikli Alanlar Yüksek Lisans Burs Programı" ve "TÜBİTAK - ARDEB 3501 - Kariyer Geliştirme Programı" tarafından desteklenmiştir.

ÖZET

ÇORUH A., Kişiselleştirilmiş Sağkalım Tahmini İçin Geniş Çaplı Kanser Verisinin Yapay Öğrenme ve Çoklu-Omik Bazlı Analizi, Hacettepe Üniversitesi; Sağlık Bilimleri Enstitüsü, Biyoinformatik Programı Yüksek Lisans Tezi, ANKARA, 2022. Kanser, dünyada en önde gelen sağlık sorunlarından bir tanesidir. Özellikle bazı kanser alt türlerinin öldürücülüğünün yüksek olması, doğru teşhis, eksiksiz takip ve etkili tedavinin önemini artırmaktadır. Kanserde hayatta kalma, hastaların teşhisinden veya belirli bir tedavinin uygulanmasından sonra hayatta kaldıkları süre olarak tanımlanabilir. Biyotıp alanında kritik öneme sahip bir konu olan hayatta kalma tahmini, ilgili göstergeler ve geçmiş hasta verisi kullanılarak gerçekleştirilmektedir. Yakın zamana kadar, araştırmacılar sağkalımı modellemek için hastaların klinik ve demografik verisi kullanmışlardır. Bu yaklaşım çerçevesinde hem bir tedaviye verilen yanıtın hem de genel olarak hastalığın ilerlemesini etkileyen hastaya özgü moleküler özelliklerin göz ardı edilmesinden dolayı, genellikle düşük bir sağkalım tahmini başarısı elde edilmektedir. Bu çalışmada, kanser hastalarının sağkalımını yüksek başarımla tahmin etmek için yeni bir hesaplama yöntemi önerdik. Bu amaçla, "Genomic Data Commons" (GDC) veri kaynağından elde edilen seçili 13 farklı kanser türünden herhangi biri için teşhis edilen hastaların çoklu omik verisi kullanılmıştır. Girdi omik veri tipleri olarak mutasyon, kopya sayısı varyasyonu (CNV), gen ifadesi ve miRNA ifadeleri seçilmiştir. Ayrıca hastaların klinik verisinin ve uygulanan ilaç bilgilerini girdi özniteliklerine dahil edilmiştir. Rastgele orman algoritmasını kullanarak 13 farklı doku/kanser tipi için spesifik ikili sınıflandırma modelleri eğitilmiştir. Sonuçlarımıza göre, birden çok türde omik veri kullanan modeller, tek omik veri tipi kullanan modellere kıyasla daha iyi tahmin performansı elde edilmiştir. Farklı tipteki omik veri tipleri arasında, dokuların çoğunda mutasyon ve gen ifade özellikleri en yüksek tahmin performansını sağlamıştır. Bu çalışma, kanser hastalarının sağkalım sürelerinin dokuya özgü tahmini için farklı moleküler veri tiplerinin ayrıntılı bir araştırması olarak literatüre katkıda bulunmaktadır.

Anahtar kelimeler: Sağkalım tahmini, makine öğrenmesi, kanser araştırmaları, çoklu omik tabanlı analizler.

ABSTRACT

ÇORUH A., Artificial Learning and Multi-Omics Based Analysis Of Large-Scale Cancer Data For Personalized Survival Predictions, Hacettepe University, Graduate School Health Sciences, Bioinformatics Program Master Degree Thesis, ANKARA, 2022. Cancer is one of the leading causes of death worldwide. The high lethality of some of the sub-types of cancer increases the importance of correct diagnosis, complete follow-up and effective treatment. Survivability in cancer can be defined as the length of time that patients live after the diagnosis and/or the administration of a certain treatment. The estimation of survival, which is a critical topic in biomedicine, is possible using relevant indicators and historical patient data. Until lately, researchers mainly used clinical and demographic data of patients to model survivability, which generally resulted in low success, due to ignoring patient-specific molecular properties that affect both the response given to a treatment and the progression of the disease in general. In this study, we proposed a new computational method to predict the survival of cancer patients. For this purpose, we utilized multi-omics data of patients diagnosed with 1 of the 13 different types of cancer, which are obtained from Genomic Data Commons (GDC) data portal. We used mutation, copy number variation (CNV), gene expression, and miRNA expression as our input omic data types. In addition, we incorporated the clinical data and administered drug information of the patients, to our input features. We utilized the random forest algorithm and trained 13 tissue/cancer specific binary classification models. According to our results, models that use multiple types of omic data achieved better prediction performance, compared to the models using a single-omic. Among different types of omics data, mutation and gene expression features provided the highest prediction performance, in the majority of the tissues. This study contributes to the literature as a detailed investigation of different molecular data types for tissue specific prediction of cancer patient survival.

Key words: Survival prediction, machine learning, cancer research, multi-omics-based analysis.

İÇİNDEKİLER

SİMGELER ve KISALTMALAR

AJCC	American Joint Committee on Cancer
BRCA	Breast İnvasive Carcinoma
CCG	Center for Cancer Genomics
CESC	Cervical squamous cell carcinoma
CNN	Convolutional Neural Networks
CNV	Copy Number Variation
COAD	Colon Adenocarcinoma
COT	Cox Orantılı Tehlike
DNA	Deoksiribonükleik Asit
DVM	Destek Vektör Makinesi
FPKM	Fragments Per Kilobase Million
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
HGP	Human Genome Project
KIRP	Kidney renal papillary cell carcinoma
KSV	Kopya Sayısı Varyasyonu
LAML	Acute Myeloid Leukemia
LIHC	Liver hepatocellular carcinoma
LOOCV	Leave One Out Cross Validation
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MAF	Mutation Annotation Format
NIH	National Institutes of Health
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
RPM	Per Million Mapped Reads
SKCM	Skin Cutaneous Melanoma
SNP	Single Nucleotide Polymorphism
STAD	Stomach adenocarcinoma
TARGET	The Therapeutically Applicable Research to Generate Effective Treatments

TCGA	The Cancer Genome Atlas
UCEC	Uterine Corpus Endometrial Carcinoma
UICC	International Union Against Cancer
YSA	Yapay Sinir Ağları

ŞEKİLLER

TABLOLAR

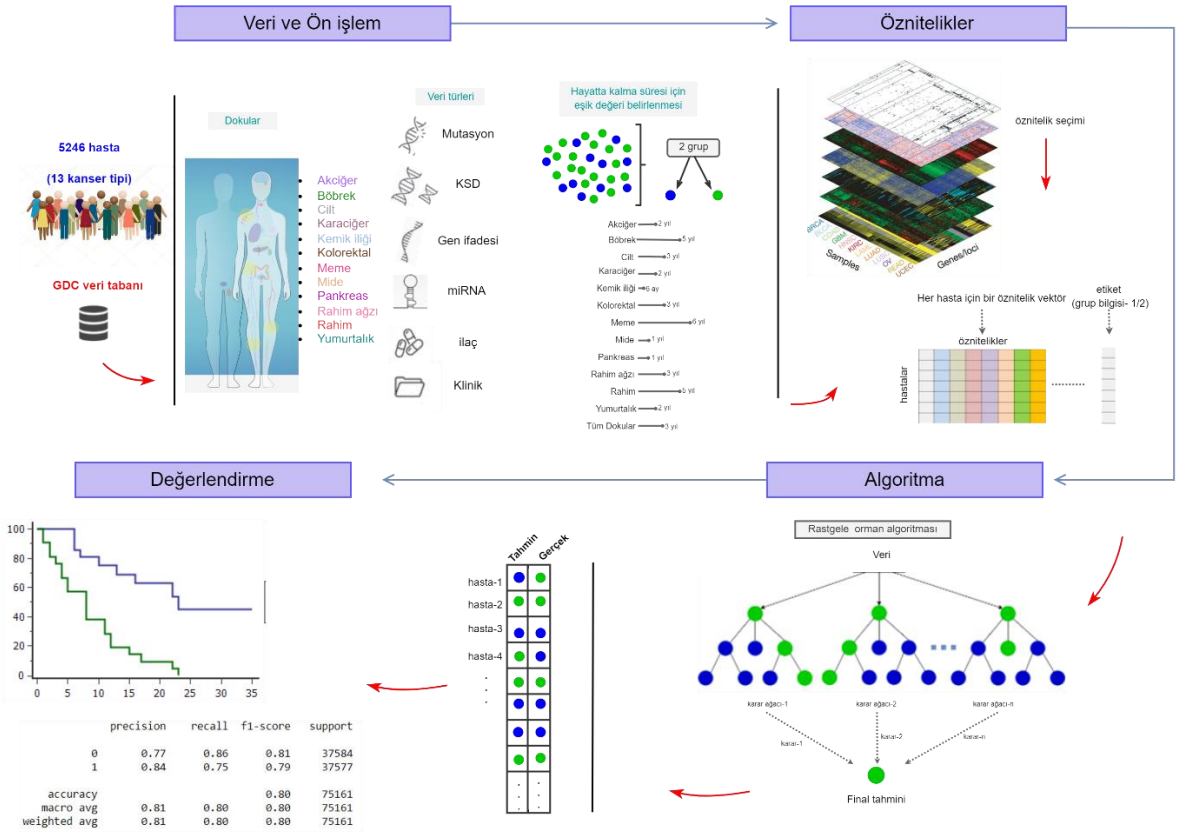
1.GİRİŞ

Kanserde sağkalım hastalar için kritik öneme sahiptir. Sağkalım hastaya uygulanacak tedavi açısından yol gösterici ve hastalığın muhtemel seyrini tahmin etmede önemlidir (1). Kanserde bireysel sağkalım tahmini önemli bir yol gösterici olması yanında kişiselleştirilmiş tıp açısından bireye özel bir yaklaşım sunmaktadır. Sağkalım modellemek için genellikle hastaların klinik ve demografik bilgilerinden yararlanılmıştır. Gelişen teknoloji ile farklı türde omik veri üretimi artmıştır ve hastaya özgü moleküler bilgilere erişmek daha kolay olmuştur. Karmaşık ve büyük yapıların önemli özelliklerini yakalaması açısından makine öğrenmesi yöntemi, omik verinin analizi için verimli olmuştur. Omik veriyi kullanarak kanserde sağkalım tahmini yapan birçok çalışma yapılmıştır. Literatürde kanserde sağkalım tahmini için tek omik veri tipi kullanan çalışmalar mevcuttur. Ancak çoklu omik veri tiplerinin kullanılması bu konuda yeterince çalışılmamıştır. Hastaların genetik yapısındaki farklılıklardan kaynaklı olarak ilaçlara verecekleri yanıt ve buna bağlı olarak sağkalım süreleri değişmektedir. Bu nedenle bunu başarılı bir şekilde modelleyen hesaplamalı yöntemlere ihtiyaç vardır.

Tez çalışmamızda mevcut ihtiyaçlar göz önüne alınarak modelleme sırasında kullanılan girdi verisinin çeşitlendirilmesiyle daha başarılı sağkalım tahmin modelleri olusturmak amaçlanmıştır.

Bu amaç doğrultusunda tez çalışmasında; kamuya açık en büyük kanser veri tabanı olan GDC veri kaynağından elde edilen 13 farklı kanser türü için hastaların çoklu omik verisi kullanılmıştır. Omik veri tipi olarak mutasyon, kopya sayısı varyasyonu (CNV), gen ifadesi ve miRNA ifadelerini kullanılmıştır. Omik veriye ek olarak, hastaların klinik ve tedavi için uygulanan ilaç bilgilerini de çalışmaya dahil edilmiştir. Sağkalım tahmini için rastgele orman algoritması kullanılarak her kanser türüne özgü ikili sınıflandırma modelleri eğitilmiştir. Çalışmada kullanılan veri türleriyle farklı veri kombinasyonları üzerinden modeller eğitilmiş ve performans sonuçları araştırılmıştır. Tüm kanser türlerinin bir arada değerlendirildiği yaklaşım ile

doku fark etmeksizin tüm kanserlerin birlikte gösterdikleri performansı da araştırılmıştır. Tüm çalışmanın iş akışı Şekil 1.1.'de gösterilmektedir.



Şekil 1.1. Model iş akışı.

Genel bilgiler bölümünde çalışma ile ilgili temel kavramlar ve kanserde sağkalımla ilgili literatür araştırması yapılmıştır. Çalışmada kullanılan verinin düzenlenmesi, modelleme ve performans metrikleri hakkında bilgilendirme gereç ve yöntemler bölümünde yapılmıştır. Eğitilen tüm modellerin performans sonuçları bulgular bölümünde verilmiştir. Tartışma bölümünde bulgularda verilen sonuçlara dayandırılarak modellerin performansları değerlendirilmiştir. Sonuçlar bölümünde tez çalışması için amaçlanan sonuçlar göz önüne alınarak elde edilen sonuçlar hakkında değerlendirme yapılmıştır.

2. GENEL BİLGİLER

2.1. Kanser

Kanser, metastaz yapma özelliğine sahip kontrolsüz hücre bölünmesi olarak tanımlanmaktadır (2). Kanserin 100'den fazla türü bulunmaktadır. Vücudun herhangi bir doku veya organında gelişebilir (3). Normalde hücreler, vücudun ihtiyaç duyduğu yeni hücreler oluşturmak için büyür ve çoğalır. Hücreler yaşlandıklarında veya hasar gördüklerinde mevcut hücreler ölür ve yerlerini yeni hücreler alır. Bazen bu düzenli süreç bozulur ve anormal veya hasarlı hücreler ölmeleri gerekirken çoğalır. Bu çoğalan kontrolsüz hücreler, doku birikmesiyle tümörleri oluşturabilir. Kanserli tümörler yakındaki dokulara yayılır veya onları istila ederler. Metastaz adı verilen bir süreçle vücuttaki farklı bölgelere yayılarak yeni tümörler oluşabilir. Kanserli tümörler malign tümörler olarak da adlandırılmaktadır.

Kanser, genetik bir hastalık olması yanında tek bir sebebe bağlı olmayan çok faktörlü bir hastalık olarak tanımlanmaktadır (4). Son yıllarda, değişen yaşam şartları ve artan yaşamsal beklentilerden kaynaklı olarak kanserin görülme sıklığı giderek artmaktadır (3). 21. yüzyılda sürekli artan insidansla giderek yaygınlaştığı bilinen kanser, en korkulan hastalıklardan biri olarak görülmektedir (3). Kanser hücrelerini diğer sağlıklı hücrelerden ayıran belirlenmiş altı kanser belirteci (*The Hallmarks of Cancer*) mevcuttur. Bu belirteçler: bölünme sinyallerinde kendi kendine yeterlilik, bölünmeyi durdurucu sinyallere karşı cevapsızlık, programlanmış hücre ölümünün önlenmesi, sınırsız çoğalma, sürekli anjiyogenez , dokuda yayılma ve metastazdır (5). Hücrenin bu değişimleri antikanser mekanizmasını başarılı bir şekilde engeller ve kanser gelişimine neden olur (5)

Dünya genelinde başlıca ölüm nedenleri arasında 2. sırada olan kanser her yaş grubu ve cinsiyette görülmektedir (6). 2020 GLOBOCAN verisine göre, dünya çapında tahmini 19.3 milyon kanser yeni vakası ve 10 milyon kanserden kaynaklı ölüm olduğu bildirilmiştir (7). Dünya çapında tahmini vakalar ve ölümler için ilk on kanser türü: meme, akciğer, prostat, cilt, kolon, mide, karaciğer, rektum, rahim ağzı ve özofagus kanseridir (7). Erkeklerde en yaygın kanser türü prostat, en çok ölümle

sonuçlanan ise akciğer kanseridir. Kadınlarda ise hem en yaygın hem de en çok ölümle sonuçlanan kanser türü meme kanseridir (7).

Genel olarak insanda kanser gelişimi, normal hücrelerin genetik değişiklikler sonucu malign hücrelere dönüşümüne ilerleyen, çok aşamalı bir süreci olarak kabul edilir (5, 8). Ancak son zamanlardaki çalışmalar, insan kanser hücrelerinin bu genetik değişikliğe ek olarak, belirgin epigenetik anormalliklere de sahip olduğunu göstermektedir (9). Epigenetik, DNA dizisindeki herhangi bir değişikliğe bağlı olmayan, DNA ve kromatin üzerinde modifikasyonlarla ilişkili gen ifadesindeki kalıtsal değişiklikler olarak tanımlanmaktadır (10). Kanser, hem genetik hem de epigenetik değişikliklerle ilişkili bir hastalık olarak kabul edilir ve her iki faktör de kanserin gelişmesi ve ilerlemesini desteklemek için birlikte çalışırlar (10).

2.1.1. Kanserde Tanı ve Tedavi Yaklaşımları

Sık görülmesi ve öldürücülüğünün yüksek olması kanserin tanısının, takibinin, tedavisinin önemini artırmaktadır. Kanser tanısı için hekimlerin kullandıkları birçok yöntem bulunmaktadır. Bir veya birden çok yaklaşım ile tanı konabilmektedir. Bu yaklaşımlar fiziki muayene, laboratuvar testleri, görüntüleme testleri ve biyopsi gibi yöntemlerdir.

Kanserde tümörün büyüklüğü ve diğer dokulara metastaz yapma durumu gibi kanserin kapsamını ifade eden kavrama evre denir. Kanserin evresini bilmek ne kadar ilerlediğini gösterir. Hastanın hayatta kalma şansını anlamaya ve hastalık için en uygun tedavi planını oluşturmaya yardımcı olur. Bunlara ek olarak hekim, hastalığın evresini öğrenmek için röntgen, laboratuvar testleri ve benzeri tanı tetkiklerini isteyebilmektedir.

Kanserde birçok evreleme sistemi mevcuttur. Genel olarak sistemler, tümörün vücutta bulunduğu yer, hücre tipi, tümör boyutu, lenf düğümlerine yayılma durumu, kanserin vücudun başka bölgesine yayılma durumu ve tümör derecesi bilgisini içerir. Belirli bir kanser türüne özgü evreleme sitemleri bulunmaktadır. Tümörün özellikleri (T), lenf nodu (N) ve metastaz (M) durumlarının değerlendirildiği TNM evreleme sistemi birçok kanser türü için ortak kullanılan bir evreleme sistemidir (11). AJCC

evreleme sistemi olarak da adlandırılmaktadır. Bu sistem Amerikan Ortak Kanser Komitesi (American Joint Committee on Cancer, AJCC) ve Uluslararası Kansere Karşı Birlik (International Union Against Cancer, UICC) tarafından oluşturulmuş ve belli aralıklar ile güncellenmektedir. Kanser vakalarını evrelere göre gruplara ayırma uygulaması, hastalığın lokalize olduğu vakalarda sağkalım oranlarının, hastalığın başka bölgeye yayıldığı vakalara göre daha yüksek olmasından kaynaklanmıştır (11). Hastalığın anatomik yaygınlığını tanımlayan TNM sistemi, üç bileşenin değerlendirilmesine dayanır:

T – Primer tümörün boyutu

N – Bölgesel lenf nodu metastazının yokluğu veya varlığı ve yaygınlığı

M – Uzak metastaz yokluğu veya varlığı

Bu üç bileşene sayıların eklenmesi, malign hastalığın yaygınlığını gösterir. Bunlar:

T0, T1, T2, T3, T4 N0, N1, N2, N3 M0, M1'dir (11).

Kanserde yaygın olarak kullanılan 3 tedavi yöntemi kullanılır. Bunlar cerrahi, radyoterapi ve kemoterapidir (12). Bu tedavi yöntemleri, konvansiyonel tedavi yöntemleri olarak bilinir ve bilimsel olarak test edilmiş, etkili ve güvenli bulunmuştur (12). Bunlar dışında hormon tedavisi, biyolojik tedavi yöntemi ve hedefe yönelik tedaviler yöntemlerine de başvurulabilir. Gerekli tedavi, hastanın hastalık durumuna göre tek bir yöntem veya birden çok yöntemin birlikte uygulanmasını şeklinde olabilir (12).

Kanserli tümör ve tümörlü dokunun vücuttan çıkarılmasına cerrahi tedavi denmektedir. Birçok kanser türünde ilk tedavi yöntemidir ve tedavi sonunda hastalıkta iyileşme sağlanabilir. Cerrahi tedavi yönteminde tanı doğrulamak amacı ile biyopsi işlemi yapılır. Biyopsi, kanserde evreleme ve ağrıyı azaltmaya yönelik bir tedavi için uygulanmaktadır (13).

Yüksek ve özel bir enerji türü olan radyasyonu, özel cihazlar yardımıyla hastalıklı organa yönlendirerek, yüksek dozda tedavi amacıyla kullanılmasına radyoterapi denilmektedir. Hastalığının erken evresindeki hastalarda ya da ameliyat

sonrası kanser nüksünü önlemek için kullanılır. Amacı kanserli hücreleri yok etmek ya da kanserli tümörün küçülmesini sağlamaktır (14).

Kemoterapi, kanser hücrelerini etkileyen ve ölmelerini sağlayan kanser ilaçları kullanılarak yapılan bir tedavidir. Kemoterapide kullanılan ilaçlar kanser hücrelerinin bölünmesini ve çoğalmasını önleyerek bu ölümsüz hücreleri hasara uğratır. Kemoterapi tek başına uygulandığı gibi ilaç kombinasyonlarıyla uygulanabilir. Bu kombinasyonlar tek ilaç kullanımına kıyasla daha yüksek etki gösterebilmektedir. Kemoterapi tedavisinde ilaç, hastalıklı hücreleri etkilediği gibi sağlıklı hücre ve dokuları da olumsuz etkilemektedir (15).

Genetik bilgi kanser için oldukça önemlidir. Kanserli ve normal dokular arasında genetik değişikliklerden kaynaklı farklar bulunmaktadır. Genetik bilgi eldesi için kullanılan DNA dizileme konusunda ilk önemli gelişme, tamamı 2003 yılında tamamlanan İnsan Genom Projesi (*Human Genome Project, HGP*) ile olmuştur. İnsan Genom Projesi, insan DNA'sını oluşturan baz çiftlerini belirlemek, insan genomunun tüm genlerini tanımlamak ve gen haritasını çıkarmak amacı güden uluslararası bir bilimsel araştırma projesidir (16). Bu proje, farklı ülke ve kurumların iş birliğiyle gerçekleştirilmiştir. Çalışma, yeni teknolojik gelişmelerin yardımıyla planlanandan erken tamamlanmıştır. İnsan genomunun ilk dizisinin yayınlanması, modern biyolojik araştırmaların en önemli noktalarından biri olarak kabul edilmektedir (16). İnsan genom projesi sonrası elde edilen tecrübe ve bilgiler ile yeni DNA dizileme yöntemleri geliştirilmiştir.

Gelişen teknolojinin yardımı ile moleküler biyoloji ve tümör biyolojisi hakkındaki artan bilgi, son yıllarda kanser tedavisi paradigmalarını önemli şekilde değiştirmiştir (17). Geleneksel klinik genetik genellikle bir kişinin aile öyküsü, etnik kökeni veya tıbbi geçmişi temelinde önceden belirlenmiş monogenik bozuklukların (tek gen hastalıkların) belirlenmesini kapsar. Buna karşılık kişiselleştirilmiş tıp yaklaşımı, hastalık riskinin birkaç gen ve aynı zamanda genomik olmayan faktörler arasındaki etkileşimden kaynaklanan çok faktörlü bozukluklarda bireysel risk bilgisi sunmak için genetik profil oluşturma stratejisini içerir (16). Standart tıbbi tedavi, bireylerin genetik değişkenliğinin hesaba katılmadığı ve popülasyon düzeyinde yapılan kohort tabanlı çalışmalar tarafından yönlendirilmiştir (18). Modern

kişiselleştirilmiş tıpta, hasta için bir tedavi planı oluşturulmadan önce bireyin genetik yapısı ve hastalık geçmişini dikkate alınır ve hedefe yönelik tedaviye uygulanır (19). Kanser için kişiselleştirilmiş tıp yaklaşımı, kanserin ve hastanın genetik profillerine dayalı olarak, hastaya doğru zamanda doğru ilacın gerekli dozunu sağlamayı amaçlar (20). Bu yaklaşım kişinin genetik farklılıklarından doğan kendine özel hastalık karakterini belirler. Bu şekilde hastanın sağlıklı doku veya organlarına daha az zarar veren, daha etkili bir tedavi programı oluşturulabilir.

2.1.2. Kanser Çalışmalarında Veri Analizi

Kanserde teşhis amacı ile yapılan laboratuvar testleri, demografik ve klinik bilgiler gibi veri türleri kanser ile ilgili çalışmalarda sıklıkla kullanılmaktadır. Klinik veri: hastalık teşhis bilgisi, hasta takip bilgisi, patoloji bilgisi, tedavi bilgisi, demografik bilgiler, laboratuvar testleri ve aile öyküsü ile ilgili bilgileri kapsar. Demografi, araştırmacıların belirli bir popülasyonun ölçülebilir istatistiklerini inceledikleri alandır. Popülasyondaki alt kümeleri belirlemek için kullanılır. Demografik veri popülasyondaki ırk, cinsiyet, yaş ve benzeri faktörleri içeren çeşitli özelliklerdir.

Sağkalım Analizi

Sağkalım analizi, bir olayın başlangıç noktasından olayın gerçekleştiği noktaya kadarki sürede üretilen verinin analizi ile yaşam süresinin hesaplanmasıdır (21). Sağkalım analizinde ölene kadar geçen süreye genel sağkalım (OS), bir tedaviden sonra hastanın tekrar hastalığa yakalanmasına kadar geçen süreye ise hastalıksız sağkalım denmektedir. Kanserde sağkalım, hastaların teşhis veya belirli bir tedavinin uygulanmasından sonra yaşadıkları süre olarak tanımlanabilir. Kanser çalışmalarında ölüme kadar geçen süre, sağkalım analizinde ilgilenilen olaydır (21). Sağkalım analiziyle ilgili zorluk, olayı sadece bazı bireylerin deneyimlemiş olması ve çalışma grubunun bir alt kümesi için hayatta kalma sürelerinin bilinmemesinden kaynaklanmaktadır. Buna sansür denir ve farklı şekillerde ortaya çıkabilir: hastanın çalışmanın sonuna kadar ölümle ilgili bir sonuca ulaşmaması, hastanın çalışma süresinde takipten çıkması, hastanın takibini imkansız hale getiren farklı bir olay yaşamış olması gibi (21). Bu tür sansürlü sağkalım süreleri, olayın gerçek süresini

önemsemez. Genel olarak üç tip sansürlü veri mevcuttur: sağdan sansürlü veri, soldan sansürlü veri ve aralık sansürlü veridir (22). Bireyin olayı takip süresi içinde ilgilenilen olay gözlenmezse ya da belirli bir süre sonra bireyden bilgi alınamıyorsa bu duruma sağdan sansür denir. Sansürleme, ilgilenilen olayın varlığını gözlemlediğimizde ancak nerede başladığını bilmediğimiz durumlarda meydana gelebilir. Buna da soldan sansür denir. Bireyin gözleme girip çıkması durumunda ise olay zamanı verisi, aralıklı sansürlü olur (21).

Sağkalım verisi genellikle sağkalım ve tehlike olarak iki olasılık açısından tanımlanır ve modellenir (21). Sağkalım olasılığı S(t) bir bireyin zaman başlangıcından (kanser teşhisi konulması gibi) belirli bir zaman olan t'ye kadar hayatta kalma olasılığıdır ve bir sağkalım analizi için esastır. Tehlike ise h(t) veya λ(t) ile gösterilir ve t anında bireyin o anda bir olay yaşama olasılığını gösterir (21).

1958 yılında önerilen Kaplan – Meier yöntemi, belirli bir olayın gerçekleşme oranının ya da olasılığının zamana bağlı olarak değişiminin analizini yapar (23). Hayatta kalma olasılığı hem sansürlü hem de sansürsüz gözlemlenen hayatta kalma sürelerinden parametrik olmayan bir şekilde tahmin edilebilir (23). Her olay için bir olasılık tahmin edilir ve bu olasılıklar bir grafiğe döküldüğünde basamak gibi görünür. Kaplan-Meier yönteminde sağkalım olasılıkları, adımsal biçiminde belirtilir.

İki veya daha fazla hasta grubunun sağkalımını karşılaştırmak için, parametrik olmayan bir test olan log-rank testi kullanılabilir (24). İki veya daha fazla sağkalım eğrisini karşılaştırmak için kullanılan en yaygın yöntemdir (24). Yöntem, her olay zamanında, her bir grup için, bir önceki olaydan bu yana beklenebilecek olay sayısını hesaplar (21). Gözlenen her olay zamanında iki grubun tehlike fonksiyonlarının tahminlerini karşılaştırır.

Sağkalım verisinin analizinde yaygın olarak kullanılan bir diğer yöntem de yaşam tablosu yöntemidir. Ölüm oranını ölçmek ve bir popülasyonun hayatta kalma deneyimini tanımlamak için kullanılan en eski tekniklerden biridir (25). Hastalar için oluşturulan yaşam tablolarına klinik yaşam tabloları denir (25). Yaşam tablosu yöntemi, yaşam sürelerinin araştırmacının belirlediği zaman aralıklarına göre

gruplayarak değerlendirildiği bir yöntemdir (26). Sağkalım eğrilerini bulmak için Kaplan – Meier yöntemi ile kullanılan en yaygın yöntemdir.

Kaplan – Meier yöntemi ve log-rank yöntemi tek değişkenli analiz örnekleridir. Ancak birden çok değişkenli durumlarda kullanılamamaları farklı yöntemlerin geliştirilmesini sağlamıştır (27). 1972 yılında Cox tarafından geliştirilen Cox orantılı tehlike (COT) modeli, tıbbi araştırmalarda, hastaların hayatta kalma süreleri ile bir veya daha fazla değişken arasındaki ilişkiyi araştırmak için istatistiksel olarak yaygın kullanılan bir modeldir (27). Hasta sağkalımını etkileyen birçok durum mevcuttur. COT modeli de bu durumu temel alarak birden çok değişkenli sağkalım analizinde kullanılmaktadır. COT model belirli bir zaman içinde belirli bir olayın (ölüm, hastalık nüksü vb.) meydana gelme riski üzerindeki etkilerini hesaplayan yarı parametrik bir modeldir. Ortak değişkenlerin sağkalım veya diğer sansürlenmiş sonuçlarla ilişkisini modelleme için en çok kullanılan prosedür haline gelmiştir (28). Tehlike fonksiyonu ile ifade edilen olay insidansı ve ortak değişkenler arası ilişkiyi tanımlayan bir sağkalım analizi regresyon modelidir (29). Tehlike, belirli bir zamandaki anlık olay olma olasılığı veya bir kişinin, olayı bir dönemde deneyimleme olasılığı olarak tanımlanabilir (29). Cox orantılı tehlike modeli matematiksel olarak Formül 2.1 ile ifade edilebilir (29).

$$h(t) = h_0(t) \cdot exp\left[b_1x_1 + b_2x_2 + \cdots + b_px_p\right] \qquad (2.1)$$

Tehlike fonksiyonu $h(t)$, etkisi ilgili katsayıların $(b1, b2, \dots, bp)$ boyutuyla ölçülen bir dizi p ortak değişkene $(x1, x2, \dots, xp)$ bağlıdır. "$h0$" terimi temel tehlike olarak adlandırılır ve tüm xi lerin sıfıra eşit olması durumunda tehlikenin değeridir. t sağkalım süresini temsil eder. $h(t)$'deki 't' bize tehlikenin zaman içinde değişebileceğini hatırlatır. COT modelinin önemli bir özelliği, temel tehlike fonksiyonunun parametrik olmayan bir şekilde tahmin edilmesidir. Bu sebeple diğer birçok istatistiksel modelden farklı olarak, sağkalım sürelerinin belirli bir istatistiksel dağılımı takip ettiği varsayılmaz (27).

COT modeli, xi değişkenleri üzerindeki tehlikenin logaritmasının çoklu doğrusal bir regresyonudur. Modelin temel varsayımını, herhangi bir gruptaki olayın

tehlikesi, bir diğerindeki tehlikenin sabit bir katıdır. Bu varsayım, gruplar için tehlike eğrilerinin orantılı olması gerektiğini ifade eder (27).

2.2. Makine Öğrenmesi

Makine öğrenmesi, hesaplamalar ve analizlerle veri üzerinden çıkarım yaparak bilgisayarda modelleme yapılmasıdır. Genel olarak performansı artırmak, doğru tahminler yapmak için deneyimi kullanan hesaplama yöntemleri olarak tanımlanır (30). İnsanların öğrenme şeklini taklit ederek oluşturulmuş sistemlerdir. Belirli bir kullanım amacı için toplanan ve analiz için uygun hale getirilen veri, öğrenen sistem için mevcut olan geçmiş bilgileri ifade eder ve bu veriyi kullanılarak bilgisayar programlanır (30). Makine öğrenmesi tekniklerinin temel amacı, sınıflandırma, tahmin veya benzeri herhangi bir görevi gerçekleştirmek amacıyla kullanılabilecek bir model üretmektir.

Bir makine öğrenmesi algoritmasının başarısı kullanılan veriye bağlıdır. Öğrenmesi, doğasından kaynaklı olarak veri analizi ve istatistiklerle ilişkidir (30). Makine öğrenmesi çok geniş bir pratik uygulama alanı sahiptir. Metin veya belge sınıflandırması, doğal dil işleme, konuşma işleme, bilgisayarla görme, hesaplamalı biyoloji ve diğer birçok sorunu içerir (30). Belirlenen tahmin problemlerinin çoğu, öğrenme problemleri olarak gösterilebilir. Yaygın kullanımı olan makine öğrenmesinin uygulama alanları günümüzde genişlemeye devam etmektedir (30). Makine öğrenmesi yöntemlerinin kullanımının yaygın olmasındaki başlıca nedenlerinden biri, verideki kalıpları otomatik olarak tanımlamasıdır (31). Birden çok heterojen özelliği entegre etmede verimlidirler. Makine öğrenmesinin bir başka gücü, oldukça karmaşık modeller oluşturma yeteneğine sahip olmasıyla karmaşık ilişkiler içeren verinin analizinde başarılı olmalarıdır (31).

Makine öğrenmesi algoritmaları, öğrenme yöntemlerine göre denetimli, denetimsiz ve yarı denetimli olarak 3 ana yönteme ayrılır (32). Denetimli yöntemlerde, model etiketli örnekler üzerinde eğitilir. Daha sonra görülmemiş örneklerle karşılaşınca etiketleri tahmin etmek için öğrenilmiş bilgi kullanılır. Sınıflandırma, regresyon ve sıralama problemleriyle ilgili en yaygın kullanımdır (30). Denetimsiz yöntemlerde, model etiket bilgisine gerek duymadan veri kümelerindeki kalıplar bulur.

Kümeleme, boyutluluk azaltma problemleri için bu yöntem kullanılabilir (30). Yarı denetimli yöntemler, bu iki yaklaşımı birleştirme prensibine dayanır. Etiketlerin tahmininde gücü artırmak için etiketlenmemiş verideki kalıplardan yararlanırlar. Melez bir yöntem olan yarı denetimli öğrenme, etiketlenmemiş veriye kolayca erişilebildiği ancak etiketlerin elde edilmesinin zor olduğu durumlarda tercih edilir. Hem denetimli hem denetimsiz öğrenmedeki problemler için kullanılabilir (30).

Bir makine öğrenmesi yöntemi uygulanırken veri örnekleri temel bileşenleri oluşturur. Her örnek çeşitli özelliklerle tanımlanır ve her özellik farklı değer türlerinden meydana gelmektedir. Veride gürültü, aykırı değerler, eksik ve yinelenen veride dolayı veri kalitesi sorunları ortaya çıkabilir. Veri kalitesi iyileştirildiğinde, tipik olarak elde edilen analizin kalitesi de iyileştirilir (33). Bu yüzden veri ön işleme basamağı analiz için önemli bir aşamadır. Belirli bir makine öğrenmesi yöntemine daha iyi uyması için veriyi değiştirmeye odaklanan veri ön işleme ile ilgili bir dizi farklı teknik veya strateji mevcuttur. Bu teknikler arasında en önemli yaklaşımlar: boyutluluk indirgeme, özellik seçimi ve özellik çıkarımıdır (33). Veri kümelerinin çok sayıda özelliğe sahip olduğu durumlarda, veri boyutunun azaltılması beraberinde birçok fayda sağlar. Makine öğrenmesi algoritmaları, veri boyutu düşük olduğunda daha iyi çalışır. Buna ek olarak, boyutluluğun azaltılması, alakasız özellikleri ortadan kaldırabilir, gürültüyü azaltabilir. Daha az özelliğin dahil edilmesi nedeniyle algoritmanın daha yüksek performanslı öğrenme modelleri oluşturması sağlayabilir (33).

Genel olarak, eldeki verinin bir alt kümesini seçerek boyut azaltma olan öznitelik seçimi için gömülü, filtre ve sarmalayıcı yaklaşımlar olmak üzere üç ana yaklaşım bulunmaktadır. Özellik çıkarma durumunda, bir veri kümesindeki tüm önemli bilgileri yakalayan ilk kümeden yeni bir özellik kümesi oluşturulabilir (33).

2.2.1. Biyolojik veri analizinde kullanılan makine öğrenmesi algoritmaları

Lojistik Regresyon

İkili sınıflandırma için yaygın bir makine öğrenmesi algoritmasıdır. Lojistik regresyon, bir sonucu belirleyen bir yada daha fazla bağımsız değişkenli veriyi analiz etmek için kullanılır (34). Modelinin temeli, 2 seviyeli bir olayın sonuç olasılığına dayanmaktadır (35). Buradaki 2 seviyeli durum "1" ve "0" olarak sınıflandırılan veriyi içerir. "1" olayın olması, "0" ise olmaması anlamına gelmektedir. Olay olasılığı, olayın olma olasılığının, olmama olasılığına oranı olarak tanımlanmaktadır (35). Lojistik regresyon modeli, esasen regresyon fonksiyonu olarak olasılıkların doğal logaritmasını almaktadır (35). Modelleme süreci, Wald testine ve olabilirlik oranı testine dayanmaktadır (36).

K-En Yakın Komşu Algoritması

En bilinen sınıflandırma yöntemlerinden biridir. K-en yakın komşu algoritması, eğitim kümesinde test nesnesine en yakın olan bir grup k nesneyi bulur. Nesne için etiket atamasını belirli bir sınıfın baskınlığına dayandırır (37). Algoritma için üç önemli unsuru bulunmaktadır. Birincisi etiketlenmiş nesneler, ikincisi bu nesneler arasındaki mesafeyi hesaplamak için bir mesafe veya benzerlik ölçüsü ve son olarak en yakın komşuların sayısı olan k değeridir (37). Etiketlenmemiş bir nesneyi sınıflandırmak için, nesnenin etiketlenmiş nesnelere olan mesafesi hesaplanır. Daha sonra en yakın komşuları tanımlanır. En yakın komşuları sınıf etiketi kullanılarak nesnenin etiketi belirlenir (37). K-en yakın komşu algoritması, anlaşılması ve uygulanması kolay bir sınıflandırma tekniğidir.

Destek Vektör Makinesi

Destek Vektör Makinesi (DVM), sınıflandırma problemleri için kullanılan oldukça etkili ve basit bir denetimli makine öğrenmesi algoritmasıdır (38). İki sınıflı bir öğrenme görevinde DVM'nin amacı, eğitim verisinde iki sınıfın üyelerini ayırt etmek için en iyi sınıflandırma fonksiyonunu bulmaktır (37). Sınıf dağılımlarının kenarında yer alan eğitim örneklerine, destek vektörlerine göre sınıflar arasında

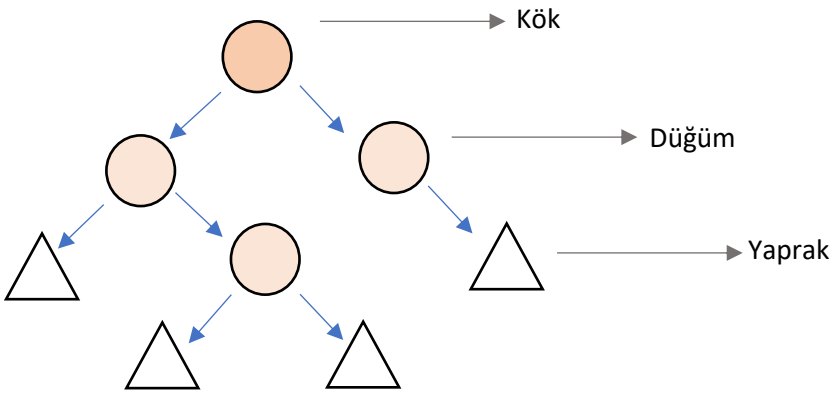
optimal bir ayırıcı hiper düzlem yerleştirme fikrine dayanmaktadır (39). Sınıflandırma için düzlemdeki iki grup arasında bir sınır çizilir ve bu durum, iki grubu ayırmayı sağlar (38). Bu sınırın tam olarak nereden çizileceği ise iki gruba dahil olan üyelerin birbirlerine en uzak oldukları yer olarak belirlenir. Az sayıda örnekleme sahip veride doğru sınıflandırma sonuçları elde etmeyi sağlamaktadır (39). DVM sınıflandırmasına yönelik temel yaklaşım, doğrusal olmayan verinin sınıflandırılmasında kullanılabilir (39).

Yapay Sinir Ağları

Yapay sinir ağı (YSA) temel olarak insan beyninin işlevini taklit eden bir yapıdadır ve bu işlevi taklit ederek problemleri kodlama ve çözme yollarından ilham alan bir sistem şeklinde tasarlanmıştır (40). Bir işlemi yürüttükten sonra diğer gözlemlerden yeni gözlemleri tahmin edebilen son derece karmaşık analitik teknikler olarak bilinen YSA'ları, karmaşık doğrusal olmayan fonksiyonları modelleyebilir. Bir YSA, ağırlıklı bağlantılarla birbirine bağlanan çok sayıda basit işlemciden oluşur (41). Her birim diğer nöron olarak adlandırılan birçok düğümden girdi alır, çıktısını başka bir düğüme iletir ve en sonunda gelen sonuçlar nihai çıktıyı verir.

Karar Ağaçları

Denetimli bir öğrenme olan karar ağaçları, sınıf etiketlerinin bilindiği bir dizi eğitim örneği analiz edilerek oluşturulur. Daha sonra görülmemiş örnekleri öğrenilen bilgilerle sınıflandırmak için kullanarak uygulanır (42). Bir karar ağacı, öğelerle ilişkili özellikler hakkında sorular sorarak veri öğelerini sınıflandırır. Her soru bir düğümde bulunur ve her düğüm, soruya olası her yanıt için bir alttaki düğüme işaret eder. Bu şekilde bir yapıda sorular ağaç şeklinde (Şekil 2.1.) bir hiyerarşi oluşturmaktadır (42). Dışarı çıkan kenarlara sahip bir düğüme dahili düğüm denir ve bunun dışında kalan tüm düğümlere yapraklar denir (43). Şekil 2.1 'de karar ağaçları için temsili gösterimde, düğümler daire ile temsil edilirken, yapraklar üçgenler ile gösterilir (43). Karar ağaçlarının ilk hücrelerine ise kök (*root*) denir.



Şekil 2.1. Karar ağaçları düğüm ve yaprak yapısına bir örnek.

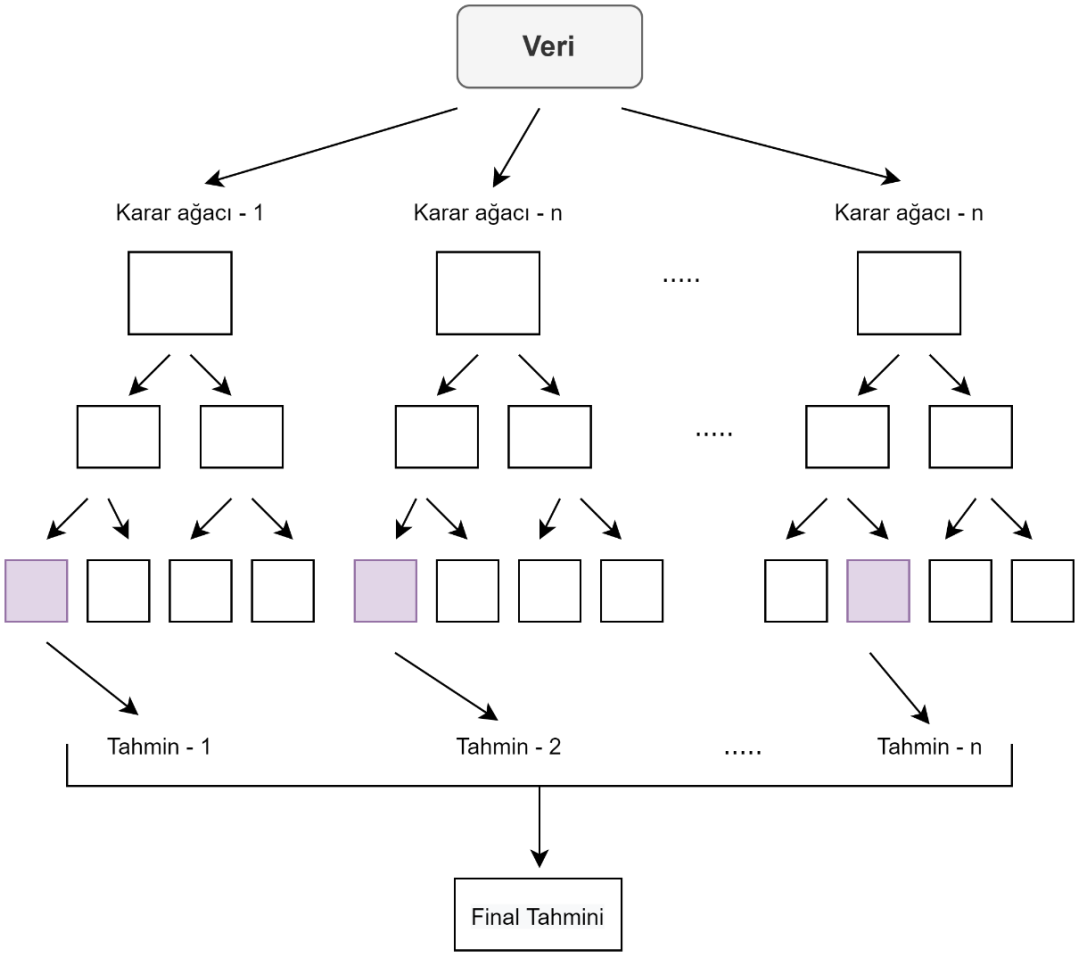
İlk hücre olan kökteki koşula göre verideki her bir gözlem, düğümler yardımıyla sınıflandırılır. Bir öğe, en üstteki düğümden bir yaprağa giden yolu izleyerek, söz konusu öğeye uygulanan cevaplara göre bir sınıfa ayrılmaktadır. (42). Her düğüm test ettiği öznitelikle etiketlenirken dalları, özniteliğin karşılık gelen değeriyle etiketlenir (43). Karar ağaçları, veriyle ilgili basit soruları anlaşılır bir şekilde bir araya getirdiği için sinir ağları ve destek vektör makineleri gibi diğer sınıflandırıcılardan bazı durumlarda daha yorumlanabilir bir yöntemdir (42).

Rastgele Orman Algoritması

Rastgele orman, birden çok karar ağacından yararlanarak çalışan denetimli bir makine öğrenme algoritmasıdır (44). Karmaşık sorunları çözmek için birçok sınıflandırıcıyı birleştiren bir teknik olan topluluk öğrenmeyi kullanan bir yöntemdir. Algoritma tarafından oluşturulan orman, torbalama toplama yoluyla eğitilir (44). Rastgele ormanda kullanılan rastgele örnekleme ve topluluk öğrenmesi, daha yüksek genellemelerin ve de daha doğru tahminlerin elde etmesini sağlar (45). Rastgele orman algoritmasında kullanılan topluluk öğrenme yöntemleri, bir dizi sınıflandırıcı olan karar ağaçlarından oluşur ve en popüler sonucu belirlemek için bunların tahminleri toplanır. En iyi bilinen topluluk yöntemleri, önyükleme toplaması olarak da bilinen 1996 yılında tanıtılan torbalama (*bagging*) yöntemidir (46). Bu yöntemde, eğitim kümesindeki rastgele bir veri örneği, tek tek veri noktalarının bir kereden fazla seçilebileceği anlamına gelen değiştirme ile seçilir (46). Bu ise varyansı azaltarak iyileşen tahmin performansı sağlayan torbalama şemasından gelir (45). Rastgele

orman, diğer bilinen makine öğrenmesi yöntemlerine kıyasla daha yüksek tahmin doğruluğu ve daha kolay model yorumlaması sağlamaktadır (45).

Algoritmanın çalışma prensibi, karar ağaçlarının tahminlerine dayalı olan, çeşitli ağaçların çıktısının ortalamasını alarak sonucu tahmin etmesidir (44). Modelde eğitim veriyi, çeşitli karar ağaçlarını eğitmek için modele verilir. Daha sonra öğrenmeyi yapan model hiç görmediği örneklerle test edilir. Test sonu Şekil 2.2'de görüldüğü gibi birçok karar ağacında tahmin sonuçları elde edilir ve sonucunda tüm bu karar ağaçlarının tahminlerini dikkate alan final sonucu, nihai sonuç olarak bildirilir.



Şekil 2.2. Rastgele orman algoritması şeması.

Rastgele orman bir karar ağacı algoritmasında meydana gelen sınırlamaları ortadan kaldırma özelliğine sahiptir (37). Ana odağında üç önemli özelliği barındırır ve bunlar: birçok uygulama için doğru tahmin, model eğitimiyle her özelliğin önemini ölçme, örnekler arası ikili yakınlığın eğitilmiş model tarafından ölçülebilmesidir (45).

Rastgele ormanda yaygın olarak kullanılan önem ölçüsü Gini önemi, elde edilen ağaçlarındaki Gini indeksinden doğrudan elde edilir (45). Permütasyon önemi, aynı anda çok değişkenli etkileşimleri ve diğer özellikleri hesaba katarak, her değişkenin etkisini kapsayan bir başka önemli özellik sıralama ölçüsüdür (45).

Son yıllarda biyoenformatikteki çeşitli problemlere başarıyla uygulanan rastgele orman algoritmasının, bu alanda yoğun tercih edilmesinin sebebi, karmaşık etkileşimleri içeren verinin, rastgele orman algoritmasının çeşitli veri türlerine uygulanabilir özelliğinden kaynaklanmaktadır. Gen ekspresyon sınıflandırması, kütle spektrumu protein ekspresyon analizi, protein-protein etkileşimi tahmini gibi çeşitli biyoenformatik problemlere uygulanmıştır ve bu alanda kullanımı yaygındır (45).

2.2.2. Biyolojik Veri Analizinde Makine Öğrenmesinin Kullanımı

Biyoenformatik, karmaşık ve yüksek boyutlu biyolojik veriyi istatistik, matematik ve bilişim teknolojilerinden yararlanarak, analiz etmek ve yorumlamak için hesaplamalı araçlar geliştiren multidisipliner bir alandır. Çeşitli genetik ve epigenetik değişikliklerle belirlenen bir hastalık olan kanserinin gelişiminde rol alan mekanizmalar hala tam olarak anlaşılamamıştır (47). Kanser biyoenformatiği; kanserli hücrelerin tanınması, kanserde metastaz oluşumu ve kanserde genom analizinin kişiselleştirilmesi gibi birçok farklı kanser çalışmasında kullanılmaktadır. Kanser biyoenformatiği, her bir hastanın gen ve protein varyasyonlarına dayalı olarak en güvenli ve en etkili tedavi stratejisini sağlayan kişiselleştirilmiş tıp uygulamalarının belirlenmesinde önemlidir. Kanser biyoenformatiği uygulanan tedavilerin etkinliğinin tahmin edilmesinde de önemli rol oynamaktadır (48).

Mevcut biyolojik veri miktarının son teknolojik gelişmeler yardımı ile hızlı bir şekilde katlanarak artması iki sorunu ortaya çıkarır. Bunlar, bilgiyi verimli depolama ve bu veriden faydalı bilgilerin çıkarılmasıdır (36). Mevcut veriden faydalı bilgilerin

eldesi, heterojen verinin altında yatan mekanizmayı biyolojik bilgiye dönüştürebilen araç ve yöntemlerin geliştirilmesi ana zorluklardan biridir (36). Biyoenformatikte makine öğrenmesi kullanımının genişlemesi üç faktörün birleşimidir ve bunlar; veri, bilgisayar ve teorik olasılık çerçevesidir (49). Biyoenformatik ve makine öğrenmesi yöntemleri biyoloji ve tıpta çeşitli sorunları çözmek için önemli bir etkiye sahiptir (49).

Genomik , biyoenformatikte önemli alanlardan biridir (36). Büyük boyutlu genomik verinin analizi için kullanışlı bir yöntem olan makine öğrenmesi, genomikte çok sayıda alanda kullanılmıştır (32). Makine öğrenmesi algoritmaları, çok çeşitli genomik dizi öğelerine açıklama eklemek için kullanılır. Bunun yanında, diğer genomik testler tarafından oluşturulan girdi verisini de kullanabilir (32). RNA dizilimi (RNA-seq) ifade verisi, kromatin verisi, protein-protein etkileşim verisi, histon modifikasyonu veya transkripsiyon faktörü bağlanma verisi gibi birçok farklı veri tahmine dayalı algoritmaların girdisi olarak kullanılabilir. Genomik dizi de dahil olmak üzere çok çeşitli veri tiplerinden herhangi biri veya daha fazlası girdi verisi olarak makine öğrenmesinde kullanılabilir (32).

2010 yılında yapılan bir çalışmada, meme kanseri veri setindeki hastalarda hastalık nüksünü belirlemek için makine öğrenmesi ve istatistiksel değerlendirme yöntemlerinin performansın değerlendirilmiştir (50). Makine öğrenmesi algoritmalarına dayalı yöntemlerin, hasta sonucunun tahmininde, istatistiksel yöntemlere kıyasla daha yüksek sonuçlar verdiğini gösterilmiştir (50). Bazı iyi bilinen algoritmalar (destek vektör makineleri, rastgele orman, bayes ağları vb.) genomik, proteomik, sistem biyolojisi ve diğer birçok alanda uygulanmıştır (36). Prokaryotik ve arke genomlarında yeni RNA genlerinin tahmini için bilinen RNA'lar arasındaki ortak özellikleri çıkarmak amacıyla sinir ağlarını ve destek vektör makinelerini kullanan bir makine öğrenmesi yaklaşımı geliştirilmiştir (51). Bu yaklaşım sonucu deneysel çalışmanın öncüsü olarak hesaplamalı tahmin edilebilecek birçok tanımlanmamış RNA olduğu ortaya koyulmuştur (51). Bir başka çalışmada ise dizi ve ifade verisi gibi çeşitli kaynaklardan yararlanıp Bayes ağlarını kullanarak operonları tahmin etmek için olasılıksal bir yaklaşım sunulmuştur (52).

2.3. Biyoenformatikte Omik Veri Tipleri

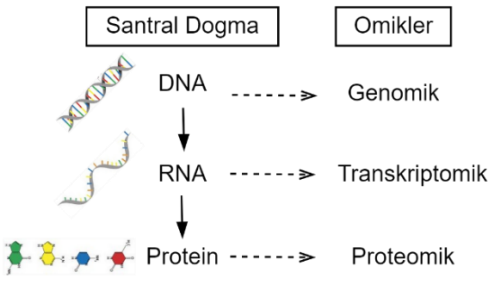
2.3.1. Omik Kavramı

Birçok farklı teknolojilerin bir araya getirilerek moleküller arasındaki ilişkileri, moleküllerin rollerini ve organizmadaki tüm hücrelerin etkilerini araştırmak için yapılan çalışmalar, omik kavramı ile ifade edilir. Moleküler bir terime "omik" ifadesinin eklenmesi, bir dizi molekülün kapsamlı şekilde değerlendirmesini ifade etmektedir (53). Büyük miktarda biyolojik verinin düşük bütçeli, yüksek verimli analizinin yapılmasına olanak sağlayan teknolojilerin gelişmesiyle mRNA, proteinler ve metabolitlere yönelik önemli çalışmalar yapılmış ve bu çalışmalar yeni "omik" araştırma alanlarının gelişmesine yol açmıştır (54). Bu omik teknolojilerde genomik, proteomik, metabolomik, transkriptomik, epigenomik gibi birçok alan mevcuttur. (54). Omik yaklaşımlar yüksek verimli ve veri odaklı yaklaşımlardır. Bu yaklaşımlarla ölçülen birden çok moleküler tür arasındaki ilişki farklı parçaların ilişkisi olarak değil, bir bütünleşik sistem gözü ile değerlendirilebilir. Yüksek verimli omik yaklaşımların ürettiği büyük miktarlardaki verinin istatistiksel ve hesaplamalı yaklaşımla analiz edilmesine ihtiyaç vardır (54, 55).

Hücresel seviyelerde farklı molekülleri analiz etmek için kullanılan omik yaklaşımlardan sadece bir tek omik yaklaşımın karmaşık bir yapıya sahip biyolojik sistemleri karakterize etmede yetersiz olabileceği belirtilmiştir (56). 2001 yılında yapılmış bir çalışmada, glikoliz kontrolünün sadece gen ifadesi ile kontrol edilmediği ve bu olayın görevlendirmesinde metabolik, proteomik ve genomik seviyeler arasında ortak bir paylaşım olduğu gösterilmiştir (57). Bu birlikte etkiden sonra çoklu omik yaklaşımlar üzerine çalışmalar yapılmıştır. Bu çoklu omik yaklaşım ile organizmanın sahip olduğu karmaşık metabolik ağların ve düzenleyici mekanizmaların belirlenebildiği gösterilmiştir. (58, 59).

2.3.2. Omik Teknolojileri

Genetik bilgi akışının bir özeti olan santral dogmanın her bir aşamasında üretilen ürünler farklı omik teknolojiler ile incelenmektedir (Şekil 2.3.).



Şekil 2.3. Santral dogma akışı.

Genomik

Bir organizmadaki yapısal ve işlevsel fonksiyonları kodlayan tüm genleri tanımlayarak genlerin birbirleri ve çevreleri ile etkileşimlerini inceleyen bilim dalına genomik denir (60). İlk olarak ortaya atılan omik disiplin olan genomik, sadece varyant ya da genlere odaklanan genetiğin aksine tüm genomun incelenmesine odaklanan bir disiplindir. Bir organizmanın genomunun incelenmesi, genomun karmaşık, biyolojik işlevini anlamak için ve genlerin statik dizilerini ortaya çıkarır (60). Genomik teknolojilerin ortaya çıkışı, hücre ve doku fonksiyonunun moleküler ayrıntılarını anlamayı kolaylaştırmıştır (60).

Transkriptomik

Bir organizmada, genom çapında üretilen tüm RNA transkriptlerinin toplamı olan transkritomu araştırmak için yapılan uygulamalara transkriptomik denir (54). Transkriptomik mRNA moleküllerinin dinamik ifadesinin nicel ölçümlerini mümkün kılmıştır (54). Her hücre, gelişiminde farklı zamanlarda ve farklı fizyolojik koşullarda farklı genleri ifade eder (60). Trankriptomik analiz, gen ifadesinin farklı organizma veya dokularda nasıl değiştiğini inceleyerek insan hastalıklarının anlaşılmasında önemli etkiye sahiptir (60). Bu bilim dalında da diğer omik alanlarında olduğu gibi yüksek verimli metotlar kullanılır.

Proteomik

Proteomik, gen ürünleri olan proteinlerin bir hücre, doku veya organizmadaki toplamının tanımlanması ve miktarının belirlenmesi için geliştirilen teknoloji uygulamalarını içerir (61). Belli bir zamanda belli bir yerde bulunan tüm proteinlerin yapıları, miktarları, translasyon sonrası modifikasyonları, işlevleri, diğer proteinlerle ve makro moleküllerle olan etkileşimi gibi konuları aydınlatmaktadır (61). Genomik teknolojideki ilerlemeler, çok büyük miktarda biyolojik veri üretmesine olanak sağlamıştır. Analitik teknolojideki gelişmelerle birlikte proteomik, birçok farklı yönün incelenmesi için giderek daha önemli olmuştur (60). Proteomik, hastalığın erken teşhisi, prognozu ve hastalık gelişiminin izlenmesi için oldukça önemlidir (61).

Metabolomik

Bir organizmada bir süre için, dokuda ortaya çıkan küçük moleküllü metabolizma ürünlerinin belirlenmesi, miktarının ölçümü ve tanımlanmasına metabolomik denilmektedir (60). Genomik, transkriptomik ve proteomik teknolojilerindeki hızlı gelişmeyle birlikte metabolomik alanına ilgi artmıştır. Hem genom tarafından kodlandığı hem de çevresel faktörlerin etkilerini taşıdığı için tıbbi uygulamalarda giderek önemli olmaktadır (60). Metabolizma ürünleri olan metabolit seviyeleri metabolik işlevi yansıtır ve normal aralığın dışındaki bozulmalar hastalık göstergesi olabilir (53).

2.4. Kanserde Sağkalım Literatürü

2.4.1. Kanserde Sağkalımın Deneysel ve Klinik Çalışmaları

Sağkalım tıbbi çalışmada ölüme kadar geçen süre ile ilgilenilen olaydır. Bununla birlikte, kanserde, bir başka önemli ölçü, tedaviye yanıt ile nüks veya nükssüz sağkalım arasındaki süredir. Olayın ne olduğunu ve gözlem süresinin ne zaman başlayıp ne zaman bittiğini belirtmek önemlidir (21). Sağkalım istatistikleri, kanser hastalarının sağkalımını ve hastalıklarının muhtemel seyrini tahmin etmek için en çok kullanılan ölçülerdir ve hastalar, klinisyenler, araştırmacılar ve politika yapıcılar tarafından büyük ilgi görmektedir (1). Göreceli sağkalım genellikle bir kanser hastasının hayatta kalmasını tahmin etmek için kullanılır. Net sağkalım olarak da

adlandırılan nispi sağkalım, kanser teşhisinin net etkisini temsil eder. Kanserin tek olası ölüm nedeni olduğunu varsayarsak hayatta kalma şansını verir. Kanser hastaları aynı zamanda rekabet eden sebeplerden de ölebildikleri için, hastanın kanserden ölme, diğer "rakip" sebeplerden ölme veya hayatta kalma şansı, kanser hastaları ve onları tedavi eden klinisyenler için daha alakalı sağkalım istatistikleridir (1).

Hayatta kalma ve ölümlülük aynı madalyonun iki yüzü olarak düşünülebilir: bir kişi ya yaşıyor ya da ölü olabilir. Ancak kanser istatistiklerine bakılırsa, hayatta kalma ve ölüm, farklı madeni paraların iki yüzü olarak tanımlanabilir (1). Mortalite, tüm nüfus (yani kanserli ve kansersiz insanlar) arasındaki kanser ölümlerinin sayısını ölçer. Popülasyondaki bir kişinin belirli bir süre, genellikle bir yıl içinde kanserden ölme şansıdır. Hayatta kalma ise, kanserli insanlar arasında yaşayan sayısıdır. Bir kanser hastasının teşhisten birkaç yıl sonra (tipik olarak beş veya 10 yıl) hayatta kalma şansıdır (1). Kanser hastaları için, ilgilenilen ana istatistik, popülasyon mortalitesi değil, bireysel sağkalımdır. Ölümlülük değil hayatta kalma, kanser hastalarının bilmek istediği soru olan "teşhisle hayatta kalma şansım nedir?" sorusunu yanıtlar (1). Sonuç olarak hayatta kalma, belirli kanser türleri ve kanser hastaları için prognoz sağlayabilen klinik bir perspektiften önemli bir istatistiktir.

Birçok klinik çalışmada, tedavilerin karşılaştırılması için birincil sonuç, hastalıkla ilgili bir olayın meydana gelme zamanıdır (62). Bu tür sonuçların görüntülenmesi için en yaygın olarak benimsenen yöntem, zamana göre olayı yaşayan hastaların oranını gösteren Kaplan-Meier sağkalım grafikleridir (62). Kanser tedavisinde nihai amaç hastayı iyileştirmektir. Hasta sağkalımındaki zamansal eğilimleri izlemek, bu alandaki performansı değerlendirmek için ideal bir yaklaşımdır (63). Meme kanserli kadınları kapsayan bir çalışmada çıkan sonuca göre, uzun süreli kemoterapinin, kemoterapi uygulanmamasına kıyasla nüks ve ölüm riskini önemli ölçüde azalttığını göstermiştir (64). Genel olarak, kemoterapi ile tedavi edilen genç kadınlar, kemoterapi almayan grupla karşılaştırıldığında, 10 yıl içinde ortalama sağkalımda önemli ölçüde artış olmuştur (64).

2000-2008 yılları arasında malign kanser teşhisi konan 15-39 yaş arası ergen ve genç yetişkin hastalardan seçilen kanser türleri, beş yıllık nispi sağkalım açısından histoloji, evre ve reseptör alt tiplerine göre analiz edilmiştir (65). Ergen ve genç

yetişkinler grubuna göre daha genç ve daha yaşlılar arasında kanser ölüm riski için tehlike oranları tahmin edilmiştir. Ergen ve genç yetişkinlerin sağkalımı, kadın meme kanseri, akut lenfoid lösemi ve akut miyeloid lösemi için daha düşük bulunmuştur (65). Bazı kanser türlerinde yaş grupları açısından sağkalım farklılıklarına vurgu yapılmıştır.

Kanserde tedavinin, hasta sağkalımı üzerine etkisini araştıran çalışmalar yayınlanmıştır. Tedavinin sağkalıma etkisini araştıran bir çalışmada laparoskopik cerrahinin, açık cerrahiye kıyasla daha yüksek iyileşme ve daha az morbidite ile ilişkili olduğu belirlenmiştir (66). 2009 yılında yayınlanan bir çalışmada, ölümcül bir kanser türü olan kolon kanseri için farklı cerrahi müdahaleler olan laparoskopik cerrahi ve açık cerrahinin hasta sağkalımı üzerindeki etkisi araştırılmıştır (66). Bir diğer çalışmada hasta omik verisi ile sağkalım analizi yapılmıştır. Çalışmada, 26 farklı kanser türünde hayatta kalma analizi yapmak için RNA sekansına dayalı transkriptomik veri kullanılmış ve kansere özgü genlerden oluşturulan imzalar, hayatta kalma ile tümör tipine özgü korelasyonlar göstermiştir (67).

2.4.2. Kanser Sağkalımında Hesaplamalı Çalışmalar

Heterojen bir hastalık olarak karakterize edilen kanserin erken teşhisi ve hastalık prognozu, hastanın ileri aşamalardaki klinik yönetimini kolaylaştıracağı için kanser araştırmalarında bir zorunluluk haline gelmiştir (68). Kanser hastalarını risk gruplarına ayırmanın önemi ile biyoenformatik alanından birçok araştırma ekibi, makine öğrenmesi yöntemlerini ele alarak kanserli durumların ilerlemesini ve tedavisini modellemek amacıyla hesaplamalı yöntemleri kullanmıştır (68). Makine öğrenmesi tekniklerinin avantajlarından biri standart istatistiksel yöntemlerin aksine, doğrusal olmayan daha karmaşık ilişkileri dikkate alarak modelleme yapmasıdır (69). Karmaşık veri kümelerinden temel özellikleri tespit etme yeteneği, makine öğrenmesi tekniklerinin önemini ortaya koymaktadır (68).

Son yıllarda, birçok farklı makine öğrenmesi teknikleri ve özellik seçim algoritmaları, hastalık prognozu ve tahmini için yaygın olarak uygulanmıştır (70). Kanser tahminiyle uğraşırken, kişi üç farklı görevle ilgilenilebilir: (i) kanser duyarlılığının tahmini, (ii) kanser nüksü tahmini ve (iii) kanser sağkalımının tahmini

(70). İlk iki durumda, kişide kanser geliştirme olasılığını ve hastada bir kanser türünü yeniden geliştirme olasılığını bulmayı amaçlar. Diğer durumda ise, kanser teşhisi veya tedavisinden sonra hastalığa özgü veya genel sağkalım gibi bir hayatta kalma sonucunun tahmini amaçlanmaktadır (70). Destek vektör makineleri ve benzeri diğer denetimli öğrenme tekniklerinin kullanımında son yıllarda kanser tahmini ve prognozuna yönelik artan bir eğilim kaydedilmiştir (71, 72). Bu sınıflandırma algoritmaları, kanser araştırmalarında ortaya çıkan çeşitli problemlerde yaygın olarak kullanılmaktadır.

Kanserde hasta prognozu ile ilgili kararlar geçmişte histolojik klinik ve popülasyona dayalı veri göz önüne alınarak verilse de bu tür parametreler sağlıklı karar vermek için yeterli bilgi sağlamazlar (68). Genomik bilimin hızla gelişmesiyle, toplanan ve tıbbi araştırma topluluğunun kullanımına sunulan çok büyük miktarda genomik kanser verisi üretilmiştir (68). Bu verinin işlenmesinde makine öğrenmesi teknikleri, bir kanser türünün gelecekteki sonuçlarını etkili bir şekilde tahmin edebilir, karmaşık veri kümelerinden aralarındaki kalıpları ve ilişkileri keşfedebilir. Aynı kanser türü arasında, farklı klinik sonuçlar, farklı tedavi yaklaşımları ve spesifik genetik bozukluklara dayalı ayrı alt gruplar bulunmaktadır. Bu da hesaplama teknikleri küçük hasta gruplarını az maliyetli ve etkili bir şekilde belirleyerek onlara yardımcı olabilecek, bireyselleştirilmiş tedavi yaklaşımının temelidir (68).

2013 yılında, meme kanseri teşhisi konmuş kadınlarda sağkalımın değerlendirilmesi için yapılan bir çalışmada bir tahmin modeli geliştirilmiştir (73). Çalışma sonunda, tahmine dayalı modelin kullanım kolaylığı ve sonuçlara daha hızlı ulaşma süresinin, meme kanseri hastaları için doğru prognoza yol açtığı belirtilmiştir (Tablo 2.1.). Bir başka çalışmada, gen ifadesi verisi ve klinik verinin yapay sinir ağları kullanılarak akciğer kanseri hastalarında hayatta kalma risk tahminleri araştırılmıştır. Yüksek risk grubundaki hastalardan düşük riskli hastalara kıyasla daha düşük bir medyan genel sağkalım elde edilmiştir (74) (Tablo 2.1.).

Kişiselleştirilmiş tedavi önerileri sağlamak için bir hastanın ortak değişkenleri ve tedavi etkinliği arasındaki etkileşimleri modellemek amacıyla yapılan bir çalışmada, klasik COT yönteminin es geçtiği karmaşık etkileşimler için derin sinir ağları kullanılmıştır. Sağkalım analizinde uygulanan makine öğrenmesi yönteminde

klasik modele göre modelin esnekliğiyle daha yüksek performansa ulaşılmış ve bireysel tedavi tahminlerinde etkili tedavi yöntemlerini önerdiği gösterilmiştir (75) (Tablo 2.1.). Benzer bir çalışmada, derin sinir ağları kullanılarak, hayatta kalma verisinin doğasında bulunan sansürü yakalayarak, sağkalım süresi ve olayının tahmini ortak dağılımını öğrenmek için bir sinir ağı eğitilmiştir (76) (Tablo 2.1.).

Kadınlarda en agresif ve ortalama sağkalımı düşük meme kanserinde doğru prognoz tahmini hastaların gereksiz ağır ve pahalı tedaviler almasının önüne geçebilir (77). Önceki çalışmalarda tahmine dayalı modeller oluşturulurken genellikle gen ifadesi verisinden yararlanılmıştır. Ancak gelişen makine öğrenmesi yöntemleri ve büyük boyutlu moleküler verinin eldesi ile meme kanserinin daha kapsamlı analizi için de yeni fırsatlar doğmuştur. Bu kapsamda meme kanserinin tanısı, tedavisi ve önlemesine yönelik yeni çabalarla iyileştirilmiş analizler ortaya çıkabilir. Bu fikirle yola çıkıp çok boyutlu bu verinin entegre edilmesiyle oluşturulmuş bir derin sinir ağı öneren çalışma yapılmıştır. Hem klinik bilgileri hem de genler arasındaki meme kanserine özgü ilişkileri entegre etmek için derin sinir ağlarının kullanılabileceği belirtilmiştir (77). Gen ifadesi, kopya sayısı varyasyonu ve klinik veriyi içeren model tek boyutlu veri içeren modellerden daha yüksek performans göstermiştir. Meme kanserinde sağkalım tahmini için farklı veri türlerinin birlikte kullanımının performansı iyileştirdiği gösterilmiştir (Tablo 2.1.).

Dünya çapında erkeklerde önde gelen ölümden sorumlu kanser türü olan karaciğer kanseri ile ilgili yapılan bir çalışmada, çoklu omik veri ve klinik verinin kullanılmasıyla oluşturulan model ile iki farklı sağkalım alt tipi tanımlanmıştır (78) (Tablo 2.1.). Çoklu omik veri ile yapılan modelin tekli omik modellere kıyasla daha yüksek performans elde ettiği belirlenmiştir.

33 kanser türünden oluşan büyük bir grup hasta verisiyle yapılan çalışmada, hasta risk tahmini için multimodal bir model sunulmuştur (79). Omik veri, klinik veri ve histopatolojik mikroskop slaytlar kullanarak tüm kanser türlerinde yüksek tahmin performansına sahip, eksik veriyi sorunsuz bir şekilde işleyen yeni bir model önerilmiştir. Hem geniş bir grup kanser türü için hem de birden çok moleküler verinin bir arada kullanılması ile elde edilmiş bir model ortaya konulmuştur (Tablo 2.1.).

2019 yılında yapılan, 20 kanser türünde sağkalım tahmini için yapılan çalışmada mRNA, miRNA, klinik veri ve histopatolojik mikroskop slaytlar kullanılmıştır (80). Her veri türü için evrişimsel sinir ağları (*Convolutional Neural Networks, CNN*) mimarisi kullanılmıştır. Modele, performansı artırdığı önceki çalışmalarda tanımlanan, Cox kayıp fonksiyonu son bir tahmin katmanı olarak eklenmiştir. Genomik ve görüntü verisinin sağkalım tahmininde önemini incelemek için farklı veri kombinasyonlarını araştırmışlar. Çalışma sonucunda çoklu verinin kullanıldığı yeni bir yaklaşım önerilmiştir. Böbrek kanserinde benzer çalışmalardan daha düşük, kemik iliğinde ise daha yüksek performansa sahip olduğu belirlenmiştir. Önerilen yöntemle eksik veri ele almıştır ve 20 kanser türünü öngörecek çalışmalarla karşılaştırılabilir bulunmuştur. Ayrıca daha çeşitli genomik verinin entegre edilmesinin daha yüksek sağkalım tahmini performansı sağlayabileceğini belirtmişlerdir. (Tablo 2.1.).

Akciğer kanseri en sık görülen kanser türüdür ve ölüm oranı yüksektir, bu da kanser türü için sağkalım tahmini geliştirme ihtiyacı doğurmuştur. Akciğer kanserinin en yaygın alt türü olan akciğer adenokarsinomu (LUAD) tüm akciğer kanserinin yaklaşık %40'ını oluşturmaktadır. LUAD sağkalım çalışmalarında daha önce sadece tek mRNA veya miRNA gibi tekli omik verinin kullanılması, sağkalım analizinde çoklu omik verinin kullanıldığı yeni bir çalışmayı doğurmuştur (81). 4 farklı omik verinin kullanıldığı çalışmada derin öğrenmeye dayalı otomatik kodlama yaklaşımı ile yeni bir model önerilmiştir. Çoklu omik veri ile yapılan modeller tekli omik veriyle yapılan modellerden daha yüksek performans göstermiştir (Tablo 2.1.).

Meme kanserinde sağkalım analizinde çoklu-omik veri kullanımı araştıran bir çalışmada, 6 farklı çoklu-omik veri kombinasyonu ile model yapılmıştır. SALMON, bir sinir ağını mRNA ve miRNA verisiyle beslemek yerine, ortak ifade analizinden elde edilen matrisleri girdi olarak alır. Böylece, girdi özelliklerini yaklaşık %99 oranında azaltarak yüksek boyutluluk sorununun üstesinden gelir. Model yapımında daha fazla omik verisi kullanıldığında performansın arttığı gözlemlenmiştir (82). **Tablo 2.1.** Kanser sağkalım tahmini için kullanılan makine öğrenmesi yöntemleriyle ilgili yayınlar.

Yayın	Makine öğrenmesi yöntemi	Kanser tipi	Veri tipi	Sağkalım tanımı
Park ve ark., 2013(73)	YSA, SVM, SSL	Meme kanseri	Klinik	5 yıldan az ve çok yaşamış iki grup
Chen ve ark., 2014 (74)	Yapay sinir ağları	Akciğer kanseri	mRNA, Klinik	Risk alt grupları
Katzman ve ark., 2018 (75)	Derin sinir ağları	Meme kanseri	mRNA, Klinik	Sağkalım olasılığı
Lee ve ark., 2018 (76)	Derin sinir ağları	Meme kanseri	mRNA, Klinik	Sağkalım olasılığı
Sun ve ark., 2019 (77)	Derin sinir ağları	Meme kanseri	mRNA, KSV, Klinik	Uzun-kısa vadeli sağkalım
Chaudhary ve ark., 2018 (78)	Otomatik kodlayıcı, K-ortalama, DVM	Karaciğer kanseri	mRNA, miRNA, DNAm, Klinik	Risk alt grupları
Vale-Silva ve ark., 2020 (79)	Yapay sinir ağları	33 kanser türü	mRNA, miRNA, DNAm, CVN, WSI	Risk oranı
Cheerla ve ark., 2019 (80)	Derin sinir ağları	20 kanser türü	mRNA, miRNA, Klinik, WSI	Genel sağkalım
Lee ve ark., 2020 (81)	Otomatik kodlayıcı, Rastgele orman	Akciğer kanseri	mRNA, miRNA, DNAm, KSV	Sağkalım alt grupları
Huang ve ark., 2019 (82)	Yapay sinir ağları	Meme kanseri	mRNA, miRNA, TMB, CNB, Klinik	Risk oranı

3. GEREÇ VE YÖNTEM

3.1. Veri Düzenleme

Veri setleri kamuya açık çevrim içi bir veri tabanı olan "*Genomic Data Commons*" (*GDC*) dan elde edildi (83) (Şekil 1.1). Ulusal Kanser Enstitüsü'ne (*National Institutes of Health, NCI*) bağlı *GDC,* onkolojide kişiselleştirilmiş tıbbı teşvik eden bir veri paylaşım platformudur. Bilinen kanser bazlı en büyük veri tabanlarından biridir. *GDC* kanser genomik çalışmaları arasında veri paylaşımını sağlayan bir havuz ve kanser bilgi tabanı sağlamaktır. Kanser Genom Atlası (*The Cancer Genome Atlas, TCGA*) ve Etkili Tedaviler Oluşturmak için Terapötik Olarak Uygulanabilir Araştırma (*The Therapeutically Applicable Research to Generate Effective Treatments, TARGET*) gibi *NCI* Kanser Genomik Merkezi'ndeki (*Center for Cancer Genomics, CCG*) çeşitli kanser genom programlarını içermektedir (84). Ulusal Sağlık Enstitüsü (*National Institutes of Health, NIH*) tarafından, kanser genomik profilleri için geniş bir atlas oluşturmak amacıyla TCGA projesi, 2006 yılında başlatılmıştır. (84). TCGA projesi, farklı disiplinlerden araştırmacıların ve birçok kurumun dahil olduğu 33 farklı kanser türünü kapsayan bir çalışmadır (85). Kanserin teşhisi, tedavisi ve önlenmesi amacıyla devamlı olarak iyileştirilen veri, herkesin kullanması için kamuya açık olarak yayınlanmıştır (85).

TCGA kapsamında incelenen her kanser tipi için elde edilen genomik, epigenomik, transkriptomik ve proteomik veriyle bir atlas oluşturularak ilgili kanser için geniş çapta bir moleküler karakterizasyon yayınlanmıştır (86). TCGA programı kapsamında, 2.5 petabayttan fazla genomik, epigenomik, transkriptomik ve proteomik veri üretmiştir. Kanser hastalarında çoklu-omik profillerini oluşturmak için bir çaba ile ortaya çıkmıştır. Çeşitli kanser türleri için kansere neden olan büyük genomik değişiklikleri kataloglamayı ve keşfetmeyi, çeşitli tümör tiplerinin ve alt tiplerinin moleküler profillerini oluşturmayı ve analiz etmeyi amaçlamaktadır (87). TCGA projesi, kanserde moleküler karakterizasyon ve zengin bir genomik veri kaynağı sunar. Sağlık ve bilim teknolojilerinde ilerleme ve klinikte kanser hastalarında tedavi şekillerinde değişim gibi birçok önemli gelişmeye sebep olmuştur (88). Proje, birçok farklı kanser türünde binlerce hastanın kanserlerindeki moleküler farklılıkları öne çıkararak, kişiselleştirilmiş tıp yaklaşımına katkıda bulunmaktadır.

Bu tez çalışmasında, 13 farklı kanser türü için çalışılmış TCGA projeleri kullanılmıştır (Bkz. Şekil 1.1). 4 farklı omik veri tipi çalışmada kullanılmıştır. Kullanılan omik veri tipleri: kopya sayısı varyasyonu (KSV), mutasyon, gen ifadesi ve miRNA ifadesi. Bu omik veri tiplerine ek olarak hastaların klinik ve ilaç kullanım bilgileri çalışmada kullanılmıştır. Veri düzenleme aşamasında çalışmada kullanılan omik veri tiplerinden bir ya da daha fazla eksik verisi olan hastalar kullanılmamıştır. Kanser türlerinde bu koşulları sağlayan hastalar belirlenmiştir. Her kanser türünde seçilen hasta sayıları Tablo 3.1.'de gösterilmiştir. Kullanılan veri türleri ile ilgili düzenlemeler maddeler halinde açıklanmıştır.

Tablo 3.1. Çalışmada kullanılan her projeden tüm omik veri tiplerine sahip hasta sayıları.

Kanser Türü	**TCGA Projesi**	**Toplam Hasta Sayısı**	**Seçilen Hasta Sayısı**
Akciğer-LUAD	LUAD	549	457
Akciğer-LUSC	LUSC	504	454
Böbrek	KIRP	291	268
Cilt	SKCM	470	238
Karaciğer	LIHC	377	300
Kemik İliği	LAML	200	107
Kolorektal	COAD	393	325
Meme	BRCA	1054	840
Mide	STAD	443	364
Pankreas	PAAD	185	166
Prostat	PRAD	500	243
Rahim Ağzı	CESC	307	279
Rahim	UCEC	560	507
Tiroit	THCA	507	477
Yumurtalık	OV	600	267

Seçilen Hasta Sayısı: veri tabanından elde edilen hastalarda çalışmada kullanılan omik veri tiplerine sahip olan hasta sayısı.

Kullanılan veri setlerinde gen ifadesi, kopya sayısı varyasyonu ve mutasyon verisinde binlerce gen için bilgi bulunmaktadır. Çalışmada tüm genler için bilgilerin kullanılması verinin boyutunu büyük olmasına neden olmuştur. Bu büyüt boyut veride daha fazla gürültü olmasını sağlamıştır. Bunun önüne geçmek için genlerin daha anlamlı bir alt kümesi ile çalışılmaya karar verilmiştir. Seçilecek genler için L1000 teknolojisi ile belirlenmiş genlerden yararlanılmıştır. L1000 teknolojisi, genom çapında mRNA ifadesini tahmin etmek için yüksek verimli bir yöntemdir. L1000, her deneyde yalnızca yaklaşık 1000 geni ölçer (89). L1000, insan hücrelerinden 978

dönüm noktası (*landmark*) genin mRNA transkript bolluğunu ölçen yüksek verimli bir gen ekspresyon tahlilidir (90). Gen ekspresyonunun hücre durumları boyunca benzer ekspresyon paternleri sergileyen gen kümeleri ile yüksek oranda korelasyona sahip olduğunu göstermiştir. Böyle bir korelasyon yapısı göz önüne alındığında, transkriptomun azaltılmış bir temsilini ölçerek herhangi bir hücresel durumu düşük maliyetle tespit etmenin mümkün olabileceği belirlenmiştir (89). L1000 gen bilgileri GEO (*Gene Expression Omnibus*) veri tabanında, platform "GPL20573" kullanılarak elde edilmiştir.

3.1.1 Klinik

TCGA projelerinde elde edilen klinik veri kullanılmıştır. Klinik veri özellikleri olarak: yaş, ırk, TNM evreleme sistemi, birincil teşhis, hastanın kanser geçmişi bilgisi, hastanın tedavi geçmişi bilgisi kullanılmıştır. Cinsiyet bilgisi tek bir cinsiyet ile ilişkili meme, yumurtalık, rahim ve rahim ağzı kanser türlerinde çıkarılmıştır. Ancak tüm kanser türlerinin birlikte değerlendirildiği model için düzenlenen klinik veride kullanılmıştır. Bazı kanser türleri için bu bilgilere ek olarak, kanser türünde önemli görülen klinik bilgiler eklenmiştir. Bunlar; meme kanserinde östrojen ve progesteron reseptör durumu, akciğer kanserinde sigara kullanım durumu gibi kansere özgü klinik bilgilerdir. Klinik veri türleri ve kullanıldığı kanser türleri Tablo 3.2.'de gösterilmiştir.

Tablo 3.2. Klinik veride öznitelikler.

Klinik Veri	**Kullanılan Model**
Yaş	Tüm kanser türleri
Cinsiyet	Meme, rahim, rahim ağzı ve yumurtalık kanseri dışındaki kanser türleri
Irk	Tüm kanser türleri
TNM evreleme sistemi	Tüm kanser türleri
Birincil teşhis	Tüm kanser türleri
Aile kanser hikayesi	Tüm kanser türleri
Önceki tedavi	Tüm kanser türleri
Önceki malignite	Tüm kanser türleri
Östrojen reseptör durumu	Meme
Progesteron reseptör durumu	Meme
Lenfatik inovasyon	Tüm kanser türleri
Vasküler invazyon	Tüm kanser türleri
Histolojik tip	Tüm kanser türleri
Sigara içme hikayesi	Akciğer
Rezidüel tümör	Tüm kanser türleri

Kanserde evreleme sistemlerinden en yaygın olarak kullanılanı Amerikan Kanser Komitesi tarafından hazırlanmış TNM evreleme sistemidir (11). Klinik veride bulunan hastanın TNM evreleme bilgisi çalışmada kullanılmıştır.

Yaş bilgisi olmayan az sayıda hasta için eksik olan yaş bilgisi yerine sıfır (0) değeri kullanılmıştır. Kategorik değişkenlerin tahminlerde kullanılabilmesi için makine öğrenmesi algoritmalarına uygun bir forma dönüştürülmesi gerekmiştir. Klinik veri için çalışmada tek sıcak kodlama (*one hot encoding*) yöntemi kullanılmıştır. Bu yöntemde kategorik değişkenler ikili (*binary*) olarak temsil edilir. Her değişkendeki kategoriler ayrı bir değişken olarak dönüştürülmüştür. Şekil 3.1'de örnek olarak cinsiyet bilgisi üzerinden tek sıcak kodlama ile ikili temsil oluşturulması gösterilmiştir. Buna benzer şekilde diğer tüm kategorik değişkenler düzenlenmiştir. Tüm hastaların veri bir araya getirilerek klinik verisi, algoritma için girdi verisi olarak hazırlanmıştır.

Hasta	Cinsiyet
1	Kadın
2	Erkek
3	Erkek
4	Kadın

⇨

Hasta	Kadın	Erkek
1	1	0
2	0	1
3	0	1
4	1	0

Şekil 3.1. Klinik özniteliklerin düzenlenmesine bir örnek.

3.1.2. Kopya Sayısı Varyasyonu

Veri tabanından her bir hasta için ayrı dosya şeklinde kopya sayısı varyasyonu (KSV) verisi elde edilmiştir. Her dosyada gen *Ensembl ID*'si ve kopya sayısı değerleri bulunmaktaydı. Gen ismi ve genin kopya sayısı bilgileri kullanılarak dokudaki seçilen tüm hastaların KSV verisi bir araya getirilmiştir. Tüm genler içinden L1000 gen listesindeki genler seçilmiştir. L1000 genlerinden veride olmayan genler veriden çıkarılmıştır ve KSV verisi algoritma için girdi verisi olarak hazırlanmıştır.

3.1.3. Gen İfadesi

Gen ifadesi verisinde her hasta için bir dosya indirilmiştir. Her dosyada gen *Ensembl ID*'si ve kilobaz milyon başına parça sayısı (*Fragments Per Kilobase Million, FPKM)* değerleri bulunmaktaydı. FPKM basit bir ifade düzeyi normalleştirme

yöntemidir. FPKM, gen uzunluğuna ve eşlenen toplam okuma sayısına göre okuma sayısını normalleştirir. Seçilen hastaların dosyaları bir araya getirilmiştir. L1000 gen listesindeki genler seçilmiştir. L1000 gen genlerinden veride olmayan genler veriden çıkarılmıştır ve gen ifadesi verisi algoritma için girdi verisi olarak hazırlanmıştır.

3.1.4. Mutasyon

İlgili projedeki mutasyon bilgisi bulunan hastaların mutasyon verisi tek bir *Mutation Annotation Format (MAF)* dosyası olarak indirilmiştir. Mutasyon verisinde sadece tek nükleotid polimorfizmi (*Single Nucleotide Polymorphism, SNP*) olan mutasyon varyant tipi kullanılmıştır.

Seçilen hastaların HUGO gen sembolü ve hastaya özel Case ID bilgisi alınarak, her hasta için mutasyona uğrayan genler ‘1’, mutasyona uğramayan genler ‘0’ olacak şekilde ifade edilerek düzenleme yapılmıştır. Oluşturulan veride L1000 gen listesindeki genler seçilmiştir. L1000 genlerinden veride olmayan genler çıkarılmıştır ve mutasyon verisi algoritma için girdi verisi olarak hazırlanmıştır. Mutasyon için oluşturulan özniteliklerin düzenlemiş son hali Şekil 3.2.’de gösterilmiştir.

Hasta	Gen-1	Gen-2	Gen-3	Gen-4	Gen-5	...	Gen-n
1	0	1	0	1	1	...	0
2	1	1	0	0	0	...	0
3	0	0	0	1	0	...	0
4	1	0	1	0	0	...	1

Şekil 3.2. Mutasyon için özniteliklerinin düzenlenmesi.

3.1.5. miRNA İfadesi

miRNA ifadesi veri tipi için hastaların ayrı dosyalarda bulunun verisi indirilmiştir. TCGA projelerinin tümünde 1881 miRNA için ölçüm bilgisi bulunmaktadır. Tüm miRNA’lar çalışmada kullanılmıştır. miRNA ifadesi verisi için ölçümlerin normalize edilmiş milyon eşlenmiş okuma başına (*per million mapped reads, RPM*) değerleri ve *miRNA_ID* bilgileri kullanılmıştır. Her dokuda seçilen hastaların verisi birleştirilerek miRNA verisi algoritma için girdi verisi olarak hazırlanmıştır.

3.1.6. İlaç Kullanım Bilgisi

Her kanser türünde hastaların tedavi sırasında kullandıkları ilaç bilgisi tek bir dosya halinde indirilmiştir. Birden fazla farklı isimle yazılan, yanlış olarak yazılan ya da yazım yanlışı olan ilaç isimleri *DrugBank* veri tabanından elde edilen bilgilerle düzenlenmiştir (91). *DrugBank*, ilaç ve ilaç hedefleri hakkında bilgi içeren kapsamlı, kamuya açık, çevrimiçi bir veri tabanıdır (91). İlaç ismi için ilacın genel ismi (*Generic Name*) dikkate alınmıştır. İlaç isimleri ve Case ID bilgisi kullanılarak veri düzenlenmiştir. Her hasta için ilgili ilacı kullanması durumunda '1', kullanmaması durumda '0' olacak şekilde ifade edilerek düzenleme yapılmıştır. Bu şekilde ilaç bilgisi algoritmanın kullanımına uygun hale getirilmiştir. Tüm hasta verisi birleştirilerek ilaç verisi algoritma için girdi verisi olarak hazırlanmıştır. Düzenlenmiş ilaç bilgisinin son haline bir örnek Şekil 3.3.'te gösterilmiştir. Her kanser türü için ilaç sayıları Tablo 3.3.'te gösterilmiştir. Tüm kanser türlerinin birleştirildiği ilaç verisinde toplam 129 ilaç için bulunmaktadır.

Hasta	İlaç-1	İlaç-2	İlaç-3	İlaç-4	...	İlaç-n
1	0	0	0	0	...	1
2	0	1	1	1	...	0
3	0	0	0	1	...	0
4	1	0	1	0	...	0

Şekil 3.3. İlaç için özniteliklerin düzenlenmesi.

Tablo 3.3. Kanser türlerine göre kullanılan ilaç sayıları.

Kanser Türü	**İlaç sayısı**
Akciğer-LUAD	16
Akciğer-LUSC	23
Böbrek	15
Cilt	27
Karaciğer	15
Kolorektal	15
Meme	40
Mide	24
Pankreas	13
Rahim Ağzı	15
Rahim	21
Yumurtalık	31

3.1.7. Sağkalım

Klinik bilgilerde yer alan hastanın son takip günü (*days_to_last_follow_up*), hayatta kalma durumu (*vital_status*), teşhis yılı (*year_of_diagnosis*) ve ölüm yılı (*year_of_death*) bilgileri kullanılarak elde edilmiştir. Sağkalım süreleri yıl olarak hesaplanmıştır. Algoritma ile ikili bir sınıflandırma yapıldığı için sağkalım süreleri dikkate alınarak iki grup oluşturulmuştur. Oluşturulan gruplarda yaşayan ve ölen hastaların aynı grupta olmasının önüne geçmek için her dokuda bir eşik değeri seçilmiştir. Farklı eşik değerleri seçildiğinde ölen-yaşayan hasta dağılımlarına bakılmıştır. Her doku için bir eşik değeri belirlenmiştir. Belirlenen eşik değerine göre seçilen hastalar çalışmada kullanılmıştır.

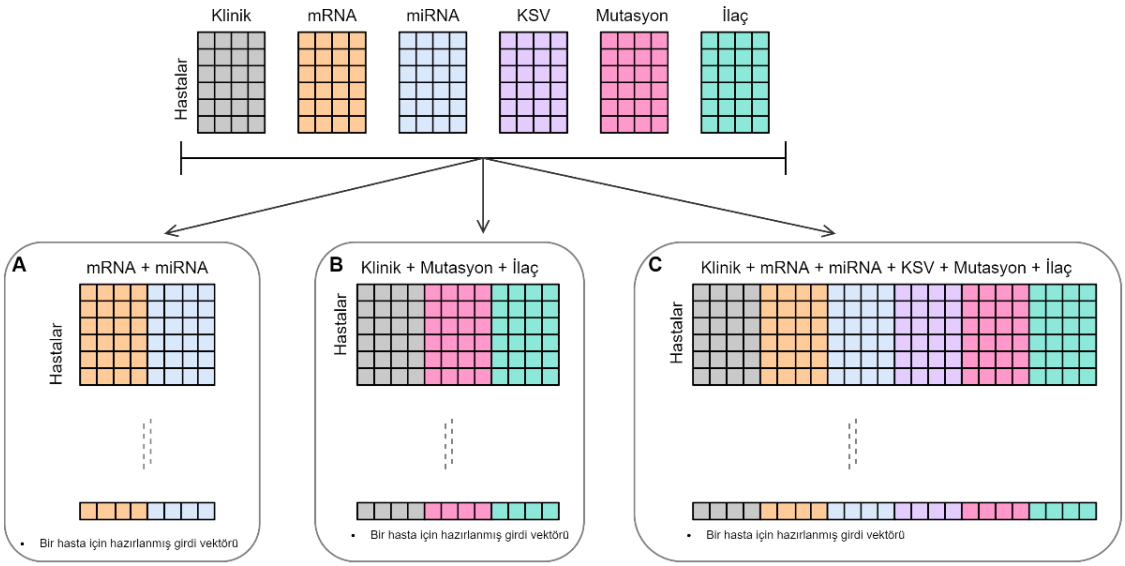
3.2. Model İçin Girdi ve Çıktı Verisinin Hazırlanması

Bu çalışmada rastgele orman algoritması ile hastaların sağkalımları tahmin edilmiştir. Bir sınıflandırma algoritması olan rastgele orman verilen girdi ve çıktı veri kombinasyonları ile eğitilmektedir. Algoritma girdi ile çıktı arasındaki ilişkiyi matematiksel olarak hesaplamaktadır. Öğrendiği bu ilişkiler ile yeni gördüğü girdi verisinin çıktısını tahmin etmektedir. Modellerimizi eğitirken ve test ederken veri tabanından elde ettiğimiz omik, klinik ve ilaç veri tipleri girdi verisi, sağkalım değerlerini çıktı verisi olarak kullanılmıştır.

Her doku için 29 farklı veri kombinasyonu ile model eğitilmiştir. Eğitilen modeller için girdi verisi model için kullanılmak istenen veri türü veya türlerine göre hazırlanmıştır. 6 farklı verinin model performanslarına etkisi araştırılmıştır. Modellere farklı veri türü kombinasyonları hazırlanırken tek veri türü, iki veri türü, üç ve daha fazla veri türü ile farklı modeller eğitilmiştir. Bu şekilde veri türlerinin birlikte ve tek olarak nasıl bir performans göstereceği araştırılmıştır. Model için kullanılan veri aynı hasta sırasına göre hazırlanmıştır. Farklı veri türleri bir araya getirilirken de aynı hasta sıraları dikkate alınmıştır. Her hastanın farklı veri türleri yan yana gelecek şekilde birleştirildi ve model için hazır hale getirilmiştir. Farklı veri türlerinin bir araya getirilmesi Şekil 3.4.'de farklı örneklerle gösterilmiştir.

Elde edilen model girdisinde her satır bir hastanın girdi vektörünü oluşturmaktadır. Şekil 3.4.'de A, B ve C durumunda farklı modellerde kullanılmak

amacıyla hazırlanan model girdisi örnekleri temsili olarak gösterilmiştir. Şekil 3.4' deki A durumunda iki omik veri tipi kullanılarak hazırlanan model girdisi örneği gösterilmiştir. Düzenlemeler sonucu seçilen genlere göre gen ifadesi ve miRNA ifadesi verisiyle iki veri türü birleştirilerek girdi hazırlanmıştır. Önce gen ifadesi bilgileri arkasından miRNA ifadesi bilgisi gelmektedir. Veride her sütun bir gen ile ilgili bilgi mevcuttur. Her satır bir hastayı temsil etmekte ve ilgili dokudaki tüm hastaların bilgileri veride bulunmaktadır. Hazırlanan model girdisinde her satır 2 omik veri tipi bilgisi olam bir hastanın girdi vektörünü temsil etmektedir.



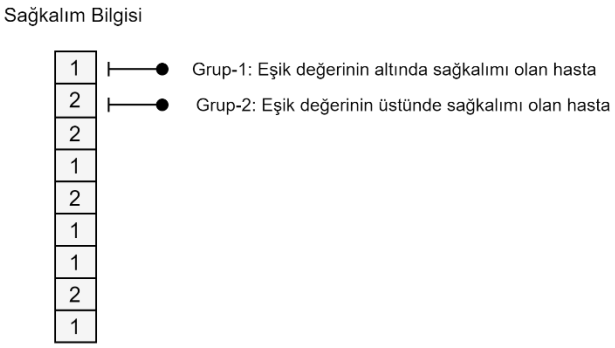
Şekil 3.4. Model girdisi için hazırlanan farklı veri kombinasyonları.

Şekil 3.4.'teki B durumunda 3 farklı veri türünün birleştirilmesi ile hazırlanan model girdi örneği gösterilmiştir. Düzenlenmiş klinik, mutasyon ve ilaç bilgisi kullanarak girdi hazırlanmıştır. Veride her sütunda farklı bir bilgi bulunmaktadır. Klinik bilgiler, belirlenen genler için mutasyon bilgileri arkasından hastaların ilaç bilgileri gelmektedir. Her satır bir hastayı temsil etmekte ve ilgili dokudaki tüm hastaların bilgileri veride bulunmaktadır. Hazırlanan model girdisinde her satır bu 3 veri bilgisini barındıran bir hastanın girdi vektörünü temsil etmektedir.

Şekil 3.4.'teki C durumunda veri tabanından elde edilen tüm veri türlerini kullanılarak hazırlanan model girdi örneği gösterilmiştir. 6 farklı veri türünün kullanılmıştır. Klinik, gen ifadesi, miRNA ifadesi, kopya sayısı varyasyonları,

mutasyon ve ilaç bilgisi verisi düzenlemeler sonunda algoritmada girdi verisi olarak kullanılabilecek hale getirilmiştir. Bu düzenlenmiş veri türleri yan yana gelecek şekilde birleştirilmiştir. Elde edilen yeni veri tüm veri türlerini kapsamaktadır. Veride her sütunda farklı bir bilgi bulunmaktadır. Her satır bir hastayı temsil etmekte ve ilgili dokudaki tüm hastaların bilgileri veride bulunmaktadır. Hazırlanan model girdisinde her satır tüm veri bilgisini barındıran bir hastanın girdi vektörünü temsil etmektedir. Verilen örneklerde olduğu gibi tüm modellerde girdi verisi aynı şekilde düzenlenip model için hazırlanmıştır.

Model çıktısı için ikili bir sınıflandırma yapılmıştır. Tahmin edilen sağkalım için her dokuda belirlenen eşik değerine göre iki grup belirlenmiştir. Her doku için hazırlanan sağkalım bilgisi vektörü modellerimizin çıktı vektörüdür. Hazırlanan çıktı vektörünün yapısı Şekil 3.5.'te gösterilmektedir. "1" ve "0" değerlerinden oluşturulan vektör iki farklı grubu temsil etmektedir. Eşik değerinin altında sağkalıma sahip hastalar grup 1'i, üstünde kalan hastalar grup 2'yi oluşturmaktadır. Girdi verisindeki hasta sırası kullanılmıştır ve her satır bir hastanın çıktı vektörüdür.



Şekil 3.5. Model için hazırlanan çıktı vektörü.

3.3. Model Eğitimi

Rastgele orman algoritmasında modelin testi ve eğitimi için veri seti ikiye bölünmüştür. Verinin %80'i eğitim, %20'si test verisi olarak değerlendirilmiştir. Tek çıkışlı çapraz doğrulayıcı (*Leave One Out Cross Validation, LOOCV)* kullanılarak çapraz doğrulama yapılmıştır. Çapraz doğrulama (*Cross-validation*), makine öğrenmesi modelinin görmediği veri üzerindeki performansını mümkün olan en doğru

şekilde değerlendirmek için kullanılan istatistiksel bir yeniden örnekleme yöntemidir. Tek çıkışlı çapraz doğrulayıcı, kat sayısının veri kümesindeki örnek sayısına eşit olduğu özel bir çapraz doğrulama durumudur. Bu yöntemde, öğrenme algoritması her örnek için bir kez uygulanır, diğer tüm örnekler bir eğitim seti olarak ve seçilen örnek tek bir test seti olarak kullanılmaktadır (92). Bu şekilde rastgele orman modelinin görmediği veri üzerindeki performansı daha objektif ve doğru bir şekilde değerlendirilebilir.

3.4. Öznitelik Seçimi

Öznitelik modeli oluştururken kullandığımız verideki her bir değişkendir. Öznitelik seçimi, tüm özniteliklerden daha iyi model performansı gösteren küçük boyutlu bir öznitelik grubunun seçildiği süreçtir. Öznitelik seçimi, yüksek boyutlu veri kümeleri için daha da önemlidir. Öznitelik sayısının çok yüksek olduğu bir veri olabilir. Bu durumda değişkenlerden hangisinin ilgili hangisinin alakasız olduğunu kolayca söylemek zorlaşır ve tüm değişkenleri dikkate alan bir model oluşturmak ve yorumlamak zordur. Bu nedenlerle özellik seçimi önemli bir görevdir (93). Yüksek boyutlu veri genellikle sınıflandırma için tahmine dayalı modellerin etkinliğini azaltan çok sayıda gereksiz ve alakasız bilgi içerebilir. Verimli ve etkili tahmin modelleri oluşturmak için, ayırt edici özniteliklere sahip alt kümeleri kullanmak iyi bir seçenektir (94).

COT modeli için öznitelik seçiminde, doğrusal bir regresyonda parametreleri tahmin eden LASSO yöntemi kullanılmıştır (95). Daha önceki çalışmalarda, COT modeli için LASSO yöntemi değerlendirilmiş ve daha az değişkenli, daha yorumlanabilir modeller verdiği belirtilmiştir (96).

Rastgele orman algoritmasında, öznitelik seçimi yapılarak eğitilen modellerde özyinelemeli öznitelik eliminasyonu (*Recursive Feature Elimination)* yöntemi kullanılmıştır. Seçilen modele uygun olacak şekilde, belirlenen öznitelik sayısına ulaşılana kadar en zayıf nitelendirdiği özniteliklerin eleyen bir yöntemidir. Hangi özniteliklerin hedef sınıfın tahmin edilmesine önemli ölçüde katkıda bulunduğunu belirlemek için sınıflandırıcı doğruluğunu kullanır. Popüler bir öznitelik seçim yöntemi olan özyinelemeli öznitelik eliminasyon algoritması, kullanım kolaylığı ve

eğitim veri setinde hedef değişkeni tahmin etmede daha çok alakalı olan öznitelikleri seçmede etkili olduğu için yaygındır. Yaklaşım, sınıflandırma ve regresyon modelleme problemlerinde öznitelik seçimi için kullanılabilir.

3.5. Performans Metrikleri

3.5.1. Yazılım Dili ve Kullanılabilirliği

Bu tez çalışmasında, Python programlama dili kullanılarak veri düzenleme ve ön işleme, model eğitimleri ve uygulamalar gerçekleştirilmiştir. Model performansları ile ilgili çeşitli hesaplamalar ve karşılaştırmalar yapılmıştır. Python dili, genel kullanım amacıyla yaygın olarak kullanılan açık kaynak kodlu, ücretsiz bir programlama dilidir. Python, bilimsel hesaplama için en popüler dillerden biri olarak, yüksek düzeyde etkileşimli doğası ve olgunlaşan bilimsel kütüphane ekosistemi sayesinde, algoritmik geliştirme ve veri analizi için önemli bir dildir (97). Birçok farklı alanda yazılım geliştirmeye müsait yapısıyla, hesaplamalı çalışmalarda da çok yaygın olarak kullanılmaktadır. Bu amaçla birçok kütüphaneye sahiptir ve bunlardan *Scikit-learn* kütüphanesi çalışmada kullanılmıştır. *Scikit-learn,* denetimli ve denetimsiz problemler için çeşitli son teknoloji makine öğrenme algoritmalarını bulunduran bir Python modülüdür (97). Veri bilimi ve makine öğrenmesi çalışmalarında kullanılan çok yönlü bir kütüphanedir. *Matplotlib* kütüphanesi, verinin ve sonuçların görselleştirilmesi için kullanılmıştır. Sayısal hesaplamaları 2 ve 3 boyutlu görsel çıktılar olarak almamızı sağlayan bir görselleştirme kütüphanesi olarak kullanılmaktadır. Uygulaması kolay ve basit olduğu için çok kullanılan kütüphanelerdendir. *Pandas* kütüphanesi, veri işleme ve analizi amaçlı olarak Python dilinde yazılmış bir başka kütüphanedir. Serileri, tabloları işlemek amacı ile verinin yapısını oluşturur, düzenler ve yapıyı işleyerek birçok analiz yapılmasına olanak sağlar. Ayrıca, Python'da veri setinin ön işleme aşamasında kullanılan birçok dahili kütüphane bulunmaktadır. Sınıflandırıcı model ve COT model, *Jupyter Notebook IDE* ortamında Python programlama dili kullanılarak geliştirilmiştir.

3.5.2. Uyumluluk Endeksi

Uyum endeksi veya C-endeksi bir algoritma tarafından yapılan tahminleri değerlendirmek için bir metriktir. Harrell C istatistiğine dayanmaktadır (98). Uyumlu çiftlerin toplam olası değerlendirme çifti sayısına bölümü olarak tanımlanmaktadır. Bir dizi olay zamanı ile tahmin edilen bir puan arasında C-endeksi hesaplanmaktadır. Uyum indeksi 0 ile 1 arasında bir değer almaktadır. 0,5 rastgele tahminleri, 1.0 mükemmel uyumu ve 0.0 mükemmel uyumsuzluğu göstermektedir. COT modelinin sonuçlarını değerlendirmek için C-endeks kullanılmaktadır.

3.5.3. Doğruluk, Kesinlik, Duyarlılık ve F1-Skor

Karışıklık matrisi (*confusion matrix*) makine öğrenmesi sınıflandırma problemi için bir performans ölçümü olarak, öngörülen ve gerçek değerlerin 4 farklı kombinasyonunu içeren bir tablodur (Şekil 3.6.).

		Tahmin değerleri	
		0	1
Gerçek değerler	0	Doğru negatif (DN)	Yanlış pozitif (YP)
	1	Yanlış negatif (YN)	Doğru pozitif (DP)

Şekil 3.6. Karışıklık matrisi.

Çalışmada, 4 performans metriği kullanılmıştır. Doğruluk (*accuracy*), doğru olarak sınıflandırılan örneklerin yüzdesidir. Aşağıdaki gibi tanımlanır:

$$Doğruluk = \frac{DP+DN}{DP+DN+YP+YN} \quad (3.1.)$$

Duyarlılık (r*ecall*), pozitif olarak tahmin edilmesi gereken işlemlerin ne kadarını pozitif olarak tahmin ettiğimizi gösteren bir metriktir. Aşağıdaki gibi tanımlanır:

$$Duyarlılık = \frac{DP}{DP+YN} \quad (3.2.)$$

Kesinlik (*precision*), pozitif olarak tahmin edilen değerlerin gerçekten kaçının pozitif olduğunu göstermektedir. Aşağıdaki gibi tanımlanır:

$$Kesinlik = \frac{DP}{DP+YP} \quad (3.3.)$$

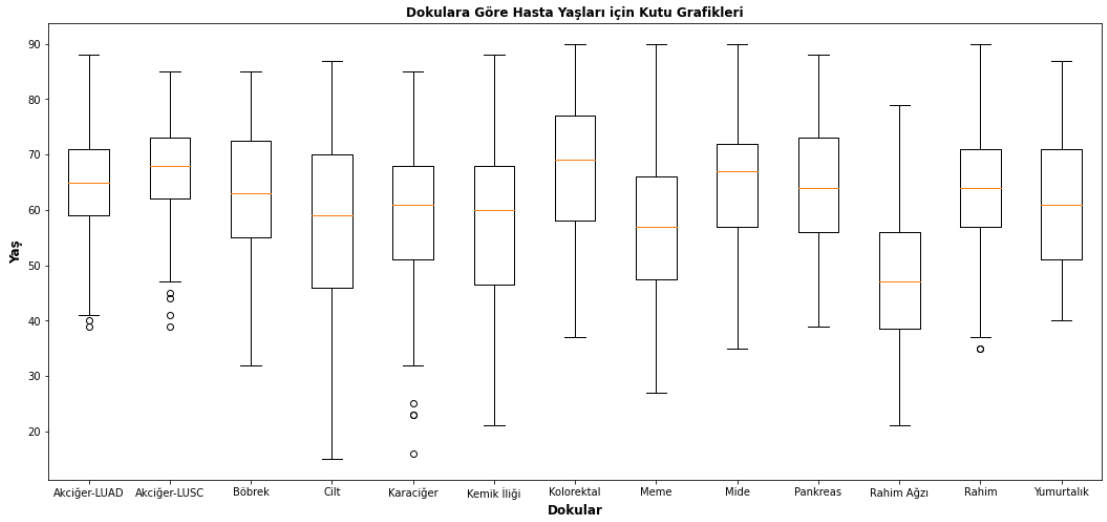
F1 skor (*f1-score*) ise bir testin doğruluğunun bir ölçüsüdür ve kesinlik ile duyarlılığın harmonik ortalamasıdır. Aşağıdaki gibi tanımlanır:

$$F1 - skor = \frac{2x\ Duyarlılık\ x\ Keskinlik}{Duyarlılık+Keskinlik} \quad (3.4.)$$

4. BULGULAR

4.1. Veri Araştırması

Farklı kanser türlerinde hastaların yaş dağılımları Şekil 4.1'de gösterilmiştir. Tüm kanser türlerinde yapılan çalışmalarda genç ve yetişkin hastalar değerlendirilmiştir. Çocuk hastalar çalışma dışı tutulmuştur. En geniş yaş aralığı cilt kanserinde görülürken, yaş ortalaması en yüksek olan kolorektal kanserdir.



Şekil 4.1. Dokulara göre hastalarda yaş dağılımı.

4.1.1. Etiket Bilgisinin Belirlenmesi

Tüm kanser türleri için tahmin edilecek sağkalım bilgisi etiket olarak düzenlenmiştir. Hastaların hayatta kalma süreleri klinik bilgiye göre belirlenmiştir ve model ile tahmini yapılacak sınıflar oluşturulmuştur. Bunun için her dokuda hastaların dahil olacağı 2 grup belirlenmiştir. 2'den fazla grup belirlenmeme sebebi grup başına hasta sayısının sınıflandırma için yetersiz kalacak olmasıdır. Benzer sağkalım süresine sahip hastaların bir kısmı çalışma sonunda ölmüş bir kısmı hayatta kalmıştır. Bu hastaların aynı gruba düşmesi yanıltıcı sonuçlara neden olabileceği için her kanserde eşik değerleri belirlenmiştir. Sağkalım sürelerine göre grupları belirlemek için her dokuda bir eşik değeri belirlenmiştir (Bkz. Şekil 1.1). Bu yapılan düzenleme ile araştırmamızı makine öğrenmesi için ikili bir sınıflandırma problemine dönüşmüştür.

Hastaların takip süreleri ve ölüm yılları baz alınarak sağkalım süreleri düzenlenmiştir. Belirlenen eşik değer yıl olarak hesaplanmıştır. Farklı eşik değerlerinde hayatta kalan ve ölen hastaların dağılımı araştırılmıştır. Kullanılacak eşik değeri belirlenirken gruplar arasında dengeli bir dağılım olmasına dikkat edilmiştir. Veri setinde dengesiz bir dağılım olduğunda daha az sayıda olan hasta grubu için yanlış sınıflandırılma sorunu ortaya çıkabilirdi. Örnek olarak ölen hastaların fazla olduğu bir veride yaşadığı bilinen hastaların da öldü olarak tahmin edilmesi gösterilebilir. Bu durum yanıltıcı sonuçlara neden olabilir. Belirlenen süreden az yaşayanlar birinci grubu, çok yaşayanlar ise ikinci grubu oluşturmaktadır. Her doku için seçilen eşik değeri ve seçilen hasta sayıları Tablo 4.1'de gösterilmiştir.

Hastaların sağkalım süreleri baz alınarak etiketlenen hastaların verisi ile rastgele orman modelleri eğitilmiştir. Sınıflandırma sonucunda hastaların sağkalımı tahmin edilmiştir.

Tablo 4.1. Dokularda eşik değerine göre hasta sayısı.

Doku	**Eşik Değeri**	**Düzenleme öncesi hasta sayısı**	**Düzenleme sonrası hasta sayısı**
Akciğer 1 (LUAD)	2 yıl	457	222
Akciğer 2 (LUSC)	2 yıl	454	248
Böbrek (KIRP)	5 yıl	268	79
Cilt (SKCM)	3 yıl	238	125
Karaciğer (LIHC)	2 yıl	300	154
Kemik İliği (LAML)	6 ay	107	75
Kolorektal (COAD)	3 yıl	325	125
Meme (BRCA)	6 yıl	840	199
Mide (STAD)	1 yıl	364	246
Pankreas (PAAD)	1 yıl	166	109
Rahim Ağzı (CESC)	3 yıl	279	123
Rahim (UCEC)	5 yıl	507	174
Yumurtalık (OV)	2 yıl	267	113
Tüm Dokular	3 yıl	5292	2402

Seçilen eşik değerini belirlemeye örnek olarak, meme kanseri için kullanılan BRCA veri setinde yapılan düzenlemeler verilmiştir. Eşik değerine göre hasta sayısı değişimini özetleyen bilgiler Tablo 4.2.'de gösterilmiştir. Veri setinde yaşayan ve ölen

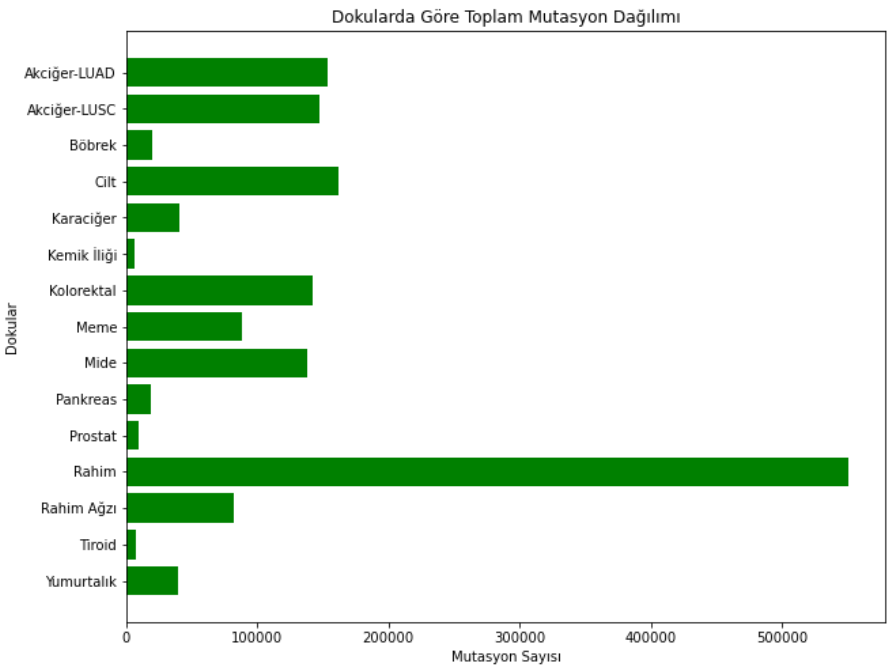
hasta sayılarının daha dengeli olduğu “6 yıl” eşik değeri olarak kullanılmıştır. Diğer kanser türlerinde de aynı yaklaşımla eşik değeri belirlenmiştir.

Tablo 4.2. BRCA veri setinde eşik değeri seçimi.

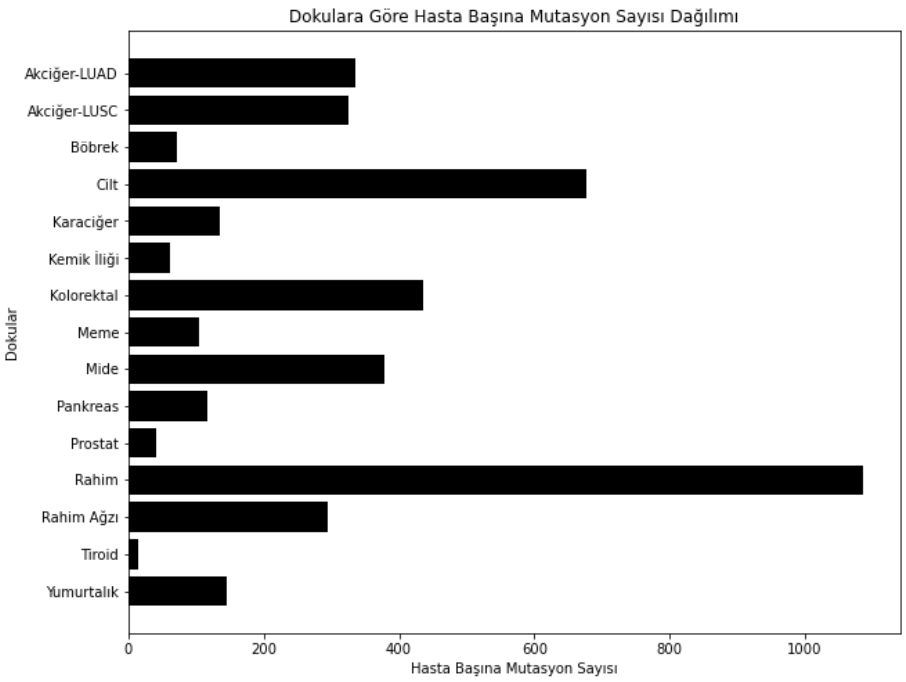
Eşik Değeri (Yıl)	Ölen hasta sayısı	Yaşayan hasta sayısı	Toplam hasta sayısı
1	110 (18)	730 (609)	840 (627)
2	110 (33)	730 (377)	840 (410)
3	110 (49)	730 (285)	840 (334)
4	110 (59)	730 (212)	840)271)
5	110 (70)	730 (160)	840 (230)
*6	110 (77)	730 (122)	840 (199)

4.1.2. Mutasyon Verisinde Araştırma

Mutasyon bilgisi bulunan gen sayısı her dokunun veri setinde farklı bulunmuştur. Doku bazında toplam mutasyon sayıları Şekil 4.2.’de, hasta başına mutasyon dağılımları Şekil 4.3.‘te gösterilmektedir. Mutasyon bilgisi olan gen sayısı en çok rahim dokusunda bulunmuştur.



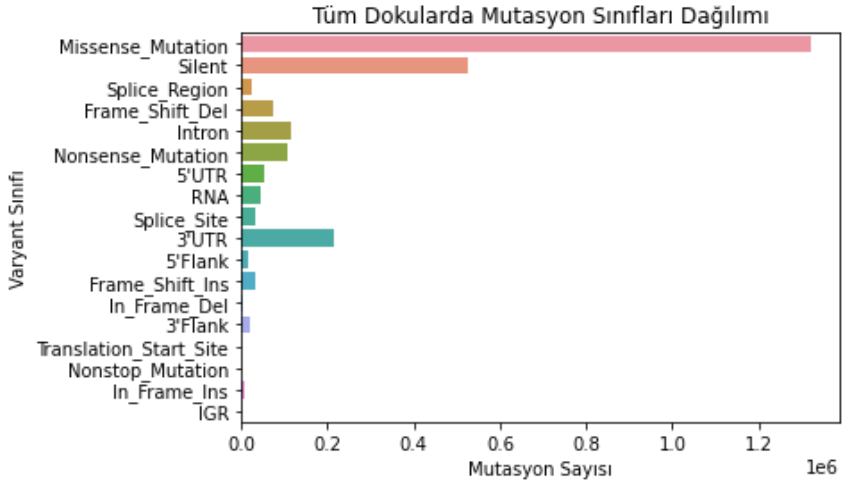
Şekil 4.2. Dokulara göre toplam mutasyon sayısı dağılımları.



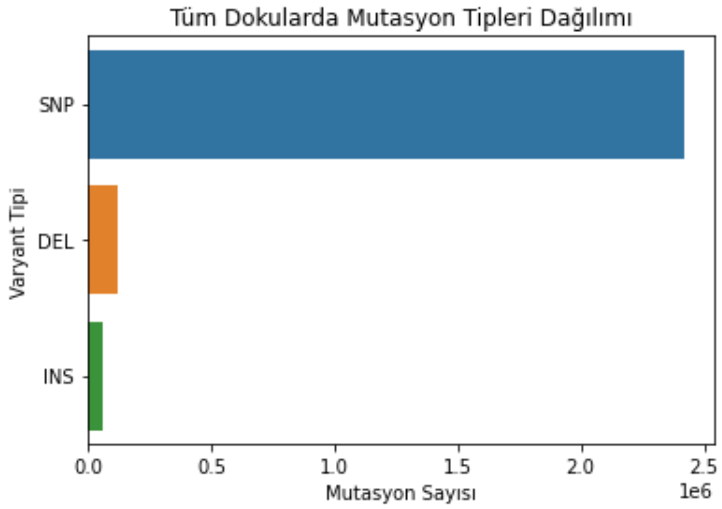
Şekil 4.3. Dokulara göre hasta başına mutasyon sayısı dağılımları.

Mutasyon için farklı varyant sınıflarının dokulara göre dağılımları Şekil 4.4.'te gösterilmektedir. Mutasyonlarda varyant tipleri içinde en çok sayıda yanlış anlamlı (*missense*) mutasyon olduğu bulunmuştur. Tüm dokularda yanlış anlamlı mutasyonların sonra en çok olan varyant sınıfı sessiz (*silent*) mutasyonlar olmuştur.

Mutasyon verisinde üç tip mutasyon varyantı gösterilmiştir. Tüm kanser türlerinin mutasyon verisine bakıldığında, en çok mutasyon varyant tipinin tek nükleotid polimorfizm (SNP) olduğu belirlenmiştir. Şekil 4.5.'te tüm dokulardaki varyant tiplerinin dağılımı gösterilmektedir.

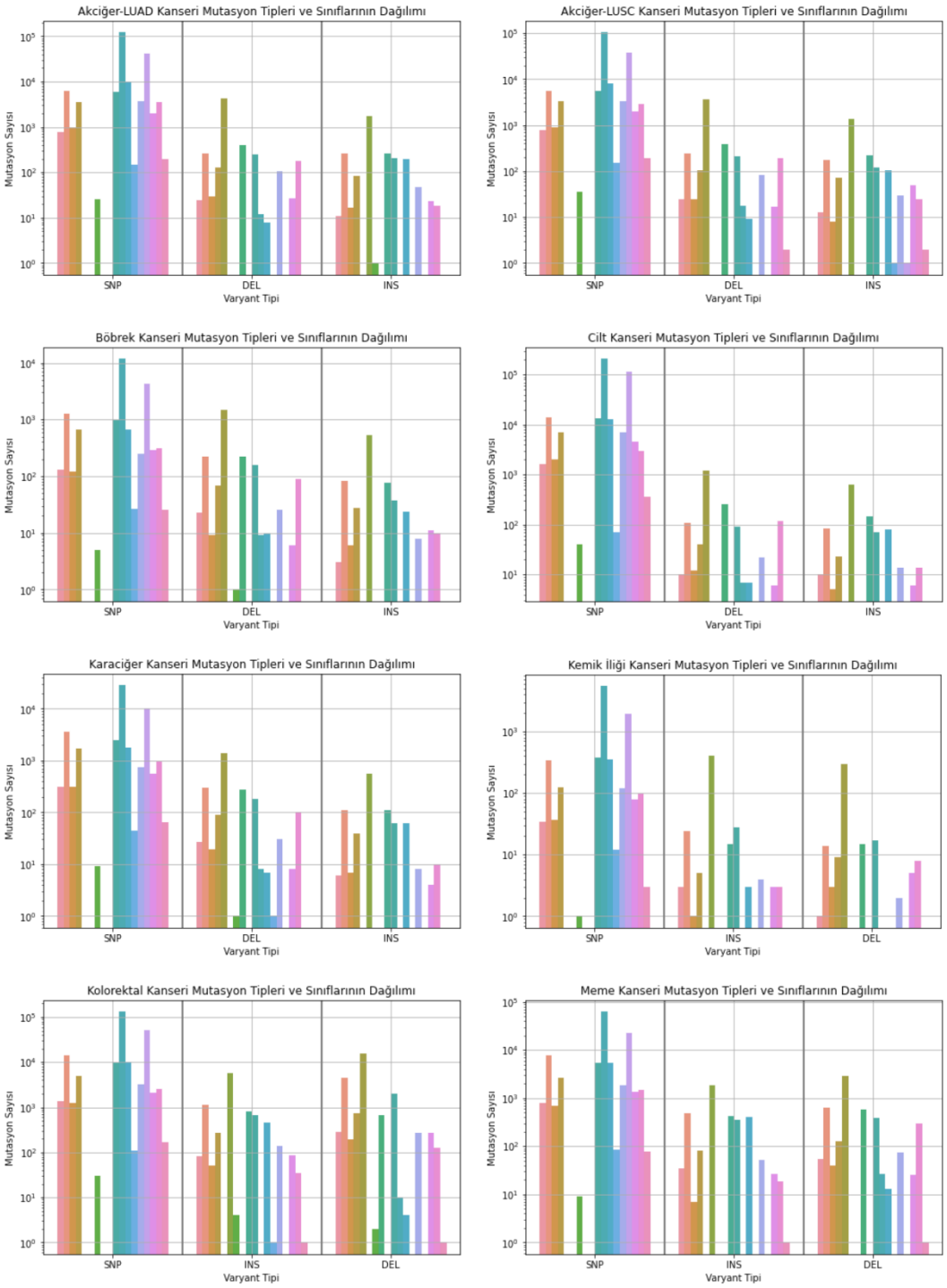


Şekil 4.4. Tüm kanser tiplerine ait hasta verisinde mutasyon sınıfları dağılımı.

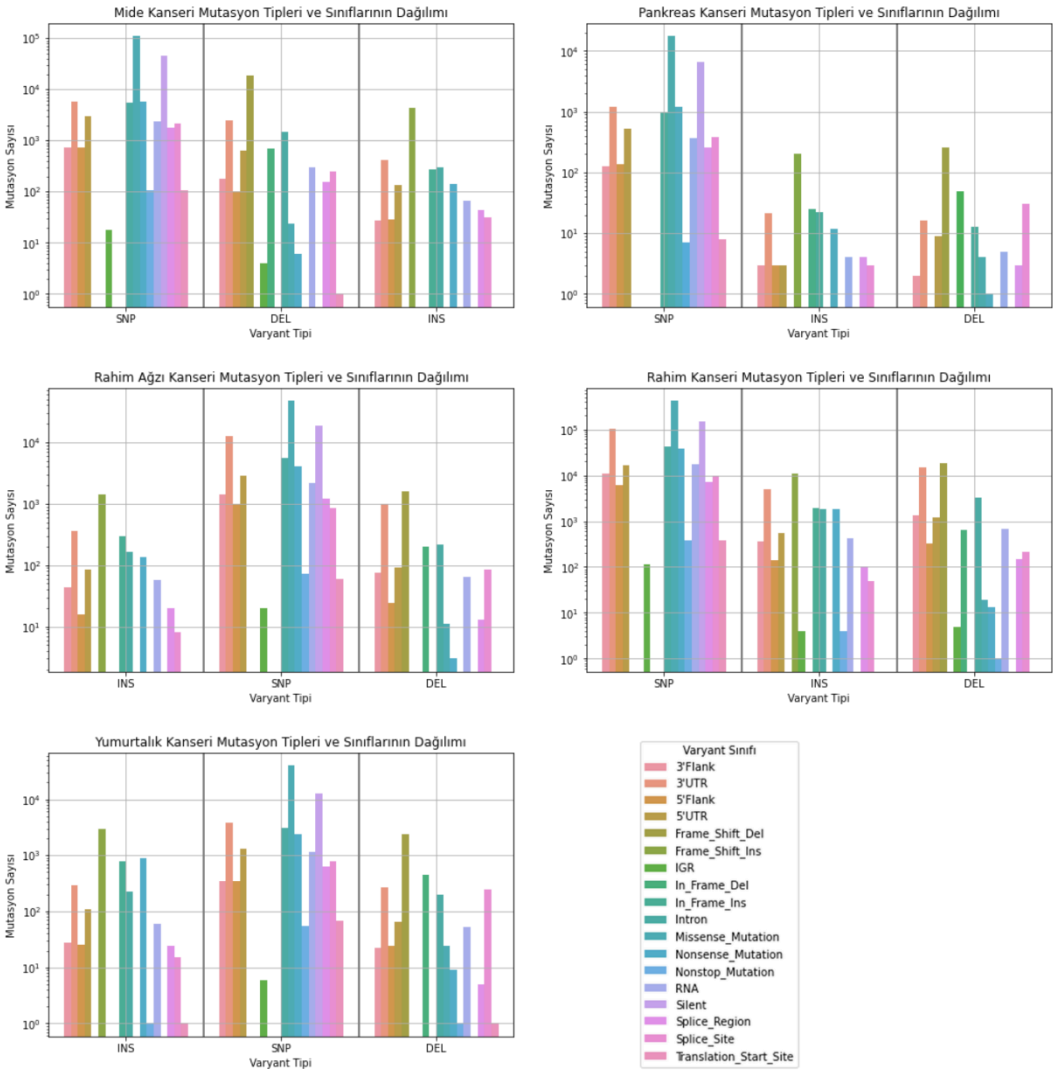


Şekil 4.5. Tüm kanser tiplerine ait hasta verisinde mutasyon tipleri dağılımı.

Farklı kanser türleri için mutasyon tipleri ve sınıflarının dağılımları Şekil 4.6.'da gösterilmiştir. Tüm dokularda en çok olan mutasyon tipi SNP tipidir. Tüm dokularda *missense* mutasyon en çok olan mutasyon varyantıdır. Yanlış anlamlı mutasyon olarak bilinen bu mutasyon gen ürünü protein üzerinde değişiklik yapar. Missense mutasyondan sonra en çok olan varyant *silent* mutasyondur. *Silent* mutasyon sessiz mutasyon olarak bilinir ve gen ürünü olan protein yapısında ve aktivitesinde değişiklik yapmaz.



Şekil 4.6. Dokular bazında mutasyon tipleri ve sınıflarının dağılımları.



Şekil 4.6. Dokular bazında mutasyon tipleri ve sınıflarının dağılımları**.** (Devam)

4.1.3. L1000 Gen Listesi ile Omik Veride Düzenleme

L1000 genleri kullanılarak mutasyon, kopya sayısı varyasyonu ve gen ifadesi verisinde düzenleme yapıldı. Büyük boyutlu verinin verimli şekilde daha az boyut kazanması sağlanmıştır. 978 genden bazıları için veri yoktur. Eksik olan bu genler veri düzenlenirken çıkarılmıştır. Omik veri tipileri için belirlenen gen sayıları Tablo 4.3.'te gösterilmiştir.

Tablo 4.3. Kanser türlerine göre üç omik veri tipinde son gen sayıları.

Kanser Türü	Gen İfadesi	KSV	Mutasyon
Akciğer 1 (LUAD)	934 (60483)	934 (60043)	908 (18424)
Akciğer 2 (LUSC)	934 (60483)	934 (60002)	904 (18553)
Böbrek	934 (60483)	934 (60053)	577 (10535)
Cilt	934 (60483)	934 (60043)	905 (18387)
Karaciğer	934 (60483)	934 (60043)	709 (14609)
Kemik İliği	934 (60483)	934 (60053)	282 (5087)
Kolorektal	934 (60483)	934 (60053)	932 (19185)
Meme	934 (60483)	934 (60043)	899 (18158)
Mide	934 (60483)	934 (60053)	926 (18665)
Pankreas	934 (60483)	934 (60053)	669 (12677)
Rahim Ağzı	934 (60483)	934 (60053)	913 (18773)
Rahim	934 (60483)	934 (60053)	963 (21197)
Yumurtalık	934 (60483)	934 (60002)	747 (14014)
Tüm Kanserler	934 (60483)	934 (60002)	963 (21775)

KSV: Kopya Sayısı Varyasyonu

4.1.4. Çapraz Doğrulama ile Model Eğitimi

Algoritma, ilk olarak eğitim seti olarak belirlenen veri ile eğitilir. Daha sonra test seti eğitimde öğrenilen bilgi test seti olarak ayrılan veriyle model test edilir. Modellerde birden çok eğitim yapıldığında performanslarda farklı doğruluk oranları elde edildi. Bu da veri seti üzerinden tek bir eğitim ile değerlendirmenin yanıltıcı olabileceğini göstermiştir. Daha isabetli makine öğrenmesi modelleri geliştirilmesi ve bunların test veri setlerinde daha yüksek performans göstermeleri için çapraz doğrulama yapılmıştır. Tüm modellerde *LOOCV* ile elde edilen sonuçlar dikkate alınmıştır. *LOOCV* ile eğitim veri setindeki her örnek için bir model eğitilmiştir. Bu şekilde model optimize edildi ve model performansının artması sağlanmıştır.

4.1.4. Cox Orantılı Tehlike Modelinde Kullanılan Veride Düzenleme

Model performansını karşılaştırmak için klasik bir sağkalım analiz yöntemi olan Cox orantılı tehlike (COT) yöntemi kullanılmıştır. COT modeli için düzenlenen veride sağkalım süresinde bir eşik değeri seçilmediği için hasta sayılarında değişme olmuştur. Her kanser türü ve tüm kanserlerin birleştirildiği veride seçilen hasta sayıları Tablo 4.4.'te verilmiştir.

Tablo 4.4. COT modeli için seçilen hasta sayıları.

Kanser Türü	Hasta Sayısı
Akciğer 1 (LUAD)	448
Akciğer 2 (LUSC)	448
Böbrek	267
Cilt	234
Karaciğer	298
Kemik İliği	102
Kolorektal	323
Meme	840
Mide	359
Pancreas	166
Rahim Ağzı	279
Rahim	506
Yumurtalık	266
Tüm Kanserler	5255

COT modeli için omik veri, L1000 gen listesine göre düzenlenmiş şekli ile ele alındı. Mutasyon, gen ifadesi ve KSV için ilgili genler seçilmiştir. COT modelinde, verinin çok boyutlu olmasından dolayı fazla uyma sıkıntısı yaşandı. Bu sorunun önüne geçmek için veride öznitelik seçimi yapılmıştır. Tablo 4.5.'te COT modeli için her veri tipinde LASSO ile seçilen öznitelik sayıları verilmiştir.

Tablo 4.5. COT modeli için LASSO ile öznitelik seçilimi sonrası kanser türleri için veri tiplerine göre öznitelik sayıları.

Kanser Türü	Gen İfadesi miRNA KSV Mutasyon Klinik İlaç	Gen İfadesi	KSV	Mutasyon	miRNA	Klinik	İlaç
Akciğer 1 (LUAD)	667 (4743)	224 (934)	150 (934)	249 (908)	518 (1881)	70	16
Akciğer 2 (LUSC)	736 (4744)	268 (934)	198 (934)	271 (904)	607 (1881)	68	23
Böbrek	493 (4377)	154 (934)	91 (934)	190 (577)	197 (1881)	36	15
Cilt	473 (4748)	175 (934)	163 (934)	305 (905)	208 (1881)	66	27
Karaciğer	382 (4515)	168 (934)	108 (934)	205 (709)	199 (1881)	42	15
Kemik İliği	99 (4041)	60 (934)	36 (934)	66 (282)	66 (1881)	10	-
Kolorektal	571 (4752)	186 (934)	98 (934)	179 (932)	459 (1881)	56	15
Meme	1120 (4763)	364 (934)	207 (934)	414 (899)	477 (1881)	75	40
Mide	213 (4749)	183 (934)	110 (934)	158 (926)	203 (1881)	52	23
Pankreas	166 (4467)	91 (934)	77 (934)	116 (669)	113 (1881)	36	13
Rahim Ağzı	615 (4719)	507 (934)	144 (934)	413 (913)	471 (1881)	42	15
Rahim	590 (4752)	254 (934)	112 (934)	228 (963)	301 (1881)	20	21
Yumurtalık	586 (4536)	438 (934)	158 (934)	220 (747)	420 (1881)	11	31
Tüm Kanserler	996 (5051)	287 (934)	125 (934)	383 (963)	559 (1881)	212	129

KSV: Kopya Sayısı Varyasyonu

4.2. Farklı Kanser Türlerinde Model Performansı Sonuçları

Farklı kanser türleri için ayrı modelleme yapılmıştır. Her kanser türünde eğitilen tüm modellerin sonuçları aşağıda tablolarla verilmiştir. 4 farklı omik veri tipi, klinik ve ilaç verisi kullanılarak farklı veri kombinasyonları bir araya getirilip modeller eğitilmiştir. Tüm kanser türlerinde 29 farklı model eğitilmiştir. Bu şekilde omiklerin tekli ve çoklu kombinasyonları ile modeller eğitilmiştir. Çoklu omik verinin sağkalım tahmini üzerine etkisi araştırılmıştır. Klinik ve ilaç veri tiplerinin omik veri tipleriyle nasıl bir performans gösterdiği araştırılmıştır. Tüm modeller aynı parametreler ile eğitilmiştir. Rastgele orman algoritması için ağaç sayısı (*n_estimators*) "100" olarak belirlenmiştir. Her modelde test verisi büyüklüğü tüm verinin %20'sini oluşturmuştur. Tüm modellerde, test ve eğitim veri kümelerini seçerken rastgele durum (*random_state*) değeri 42 olarak seçilmiştir.

4.2.1. Akciğer Kanseri

Akciğer kanseri için iki en yaygın alt tür kullanılmıştır. Skuamöz hücreli karsinom ve adenokarsinom, küçük hücreli olmayan akciğer kanserinin iki ana histolojik tipidir (99). Ayrıca iki histolojik tip arasında genel sağkalımda önemli farklılıklar gözlendiği gösterilmiştir (99). Bu bilgilerden yola çıkarak bu iki çalışmanın ayrı modellerle değerlendirilmesine karar verilmiştir. İki modelin sonuç performansları ayrı tablolarda verilmiştir.

LUAD projesi

Tablo 4.6.'de akciğer kanseri LUAD verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Adenokarsinom alt türü için kullanılan LUAD veri seti ile eğitilen model sonuçlarına bakıldığında en yüksek performans, mutasyon, klinik ve ilaç bilgisinin kullanıldığı modelde gözlenmiştir (doğruluk = 0,68). Bu üç veri kombinasyonun yanına gen ifadesi verisi de eklendiğinde performans kısmi bir azalma yaşasa da diğer modellerden daha yüksek bir sonuç elde etmiştir. KSV ve mutasyon verisinin olduğu modellerde daha düşük performans gözlenmiştir. 3 omik veri tipinin kullanıldığı modele kıyasla omik veri yanına klinik bilginin eklendiği modelde performans daha yüksektir. Omik verinin klinik ve ilaç verisi ile kullanılması sonucu sağkalım tahmin performansının daha yüksek olduğu gözlenmiştir.

Tablo 4.6. Akciğer (LUAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,59	0,60	0,59	0,60
Klinik	0,64	0,64	0,64	0,64
Gen ifadesi	0,62	0,63	0,63	0,63
KSV	0,53	0,54	0,53	0,54
Mutasyon	0,48	0,51	0,49	0,51
miRNA	0,52	0,54	0,52	0,54
İlaç	0,50	0,55	0,54	0,55
miRNA + KSV + Mutasyon + Klinik + İlaç	0,54	0,56	0,55	0,56
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,60	0,61	0,61	0,61
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,56	0,58	0,57	0,58
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,56	0,58	0,57	0,58
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,56	0,58	0,57	0,58
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,59	0,59	0,59	0,59
Gen ifadesi + Mutasyon	0,61	0,62	0,61	0,62
Gen ifadesi + KSV	0,57	0,58	0,57	0,58
Mutasyon + KSV	0,50	0,51	0,50	0,51
Gen ifadesi + İlaç	0,63	0,64	0,64	0,64
Mutasyon + İlaç	0,54	0,55	0,54	0,55
Gen ifadesi + miRNA	0,58	0,59	0,59	0,59
Gen ifadesi + Mutasyon + KSV	0,60	0,61	0,60	0,61
Gen ifadesi + Mutasyon + İlaç	0,61	0,62	0,62	0,62
Mutasyon + KSV + İlaç	0,51	0,52	0,51	0,52
Gen ifadesi + İlaç + Klinik	0,59	0,59	0,59	0,59
Mutasyon + İlaç + Klinik	0,67	0,68	0,67	0,68
İlaç + Klinik + miRNA	0,55	0,56	0,55	0,56
KSV + İlaç + Klinik	0,56	0,57	0,56	0,57
Mutasyon + İlaç + Klinik + Gen ifadesi	0,64	0,65	0,65	0,65
Mutasyon + İlaç + Klinik + KSV	0,56	0,57	0,56	0,57
Mutasyon + İlaç + Klinik + miRNA	0,55	0,57	0,56	0,57

KSV: Kopya Sayısı Varyasyonu

LUSC projesi

Tablo 4.7.'de akciğer kanseri LUSC verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Skuamöz hücreli karsinom alt türü için kullanılan LUSC veri seti ile eğitilen model sonuçlarına bakıldığında en yüksek performans miRNA, KSV, mutasyon, klinik ve ilaç bilgisinin kullanıldığı modelde gözlenmiştir (doğruluk = 0,62). Sağkalım tahmininde ilaç bilgisinin model performansını yükselttiği gözlenmiştir. 3 omik veri tipinin kullanıldığı model diğer modellerden düşük performansa sahiptir.

Tablo 4.7. Akciğer (LUSC veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	**F1-skor**	**Doğruluk**	**Kesinlik**	**Duyarlılık**
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,57	0,57	0,57	0,57
Klinik	0,58	0,58	0,58	0,58
Gen ifadesi	0,48	0,49	0,48	0,49
KSV	0,55	0,55	0,55	0,55
Mutasyon	0,49	0,49	0,49	0,49
miRNA	0,55	0,56	0,56	0,56
İlaç	0,48	0,52	0,56	0,52
miRNA + KSV + Mutasyon + Klinik + İlaç	0,61	0,62	0,62	0,62
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,54	0,55	0,54	0,55
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,55	0,56	0,55	0,56
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,59	0,59	0,59	0,59
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,58	0,59	0,59	0,59
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,52	0,52	0,52	0,52
Gen ifadesi + Mutasyon	0,49	0,50	0,49	0,50
Gen ifadesi + KSV	0,53	0,54	0,53	0,54
Mutasyon + KSV	0,55	0,55	0,55	0,55
Gen ifadesi + İlaç	0,52	0,53	0,52	0,53
Mutasyon + İlaç	0,53	0,54	0,53	0,54
Gen ifadesi + miRNA	0,54	0,55	0,55	0,55
Gen ifadesi + Mutasyon + KSV	0,52	0,53	0,53	0,53
Gen ifadesi + Mutasyon + İlaç	0,50	0,51	0,50	0,51
Mutasyon + KSV + İlaç	0,57	0,57	0,57	0,57

Tablo 4.7. Akciğer (LUSC veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + İlaç + Klinik	0,46	0,47	0,46	0,47
Mutasyon + İlaç + Klinik	0,60	0,61	0,61	0,61
İlaç + Klinik + miRNA	0,57	0,58	0,58	0,58
KSV + İlaç + Klinik	0,56	0,56	0,56	0,56
Mutasyon + İlaç + Klinik + Gen ifadesi	0,50	0,52	0,51	0,52
Mutasyon + İlaç + Klinik + KSV	0,56	0,56	0,56	0,56
Mutasyon + İlaç + Klinik + miRNA	0,57	0,58	0,58	0,58

KSV: Kopya Sayısı Varyasyonu

4.2.2. Böbrek Kanser

Tablo 4.8.'de böbrek kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Böbrek kanseri için eğitilen modellerden en yüksek performansı gen ifadesi, miRNA, mutasyon, klinik ve ilaç verisinin kullanıldığı model göstermiştir (doğruluk = 0,80). Benzer bir performansa sahip diğer model iki omik veri tipi gen ifadesi ve miRNA verisinin birlikte kullanıldığı modeldir (doğruluk = 0,80). Gen ifadesi ve KSV verisine ek olarak ilaç ve klinik verinin yüksek bir model performansı elde ettiği gösterilmiştir (doğruluk = 0,78). Sadece mutasyon bilgisinin kullanıldığı model en düşük model performansına sahiptir (doğruluk = 0,32). Gen ifadesinin olduğu modellerde daha yüksek performans gözlenmiştir.

Tablo 4.8. Böbrek (KIRP veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,74	0,75	0,75	0,75
Klinik	0,65	0,66	0,66	0,66
Gen ifadesi	0,73	0,73	0,74	0,73
KSV	0,60	0,61	0,60	0,61
Mutasyon	0,32	0,32	0,32	0,32
miRNA	0,73	0,73	0,74	0,73
İlaç	0,59	0,65	0,71	0,65
miRNA + KSV + Mutasyon + Klinik + İlaç	0,74	0,75	0,76	0,75
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,79	0,80	0,81	0,80
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,76	0,77	0,78	0,77

Tablo 4.8. Böbrek (KIRP veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,76	0,77	0,78	0,77
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,74	0,75	0,76	0,75
Gen ifadesi + Mutasyon	0,70	0,71	0,71	0,71
Gen ifadesi + KSV	0,73	0,73	0,74	0,73
Mutasyon + KSV	0,66	0,68	0,70	0,68
Gen ifadesi + İlaç	0,75	0,76	0,77	0,76
Mutasyon + İlaç	0,48	0,48	0,49	0,48
Gen ifadesi + miRNA	0,79	0,80	0,82	0,80
Gen ifadesi + Mutasyon + KSV	0,75	0,76	0,77	0,76
Gen ifadesi + Mutasyon + İlaç	0,73	0,73	0,74	0,73
Mutasyon + KSV + İlaç	0,67	0,68	0,69	0,68
Gen ifadesi + İlaç + Klinik	0,76	0,77	0,78	0,77
Mutasyon + İlaç + Klinik	0,69	0,70	0,70	0,70
İlaç + Klinik + miRNA	0,67	0,67	0,67	0,67
KSV + İlaç + Klinik	0,64	0,66	0,66	0,66
Mutasyon + İlaç + Klinik + Gen ifadesi	0,74	0,75	0,76	0,75
Mutasyon + İlaç + Klinik + KSV	0,78	0,78	0,80	0,78
Mutasyon + İlaç + Klinik + miRNA	0,75	0,76	0,77	0,76

KSV: Kopya Sayısı Varyasyonu

4.2.3. Cilt Kanseri

Tablo 4.9.'da cilt kanseri veri ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Cilt kanseri için eğitilen modellerden en yüksek performansı gen ifadesi, klinik ve ilaç verisinin kullanıldığı model vermiştir (doğruluk = 0,66). Klinik verinin tek kullanıldığı model de benzer bir performans göstermiştir (doğruluk = 0,66). Tüm veri tiplerinin kullanılarak eğitildiği modelin performansı, KSV, miRNA ve Mutasyon verisi kullanılarak eğitildiği tek omik modellere kıyasla daha yüksektir (doğruluk = 0,61).

Tablo 4.9. Cilt (SKCM veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,60	0,61	0,60	0,61
Klinik	0,66	0,66	0,67	0,66
Gen ifadesi	0,62	0,62	0,62	0,62
KSV	0,50	0,53	0,51	0,53
Mutasyon	0,48	0,48	0,48	0,48
miRNA	0,56	0,58	0,57	0,58
İlaç	0,42	0,52	0,45	0,52
miRNA + KSV + Mutasyon + Klinik + İlaç	0,54	0,58	0,57	0,58
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,64	0,64	0,64	0,64
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,59	0,61	0,61	0,61
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,60	0,61	0,60	0,61
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,64	0,65	0,65	0,65
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,57	0,58	0,57	0,58
Gen ifadesi + Mutasyon	0,57	0,58	0,57	0,58
Gen ifadesi + KSV	0,62	0,63	0,63	0,63
Mutasyon + KSV	0,50	0,53	0,51	0,53
Gen ifadesi + İlaç	0,61	0,62	0,61	0,62
Mutasyon + İlaç	0,48	0,50	0,48	0,50
Gen ifadesi + miRNA	0,61	0,62	0,61	0,62
Gen ifadesi + Mutasyon + KSV	0,62	0,63	0,63	0,63
Gen ifadesi + Mutasyon + İlaç	0,62	0,62	0,62	0,62
Mutasyon + KSV + İlaç	0,46	0,50	0,47	0,50
Gen ifadesi + İlaç + Klinik	0,65	0,66	0,66	0,66
Mutasyon + İlaç + Klinik	0,54	0,56	0,55	0,56
İlaç + Klinik + miRNA	0,60	0,61	0,60	0,61
KSV + İlaç + Klinik	0,51	0,53	0,51	0,53
Mutasyon + İlaç + Klinik + Gen ifadesi	0,60	0,60	0,60	0,60
Mutasyon + İlaç + Klinik + KSV	0,62	0,62	0,62	0,62
Mutasyon + İlaç + Klinik + miRNA	0,61	0,62	0,62	0,62

KSV: Kopya Sayısı Varyasyonu

4.2.4. Karaciğer Kanseri

Tablo 4.10.'da karaciğer kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Karaciğer kanserinde gen ifadesi verisi ile eğitilen modellerde daha yüksek model performansı gözlenmiştir. Gen ifadesi-KSV, gen

ifadesi-ilaç, gen ifadesi-miRNA ve gen ifadesi-ilaç-klinik modellerinin performansları benzer çıkmıştır (doğruluk = 0,68). Gen ifadesinin tek başına da gösterdiği performans (doğruluk = 0,65) diğer veri tipleri ile yapılan modellere göre daha düşüktür olduğu gözlenmiştir.

Tablo 4.10. Karaciğer (LIHC veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	**F1-skor**	**Doğruluk**	**Kesinlik**	**Duyarlılık**
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,64	0,64	0,64	0,64
Klinik	0,53	0,53	0,53	0,53
Gen ifadesi	0,65	0,65	0,65	0,65
KSV	0,57	0,57	0,57	0,57
Mutasyon	0,58	0,60	0,60	0,60
miRNA	0,64	0,65	0,65	0,65
İlaç	0,42	0,55	0,49	0,55
miRNA + KSV + Mutasyon + Klinik + İlaç	0,60	0,62	0,61	0,62
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,62	0,63	0,63	0,63
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,68	0,68	0,68	0,68
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,67	0,68	0,67	0,68
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,67	0,67	0,67	0,67
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,66	0,67	0,67	0,67
Gen ifadesi + Mutasyon	0,63	0,63	0,63	0,63
Gen ifadesi + KSV	0,68	0,68	0,68	0,68
Mutasyon + KSV	0,61	0,62	0,62	0,62
Gen ifadesi + İlaç	0,68	0,68	0,68	0,68
Mutasyon + İlaç	0,56	0,59	0,59	0,59
Gen ifadesi + miRNA	0,68	0,68	0,68	0,68
Gen ifadesi + Mutasyon + KSV	0,65	0,66	0,65	0,66
Gen ifadesi + Mutasyon + İlaç	0,66	0,66	0,66	0,66
Mutasyon + KSV + İlaç	0,58	0,59	0,58	0,59
Gen ifadesi + İlaç + Klinik	0,68	0,68	0,68	0,68
Mutasyon + İlaç + Klinik	0,64	0,66	0,66	0,66
İlaç + Klinik + miRNA	0,60	0,62	0,61	0,62
KSV + İlaç + Klinik	0,57	0,58	0,57	0,58
Mutasyon + İlaç + Klinik + Gen ifadesi	0,68	0,68	0,68	0,68
Mutasyon + İlaç + Klinik + KSV	0,65	0,65	0,65	0,65
Mutasyon + İlaç + Klinik + miRNA	0,65	0,65	0,65	0,65

KSV: Kopya Sayısı Varyasyonu

4.2.5. Kemik İliği Kanseri

Tablo 4.11.'de kemik iliği verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Kemik iliği kanseri diğer kanser türleri içinde en yüksek performansı gösteren modellerden biridir. Gen ifadesi, mutasyon, miRNA ve klinik veri ile eğitilen model en yüksek performansı göstermiştir (doğruluk = 0,80). Gen ifadesi, mutasyon ve miRNA modelinin performansı (doğruluk = 0,76), klinik verinin eklenmesi ile artmıştır. Üç omik veri tipi ile eğitilen modelin tek ve ikili omik veri tipi ile eğitilen modellerden daha yüksek performans gösterdiği gözlenmiştir. Tüm veri tiplerinin kullanılarak eğitildiği model diğer modellere göre daha yüksek sonuç vermiştir (doğruluk = 0,68). Bu model diğer kanserlerden farklı olarak gen ifadesi, mutasyon, KSV, miRNA ve klinik bilgiler kullanılarak eğitilmiştir. Çünkü kemik iliği kanseri için ilaç bilgisi yoktur. Bu yüzden bu kanserde omik verinin ilaç bilgisi ile değerlendirilmesi yapılamamıştır.

Tablo 4.11. Kemik iliği (LAML veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	**F1-skor**	**Doğruluk**	**Kesinlik**	**Duyarlılık**
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,75	0,76	0,76	0,76
Klinik	0,56	0,56	0,56	0,56
Gen ifadesi	0,75	0,76	0,76	0,76
KSV	0,51	0,52	0,51	0,52
Mutasyon	0,45	0,52	0,42	0,52
miRNA	0,66	0,69	0,70	0,69
miRNA + KSV + Mutasyon + Klinik	0,62	0,67	0,67	0,67
Gen ifadesi + miRNA + Mutasyon + Klinik	0,68	0,69	0,69	0,69
Gen ifadesi + miRNA + KSV + Klinik	0,73	0,75	0,75	0,75
Gen ifadesi + miRNA + KSV + Mutasyon	0,69	0,71	0,70	0,71
Gen ifadesi + Mutasyon	0,68	0,71	0,71	0,71
Gen ifadesi + KSV	0,66	0,68	0,67	0,68
Mutasyon + KSV	0,54	0,53	0,54	0,53
Mutasyon + miRNA	0,71	0,72	0,72	0,72
KSV + miRNA	0,73	0,75	0,76	0,75
Gen ifadesi + miRNA	0,71	0,73	0,74	0,73
Gen ifadesi + Mutasyon + KSV	0,67	0,69	0,69	0,69
Gen ifadesi + Mutasyon + Klinik	0,65	0,67	0,66	0,67

Tablo 4.11. Kemik iliği (LAML veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Mutasyon + KSV + Klinik	0,55	0,56	0,55	0,56
Gen ifadesi + KSV + Klinik	0,71	0,72	0,72	0,72
Gen ifadesi + Mutasyon + miRNA	0,75	0,76	0,76	0,76
Gen ifadesi + KSV + miRNA	0,67	0,69	0,69	0,69
Mutasyon + KSV + miRNA	0,71	0,73	0,75	0,73
Klinik + miRNA + Gen ifadesi	0,72	0,73	0,73	0,73
Klinik + miRNA + KSV	0,63	0,68	0,69	0,68
Klinik + miRNA + Mutasyon	0,69	0,72	0,74	0,72
Gen ifadesi + Mutasyon + Klinik + KSV	0,74	0,75	0,74	0,75
Gen ifadesi + Mutasyon + Klinik + miRNA	0,80	0,80	0,80	0,80

KSV: Kopya Sayısı Varyasyonu

4.2.6. Kolorektal Kanser

Tablo 4.12.'de kolorektal kanser verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Kolorektal kanserde en yüksek sonuç, klinik verinin kullanılarak eğitildiği modelde gözlenmiştir (doğruluk = 0,70). Mutasyon, klinik ve ilaç bilgisinin kullanıldığı modelin üç veri ile yapılan en yüksek performansa sahip olan model olduğu gözlenmiştir (doğruluk = 0,63). Tüm veri tiplerinin bir arada değerlendirildiği veriyle eğitildiği model gen ifadesi, KSV, mutasyon ve miRNA verisi ile eğitildiği tek omik veri kullanılan modellerden daha yüksek bir performans gözlenmiştir (doğruluk = 0,60).

Tablo 4.12. Kolorektal (COAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,60	0,60	0,60	0,60
Klinik	0,70	0,70	0,70	0,70
Gen ifadesi	0,56	0,56	0,56	0,56
KSV	0,53	0,53	0,53	0,53
Mutasyon	0,46	0,46	0,46	0,46
miRNA	0,54	0,54	0,54	0,54
İlaç	0,39	0,44	0,41	0,44

Tablo 4.12. Kolorektal (COAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
miRNA + KSV + Mutasyon + Klinik + İlaç	0,59	0,59	0,59	0,59
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,60	0,60	0,60	0,60
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,54	0,54	0,54	0,54
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,51	0,51	0,51	0,51
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,54	0,54	0,54	0,54
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,52	0,52	0,52	0,52
Gen ifadesi + Mutasyon	0,57	0,58	0,58	0,58
Gen ifadesi + KSV	0,50	0,50	0,50	0,50
Mutasyon + KSV	0,51	0,51	0,51	0,51
Gen ifadesi + İlaç	0,60	0,60	0,60	0,60
Mutasyon + İlaç	0,46	0,47	0,47	0,47
Gen ifadesi + miRNA	0,61	0,61	0,61	0,61
Gen ifadesi + Mutasyon + KSV	0,54	0,54	0,54	0,54
Gen ifadesi + Mutasyon + İlaç	0,49	0,49	0,49	0,49
Mutasyon + KSV + İlaç	0,50	0,50	0,50	0,50
Gen ifadesi + İlaç + Klinik	0,58	0,58	0,58	0,58
Mutasyon + İlaç + Klinik	0,63	0,63	0,64	0,63
İlaç + Klinik + miRNA	0,58	0,58	0,58	0,58
KSV + İlaç + Klinik	0,56	0,56	0,56	0,56
Mutasyon + İlaç + Klinik + Gen ifadesi	0,49	0,49	0,49	0,49
Mutasyon + İlaç + Klinik + KSV	0,56	0,56	0,56	0,56
Mutasyon + İlaç + Klinik + miRNA	0,57	0,57	0,57	0,57

KSV: Kopya Sayısı Varyasyonu

4.2.7. Meme Kanseri

Tablo 4.13.'te meme kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Meme kanseri, diğer kanser türleri içinde en yüksek performansı gösteren modellerden biridir. Mutasyon, ilaç ve klinik veri kullanarak eğitilen model en yüksek performansı gözlenmiştir (doğruluk = 0,83). Mutasyon ve ilaç bilgisinin olduğu model performansı (doğruluk = 0,73), klinik bilginin eklenmesi ile artmıştır. Meme kanserinde omik veri içinde mutasyon verisi ile eğitilen modeller daha yüksek sonuç vermiştir. Üç omik veri tipi; mutasyon, gen ifadesi ve KSV

kullanılarak eğitilen model diğer tek omik veri tipiyle eğitilen modellerden daha yüksek performans göstermiştir (doğruluk = 0,65).

Tablo 4.13. Meme (BRCA veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	**F1-skor**	**Doğruluk**	**Kesinlik**	**Duyarlılık**
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,60	0,64	0,63	0,64
Klinik	0,82	0,83	0,83	0,83
Gen ifadesi	0,63	0,64	0,63	0,64
KSV	0,57	0,59	0,57	0,59
Mutasyon	0,54	0,58	0,54	0,58
miRNA	0,56	0,60	0,57	0,60
İlaç	0,71	0,71	0,72	0,71
miRNA + KSV + Mutasyon + Klinik + İlaç	0,64	0,67	0,67	0,67
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,68	0,71	0,73	0,71
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,66	0,69	0,69	0,69
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,58	0,63	0,61	0,63
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,62	0,65	0,64	0,65
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,65	0,67	0,67	0,67
Gen ifadesi + Mutasyon	0,67	0,69	0,69	0,69
Gen ifadesi + KSV	0,64	0,66	0,65	0,66
Mutasyon + KSV	0,58	0,60	0,58	0,60
Gen ifadesi + İlaç	0,64	0,66	0,65	0,66
Mutasyon + İlaç	0,72	0,73	0,72	0,73
Gen ifadesi + miRNA	0,62	0,65	0,63	0,65
Gen ifadesi + Mutasyon + KSV	0,62	0,65	0,63	0,65
Gen ifadesi + Mutasyon + İlaç	0,67	0,68	0,68	0,68
Mutasyon + KSV + İlaç	0,62	0,64	0,63	0,64
Gen ifadesi + İlaç + Klinik	0,69	0,70	0,70	0,70
Mutasyon + İlaç + Klinik	0,82	0,83	0,84	0,83
İlaç + Klinik + miRNA	0,67	0,70	0,72	0,70
KSV + İlaç + Klinik	0,68	0,69	0,69	0,69
Mutasyon + İlaç + Klinik + Gen ifadesi	0,64	0,67	0,67	0,67
Mutasyon + İlaç + Klinik + KSV	0,70	0,71	0,71	0,71
Mutasyon + İlaç + Klinik + miRNA	0,63	0,67	0,68	0,67

KSV: Kopya Sayısı Varyasyonu

4.2.8. Mide Kanseri

Tablo 4.14.'te mide kanseri veri ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Mide kanserinde en yüksek model performans, gen ifadesi, mutasyon, miRNA, ilaç ve klinik veri kullanarak eğitilen modelde gözlenmiştir (doğruluk = 0,66). Tüm veri tipleri ile eğitilen model gen ifadesi dışında tek omik veri tipi ile eğitilen modellerden daha yüksek sonuç vermiştir (doğruluk = 0,62).

Tablo 4.14. Mide (STAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	**F1-skor**	**Doğruluk**	**Kesinlik**	**Duyarlılık**
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,57	0,62	0,59	0,62
Klinik	0,58	0,59	0,58	0,59
Gen ifadesi	0,60	0,64	0,62	0,64
KSV	0,57	0,59	0,57	0,59
Mutasyon	0,48	0,48	0,47	0,48
miRNA	0,55	0,61	0,57	0,61
İlaç	0,52	0,61	0,56	0,61
miRNA + KSV + Mutasyon + Klinik + İlaç	0,59	0,64	0,62	0,64
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,63	0,66	0,65	0,66
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,58	0,62	0,60	0,62
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,58	0,63	0,61	0,63
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,57	0,61	0,59	0,61
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,55	0,60	0,57	0,60
Gen ifadesi + Mutasyon	0,56	0,61	0,58	0,61
Gen ifadesi + KSV	0,55	0,59	0,55	0,59
Mutasyon + KSV	0,57	0,60	0,57	0,60
Gen ifadesi + İlaç	0,58	0,61	0,59	0,61
Mutasyon + İlaç	0,56	0,56	0,56	0,56
Gen ifadesi + miRNA	0,54	0,60	0,56	0,60
Gen ifadesi + Mutasyon + KSV	0,57	0,61	0,58	0,61
Gen ifadesi + Mutasyon + İlaç	0,60	0,63	0,61	0,63
Mutasyon + KSV + İlaç	0,56	0,59	0,56	0,59
Gen ifadesi + İlaç + Klinik	0,59	0,64	0,63	0,64

Tablo 4.14. Mide (STAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Mutasyon + İlaç + Klinik	0,59	0,61	0,59	0,61
İlaç + Klinik + miRNA	0,54	0,59	0,55	0,59
KSV + İlaç + Klinik	0,56	0,59	0,56	0,59
Mutasyon + İlaç + Klinik + Gen ifadesi	0,57	0,62	0,59	0,62
Mutasyon + İlaç + Klinik + KSV	0,61	0,64	0,62	0,64
Mutasyon + İlaç + Klinik + miRNA	0,54	0,59	0,56	0,59

KSV: Kopya Sayısı Varyasyonu

4.2.9. Pankreas Kanseri

Tablo 4.15.'te pankreas kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Pankreas kanserinde sadece ilaç verisiyle eğitilen modelde en yüksek performans gözlenmiştir (doğruluk = 0,62). Mutasyon, klinik ve ilaç verisi ile eğitilen modelde de benzer sonuçlar gözlenmiştir (doğruluk = 0,61). Mutasyon, miRNA, klinik ve ilaç verisi ile eğitilen model (doğruluk = 0,58), tek omik veri tipi ile eğitilen miRNA (doğruluk = 0,56), mutasyon (doğruluk = 0,50) ve KSV (doğruluk = 0,44) modellerinden dahi yüksek performans gözlenmiştir.

Tablo 4.15. Pankreas (PAAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,50	0,50	0,50	0,50
Klinik	0,55	0,55	0,55	0,55
Gen ifadesi	0,58	0,58	0,58	0,58
KSV	0,44	0,44	0,44	0,44
Mutasyon	0,51	0,50	0,51	0,50
miRNA	0,56	0,56	0,56	0,56
İlaç	0,62	0,62	0,62	0,62
miRNA + KSV + Mutasyon + Klinik + İlaç	0,51	0,51	0,51	0,51
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,54	0,54	0,54	0,54
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,54	0,54	0,54	0,54
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,49	0,50	0,49	0,50

Tablo 4.15. Pankreas (PAAD veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,52	0,52	0,52	0,52
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,48	0,49	0,48	0,49
Gen ifadesi + Mutasyon	0,54	0,54	0,54	0,54
Gen ifadesi + KSV	0,53	0,53	0,53	0,53
Mutasyon + KSV	0,44	0,44	0,44	0,44
Gen ifadesi + İlaç	0,48	0,48	0,48	0,48
Mutasyon + İlaç	0,52	0,54	0,54	0,54
Gen ifadesi + miRNA	0,56	0,56	0,56	0,56
Gen ifadesi + Mutasyon + KSV	0,51	0,51	0,51	0,51
Gen ifadesi + Mutasyon + İlaç	0,51	0,51	0,52	0,51
Mutasyon + KSV + İlaç	0,49	0,49	0,49	0,49
Gen ifadesi + İlaç + Klinik	0,53	0,53	0,53	0,53
Mutasyon + İlaç + Klinik	0,62	0,61	0,62	0,61
İlaç + Klinik + miRNA	0,58	0,58	0,58	0,58
KSV + İlaç + Klinik	0,49	0,49	0,49	0,49
Mutasyon + İlaç + Klinik + Gen ifadesi	0,52	0,52	0,52	0,52
Mutasyon + İlaç + Klinik + KSV	0,49	0,49	0,49	0,49
Mutasyon + İlaç + Klinik + miRNA	0,57	0,58	0,58	0,58

KSV: Kopya Sayısı Varyasyonu

4.2.10. Rahim Ağzı Kanseri

Tablo 4.16.'da rahim ağzı kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Rahim ağzı kanserinde en yüksek model performansı gen ifadesi, kopya sayısı varyasyonu, miRNA, klinik ve ilaç verisi ile eğitilen model göstermiştir (doğruluk = 0,68). Tüm veri tipleri ile eğitilen model (doğruluk = 0,67), tek omik veri tipi ile eğitilen mutasyon (doğruluk = 0,48), KSV (doğruluk = 0,54), gen ifadesi (doğruluk = 0,59) ve miRNA (doğruluk = 0,6) modellerinden daha yüksek performans göstermiştir. Tüm veri ile eğitilen model ve tüm veriden gen ifadesinin çıkarıldığı veriyle eğitilen model performansı eşit çıkmıştır.

Tablo 4.16. Rahim ağzı (CESC veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,66	0,67	0,66	0,67
Klinik	0,58	0,59	0,58	0,59
Gen ifadesi	0,57	0,59	0,58	0,59
KSV	0,54	0,54	0,54	0,54
Mutasyon	0,44	0,48	0,54	0,48
miRNA	0,59	0,60	0,59	0,60
İlaç	0,58	0,64	0,70	0,64
miRNA + KSV + Mutasyon + Klinik + İlaç	0,65	0,67	0,67	0,67
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,64	0,65	0,65	0,65
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,67	0,68	0,69	0,68
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,62	0,63	0,63	0,63
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,63	0,64	0,64	0,64
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,62	0,63	0,63	0,63
Gen ifadesi + Mutasyon	0,59	0,60	0,59	0,60
Gen ifadesi + KSV	0,58	0,60	0,59	0,60
Mutasyon + KSV	0,55	0,55	0,55	0,55
Gen ifadesi + İlaç	0,56	0,57	0,56	0,57
Mutasyon + İlaç	0,51	0,51	0,54	0,51
Gen ifadesi + miRNA	0,56	0,58	0,57	0,58
Gen ifadesi + Mutasyon + KSV	0,60	0,61	0,60	0,61
Gen ifadesi + Mutasyon + İlaç	0,63	0,63	0,63	0,63
Mutasyon + KSV + İlaç	0,56	0,56	0,56	0,56
Gen ifadesi + İlaç + Klinik	0,61	0,63	0,62	0,63
Mutasyon + İlaç + Klinik	0,59	0,59	0,60	0,59
İlaç + Klinik + miRNA	0,59	0,60	0,59	0,60
KSV + İlaç + Klinik	0,56	0,57	0,56	0,57
Mutasyon + İlaç + Klinik + Gen ifadesi	0,60	0,62	0,61	0,62
Mutasyon + İlaç + Klinik + KSV	0,59	0,60	0,59	0,60
Mutasyon + İlaç + Klinik + miRNA	0,62	0,63	0,62	0,63

KSV: Kopya Sayısı Varyasyonu

4.2.11. Rahim Kanseri

Tablo 4.17.'de rahim kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Rahim kanserinde en yüksek model performansı mutasyon, miRNA, klinik ve ilaç verisiyle eğitilen model gözlenmiştir (doğruluk = 0,74). Mutasyon, klinik ve ilaç bilgisinin kullanıldığı modele miRNA eklenince model performansı yükselmiştir (doğruluk = 0,72). Tüm veri tipleri ile eğitilen modelin tek omik veri tipi ile eğitilen modellerden daha yüksek performans sahip olduğu gözlenmiştir (doğruluk = 0,67).

Tablo 4.17. Rahim (UCEC veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,67	0,67	0,67	0,67
Klinik	0,67	0,67	0,67	0,67
Gen ifadesi	0,66	0,66	0,66	0,66
KSV	0,60	0,60	0,60	0,60
Mutasyon	0,66	0,66	0,66	0,66
miRNA	0,64	0,65	0,65	0,65
İlaç	0,62	0,64	0,64	0,64
miRNA + KSV + Mutasyon + Klinik + İlaç	0,66	0,66	0,66	0,66
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,69	0,69	0,69	0,69
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,66	0,67	0,66	0,67
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,67	0,67	0,67	0,67
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,66	0,67	0,66	0,67
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,67	0,67	0,67	0,67
Gen ifadesi + Mutasyon	0,69	0,69	0,69	0,69
Gen ifadesi + KSV	0,65	0,65	0,65	0,65
Mutasyon + KSV	0,58	0,57	0,58	0,57
Gen ifadesi + İlaç	0,62	0,62	0,62	0,62
Mutasyon + İlaç	0,69	0,69	0,69	0,69
Gen ifadesi + miRNA	0,68	0,68	0,68	0,68
Gen ifadesi + Mutasyon + KSV	0,68	0,68	0,68	0,68
Gen ifadesi + Mutasyon + İlaç	0,69	0,69	0,69	0,69
Mutasyon + KSV + İlaç	0,62	0,61	0,62	0,61
Gen ifadesi + İlaç + Klinik	0,67	0,67	0,67	0,67
Mutasyon + İlaç + Klinik	0,72	0,72	0,72	0,72

Tablo 4.17. Rahim (UCEC veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
İlaç + Klinik + miRNA	0,67	0,67	0,67	0,67
KSV + İlaç + Klinik	0,58	0,58	0,59	0,58
Mutasyon + İlaç + Klinik + Gen ifadesi	0,67	0,67	0,67	0,67
Mutasyon + İlaç + Klinik + KSV	0,61	0,61	0,61	0,61
Mutasyon + İlaç + Klinik + miRNA	0,73	0,74	0,74	0,74

KSV: Kopya Sayısı Varyasyonu

4.2.12. Yumurtalık Kanseri

Tablo 4.18.'de yumurtalık kanseri verisi ile eğitilmiş tüm modellerin performans sonuçları gösterilmektedir. Yumurtalık kanserinde sadece ilaç veriyle eğitilen model performansının, tüm modeller arasında en yüksek olduğu gözlenmiştir (doğruluk = 0,73). Mutasyon, klinik ve ilaç veriyle eğitilen modelin (doğruluk = 0,68) gen ifadesi, klinik ve ilaç verisi ile eğitilen modelden (doğruluk = 0,66) daha yüksek performans olduğu gözlenmiştir. İlaç verisiyle eğitilen modellerin daha yüksek performansa sahip olduğu gözlenmiştir. Tüm veri tiplerinin kullanıldığı modelin, tek omik veri tipiyle eğitilen miRNA, KSV ve mutasyon modellerinden daha yüksek performansa sahip olduğu gözlenmiştir.

Tablo 4.18. Yumurtalık (OV veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri.

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,55	0,55	0,55	0,55
Klinik	0,45	0,45	0,45	0,45
Gen ifadesi	0,59	0,59	0,59	0,59
KSV	0,50	0,50	0,50	0,50
Mutasyon	0,46	0,49	0,49	0,49
miRNA	0,55	0,55	0,55	0,55
İlaç	0,73	0,73	0,74	0,73
miRNA + KSV + Mutasyon + Klinik + İlaç	0,49	0,50	0,49	0,50
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,52	0,52	0,52	0,52

Tablo 4.18. Yumurtalık (OV veri seti) kanser tipine ait hasta verisini içerecek şekilde oluşturulan model performans değerleri. (Devam)

Model	F1-skor	Doğruluk	Kesinlik	Duyarlılık
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,58	0,58	0,58	0,58
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,50	0,50	0,50	0,50
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,57	0,58	0,58	0,58
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,58	0,58	0,58	0,58
Gen ifadesi + Mutasyon	0,58	0,58	0,58	0,58
Gen ifadesi + KSV	0,57	0,57	0,57	0,57
Mutasyon + KSV	0,46	0,47	0,47	0,47
Gen ifadesi + İlaç	0,57	0,57	0,57	0,57
Mutasyon + İlaç	0,68	0,68	0,68	0,68
Gen ifadesi + miRNA	0,59	0,59	0,59	0,59
Gen ifadesi + Mutasyon + KSV	0,50	0,50	0,50	0,50
Gen ifadesi + Mutasyon + İlaç	0,59	0,59	0,59	0,59
Mutasyon + KSV + İlaç	0,49	0,50	0,49	0,50
Gen ifadesi + İlaç + Klinik	0,62	0,62	0,62	0,62
Mutasyon + İlaç + Klinik	0,66	0,66	0,66	0,66
İlaç + Klinik + miRNA	0,57	0,58	0,58	0,58
KSV + İlaç + Klinik	0,52	0,52	0,52	0,52
Mutasyon + İlaç + Klinik + Gen ifadesi	0,61	0,61	0,61	0,61
Mutasyon + İlaç + Klinik + KSV	0,48	0,49	0,49	0,49
Mutasyon + İlaç + Klinik + miRNA	0,57	0,58	0,58	0,58

KSV: Kopya Sayısı Varyasyonu

4.3. Tüm Kanserler İçin Oluşturulan Modelin Performans Sonuçları

Kanser türlerinin ayrı ayrı değerlendirilmesinin yanı sıra tüm kanser türlerinin bir araya getirilerek eğitildiği modellerde ele alınmıştır. Ayrı ayrı modellenen 13 kanser türüne ek olarak prostat ve tiroit kanseri verisi de bu modellerde kullanılmıştır. Tablo 4.19.'da tüm kanser türlerini içeren veri ile eğitilen tüm modellerin performans sonuçları gösterilmektedir. En yüksek performans mutasyon, klinik ve ilaç veriyle eğitilen modelde gözlenmiştir (doğruluk = 0,77). Tüm veri tipleriyle eğitilen model, diğer doku bazlı modellere kıyasla daha yüksek performans göstermiştir (doğruluk = 0,75). Gen ifadesi, mutasyon ve KSV verisiyle eğitilen modelden de benzer sonuçlar elde edilmiştir (doğruluk = 0,75). Tüm kanserlerin birleştirilmesiyle oluşan veride

mutasyon ve ilaç verisi en az bilgilendirici iken, gen ifadesi verisinin en bilgilendirici olduğu gözlenmiştir.

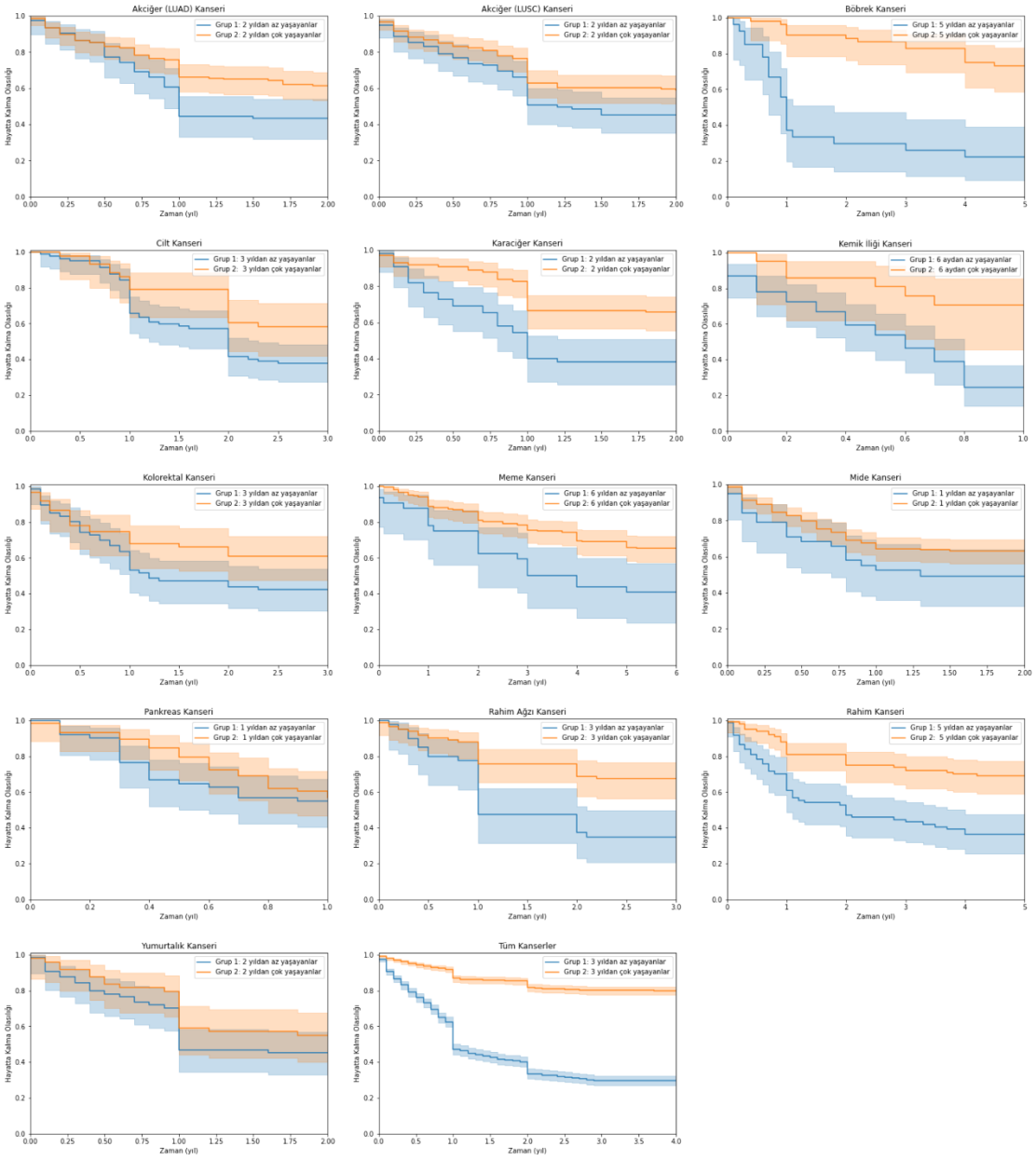
Tablo 4.19. Tüm kanser tiplerine ait hasta verisini içerecek biçimde oluşturulan sağkalım tahmin modelinin performans değerleri.

Model	**F1-skor**	**Doğruluk**	**Kesinlik**	**Duyarlılık**
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik + İlaç	0,76	0,75	0,76	0,75
Klinik	0,74	0,74	0,74	0,74
Gen ifadesi	0,75	0,75	0,75	0,75
KSV	0,64	0,64	0,65	0,64
Mutasyon	0,58	0,58	0,58	0,58
miRNA	0,72	0,72	0,72	0,72
İlaç	0,57	0,62	0,64	0,62
miRNA + KSV + Mutasyon + Klinik + İlaç	0,74	0,73	0,74	0,73
Gen ifadesi + miRNA + Mutasyon + Klinik + İlaç	0,75	0,75	0,75	0,75
Gen ifadesi + miRNA + KSV + Klinik + İlaç	0,74	0,74	0,75	0,74
Gen ifadesi + miRNA + KSV + Mutasyon + İlaç	0,74	0,74	0,75	0,74
Gen ifadesi + miRNA + KSV + Mutasyon + Klinik	0,74	0,74	0,75	0,74
Gen ifadesi + KSV + Mutasyon + Klinik + İlaç	0,75	0,75	0,76	0,75
Gen ifadesi + Mutasyon	0,74	0,74	0,75	0,74
Gen ifadesi + KSV	0,75	0,75	0,75	0,75
Mutasyon + KSV	0,65	0,64	0,65	0,64
Gen ifadesi + İlaç	0,75	0,75	0,75	0,75
Mutasyon + İlaç	0,64	0,64	0,64	0,64
Gen ifadesi + miRNA	0,74	0,74	0,74	0,74
Gen ifadesi + Mutasyon + KSV	0,75	0,75	0,75	0,75
Gen ifadesi + Mutasyon + İlaç	0,74	0,74	0,75	0,74
Mutasyon + KSV + İlaç	0,66	0,66	0,66	0,66
Gen ifadesi + İlaç + Klinik	0,75	0,75	0,75	0,75
Mutasyon + İlaç + Klinik	0,77	0,77	0,77	0,77
İlaç + Klinik + miRNA	0,73	0,73	0,73	0,73
KSV + İlaç + Klinik	0,73	0,73	0,73	0,73
Mutasyon + İlaç + Klinik + Gen ifadesi	0,75	0,75	0,75	0,75
Mutasyon + İlaç + Klinik + KSV	0,73	0,72	0,73	0,72
Mutasyon + İlaç + Klinik + miRNA	0,73	0,73	0,74	0,73

KSV: Kopya Sayısı Varyasyonu

4.4. Tüm Öznitelikleri İçeren Model Sağkalım Eğrileri

Kanser türlerinde tüm veri tiplerinin kullanılarak eğitilen modellerin tahmin edilen grupları için sağkalım eğrileri Şekil 4.7.'de gösterilmektedir. Modellerin değerlendirildiği Log-rank test sonuçları Tablo 4.20.'da gösterilmektedir.



Şekil 4.7. Kanser sağkalım tahmin modellerinin yaptığı sınıflama sonucunda birbirinden ayrılan 2 hasta grubunun gerçek hayatta kalma sürelerini gösteren eğriler.

Tablo 4.20. Model Log-rank test sonuçları.

Kanser Türü	**Test istatistik**	**p**	**-log2(p)**
Akciğer (LUAD)	6,65	**0,01**	6,65
Akciğer (LUSC)	4,55	**0,03**	4,92
Böbrek	30,02	**<0,005**	24,48
Cilt	4,78	**0,03**	5,12
Karaciğer	13,89	**<0,005**	12,33
Kemik iliği	9,63	**<0,005**	9,03
Kolorektal	4,17	**0,04**	4,6
Meme	8,51	**<0,005**	8,15
Mide	3,69	0,05	4,19
Pankreas	0,14	0,71	0,5
Rahim ağzı	11,87	**<0,005**	10,78
Rahim	20,68	**<0,005**	17,49
Yumurtalık	1,35	0.25	2,03
Tüm Kanserler	668,11	**<0,005**	486,96

10 kanserde ve tüm kanser türlerinin birlikte değerlendirildiği modelde long-rank p değeri < 0,05 olarak bulunmuştur. Bu sonuçlar tahmin edilen grupların sağkalımları arasında istatistiksel olarak anlamlı bir fark olduğunu göstermektedir. Mide, pankreas ve yumurtalık kanserinde p değeri > 0,05 olarak bulunmuştur. Bu 3 kanser türünde sağkalım grupları arasında istatistiksel olarak anlamlı bir fark yoktur.

4.5. Önerilen Model ve COT Model Performansları

Rastgele orman modeli ve COT model performanslarını karşılaştırmak için eğitilen modellerin C-endeks değerleri verilmiştir. Her kanser türünde gen ifadesi, miRNA, KSV, mutasyon, klinik ve ilaç bilgisinin bir arada kullanılmasıyla eğitilen model değerlendirilmiştir. Ayrıca 13 farklı kanser türü için ayrı ayrı veri tipi ile eğitilmiş modeller değerlendirilmiştir. Tablo 4.21.'de COT modeli ile eğitilmiş C-endeks değeri gösterilmiştir. Tablo 4.22.'de rastgele orman ile eğitilmiş modellerin C-endeks değerleri gösterilmiştir.

Tablo 4.21. Kanser türlerinde kullanılan veri tipine göre COT modellerinde C-endeks değerleri.

Kanser Türü	**Gen ifadesi miRNA KSV Mutasyon Klinik İlaç**	**Klinik**	**Gen ifadesi**	**KSV**	**Mutasyon**	**miRNA**	**İlaç**
Kemik iliği	0,95	0,68	0,89	0,71	0,72	0,93	//
Meme	0,97	0,82	0,93	0,80	0,85	0,95	0,77
Rahim ağzı	0,94	0,77	0,97	0,84	0,96	0,99	0,57
Kolorektal	0,99	0,78	0,91	0,76	0,91	0,96	0,60
Böbrek	0,99	0,88	0,94	0,89	0,92	0,97	0,71
Karaciğer	0,97	0,67	0,90	0,82	0,92	0,95	0,52
Akciğer (LUAD)	0,98	0,76	0,87	0,78	0,89	0,89	0,57
Akciğer (LUSC)	0,98	0,69	0,85	0,78	0,89	0,96	0,58
Yumurtalık	0,99	0,61	0,96	0,84	0,88	0,97	0,72
Pankreas	0,94	0,66	0,88	0,77	0,83	0,90	0,64
Cilt	0,99	0,78	0,89	0,78	0,94	0,92	0,57
Mide	0,92	0,70	0,87	0,76	0,85	0,88	0,59
Rahim	0,98	0,68	0,91	0,78	0,89	0,95	0,60
Tüm Kanserler	0,86	0,79	0,78	0,65	0,68	0,78	0,61

KSV: Kopya Sayısı Varyasyonu

Tablo 4.22. Kanser türlerinde kullanılan veri tipine göre rastgele orman modellerinde C-endeks değerleri.

Kanser Türü	Gen ifadesi miRNA KSV Mutasyon Klinik İlaç	Klinik	Gen ifadesi	KSV	Mutasyon	miRNA	İlaç
Kemik iliği	0,69	0,53	0,72	0,48	0,44	0,63	//
Meme	0,57	0,80	0,59	0,54	0,51	0,53	0,70
Rahim ağzı	0,65	0,57	0,56	0,54	0,52	0,58	0,59
Kolorektal	0,60	0,70	0,56	0,53	0,46	0,54	0,44
Böbrek	0,73	0,65	0,72	0,59	0,32	0,72	0,61
Karaciğer	0,63	0,53	0,64	0,56	0,57	0,63	0,50
Akciğer (LUAD)	0,58	0,64	0,62	0,52	0,49	0,52	0,52
Akciğer (LUSC)	0,56	0,57	0,48	0,55	0,49	0,55	0,54
Yumurtalık	0,55	0,45	0,59	0,50	0,49	0,55	0,73
Pankreas	0,50	0,55	0,58	0,44	0,50	0,56	0,62
Cilt	0,59	0,66	0,61	0,51	0,48	0,56	0,48
Mide	0,54	0,56	0,57	0,54	0,45	0,53	0,52
Rahim	0,66	0,66	0,66	0,60	0,66	0,64	0,62
Tüm Kanserler	0,75	0,74	0,75	0,64	0,57	0,72	0,58

KSV: Kopya Sayısı Varyasyonu

4.6. Öznitelik Seçimi ile Elde Edilen Model Sonuçları

Öznitelik seçiminin performansa etkisini araştırmak için tüm özniteliklerin kullanılarak eğitilen modellerden en yüksek sonucu veren iki kanser türü seçilmiştir. Böbrek ve kemik iliği kanseri değerlendirilmiştir. Diğer kanser türlerine göre daha yüksek performans elde edildiği için bu iki tür üzerinden değerlendirme yapılmıştır. İki kanser türü için L1000 veri setindeki tüm özniteliklerle eğitilen modellerin performansları Tablo 4.23.'te gösterilmiştir.

Tablo 4.23. L1000 veri setinde tüm veriyle eğitilen model performansları.

	Böbrek	**Kemik iliği**
F1-skor	0,74	0,73
Doğruluk	0,75	0,75
Kesinlik	0,75	0,75
Duyarlılık	0,75	0,75

Her kanser için İki veri setinde öznitelik seçimi yapılmıştır. Birincisi L1000 genlerinin seçildiği veri seti ve diğeri ise hiçbir seçim yapılmamış tüm gen bilgilerinin olduğu veri setidir. Öznitelik seçiminin model performansı üzerine etkisini araştırmak için böbrek ve kemik iliği dokusunda, özyinelemeli öznitelik eliminasyonu yöntemi kullanılarak, öznitelikler seçilmiştir. L1000 veri setinde algoritma böbrek kanserinde 4378 olan toplam öznitelikten 2188'i önemli olarak belirlemiştir. Kemik iliği kanserinde ise 4042 olan toplam öznitelikten 2181'i önemli olarak belirlenmiştir. Böbrek için tüm özniteliklerle eğitilen model ve seçilen öznitelikler ile eğitilen modellerin performans sonuçları Tablo 4.24.'te gösterilmiştir. Kemik iliği için özniteliklerle eğitilen model ve seçilen öznitelikler ile eğitilen modellerin performans sonuçları Tablo 4.25.'te gösterilmiştir.

Tablo 4.24. Böbrek (KIRP veri seti) için öznitelik seçimi ile yapılan modellerin doğruluk oranları.

Öznitelik sayısı	**Tüm genler veri seti**	**L1000 veri seti**
Tüm öznitelikler	0,72	0,75
50%	0,71	0,76
1000	0,7	0,73
100	0,76	0,77

Tablo 4.25. Kemik iliği (LAML veri seti) için öznitelik seçimi ile yapılan modellerin doğruluk oranları.

Öznitelik sayısı	**Tüm genler veri seti**	**L1000 veri seti**
Tüm öznitelikler	0,69	0,76
50%	0,64	0,75
1000	0,68	0,71
100	0,73	0,79

Tüm genlerin olduğu veri setinde, veri tipleri için öznitelik seçimiyle belirlenen özniteliklerle eğitilen modellerin performansları araştırılmıştır. Tablo 4.26.'da model sonuçları gösterilmiştir. Sonuçlar L1000 veri setinin tüm genler veri setinden daha iyi performansa sebep olduğunu göstermektedir.

Tablo 4.26. Tüm genler veri setinde veri tipine göre öznitelik seçimi ile model doğruluk oranları.

Model	**Böbrek**	**Kemik iliği**
Klinik	0,58	0,61
Mutasyon	0,56	0,63
KSV	0,59	0,57
Gen ifadesi	0,72	0,73
miRNA	0,75	0,69
İlaç	0,65	-

4.7. Rastgele Orman Modeli ve Derin Sinir Ağları Modeli Karşılaştırması

Yakın zamanda yapılmış çoklu kanser türlerinde sağkalım tahmini için yapay öğrenme bazlı yöntem ile bu tez çalışmasında rastgele orman algoritması kullanılarak geliştirilen yeni yöntemin C-endeks performans sonuçları Tablo 4.27.'de gösterilmektedir.

Tablo 4.27. Derin öğrenme bazlı bir çalışma ve rastgele orman modelinin 13 kanser türü için modellerin C-endeks değerleri.

TCGA Projesi	Derin Sinir Ağları Modeli*	Rastgele Orman Modeli**
LAML	0,66	0,69
BRCA	0,73	0,57
CESC	0,74	0,65
COAD	0,77	0,60
KIRP	0,61	0,73
LIHC	0,68	0,63
LUAD	0,73	0,58
LUSC	0,62	0,56
OV	0,59	0,55
PAAD	0,59	0,50
SKCM	0,58	0,59
STAD	0,8	0,54
UCEC	0,66	0,66

*Derin Sinir Ağları Modeli: “Gen ifadesi, miRNA, Klinik” veri tipleri ile oluşturulan

**Rastgele Orman Modeli: “Gen ifadesi, miRNA, Klinik, Mutasyon, KSV, İlaç” veri tipleri ile oluşturulan model

KSV: Kopya sayısı Varyasyonu

5. TARTIŞMA

Çalışmada, kanser hastalarının sağkalımını tahmin etmek için rastgele orman algoritması ile hasta omik veri kullanılarak yeni bir model önerilmiştir. Bunun için, 13 farklı kanser türünde dört farklı omik veri tipi ele kullanılmıştır. Eğitilen modellerin model performanslar sonuçları bulgular bölümde gösterilmiştir.

Akciğer kanseri adenokarsinom alt türünde sağkalım tahmininde ilaç bilgisinin model performansını yükselttiği düşünülmektedir. Ayrıca klinik verinin kullanıldığı model performansına bakılınca LUAD veri seti için klinik verinin sağkalım tahmini için önemli olduğu düşünülmektedir. Çoklu omik verinin kullanılması, tek omik veri tipinin kullanılmasına kıyasla akciğer LUAD verisinde sağkalım tahmin etmede daha yüksek performans göstermiştir. Akciğer kanseri adenokarsinom alt türünde elde edilen sonuçlar Lee ve ark. (81)'nın LUAD veri setinde çoklu omik veri ile yaptıkları çalışmanın sonucu ile benzerdir. Akciğer kanserinin skuamöz hücreli karsinom alt türünde (LUSC veri seti) omik verinin klinik ve ilaç bilgisi ile kullanılmasının sağkalım tahmini daha yüksek performansı gösterdiği düşünülmektedir. LUSC veri seti ile eğitilen modellerin sonuçlarına bakıldığında gen ifadesi verisinin olmadığı modelin yüksek performansa ve bulunduğu modellerin düşük performansa sahip olduğu gözlenmiştir. LUSC için gen ifadesi verisinin sağkalım üzerinde diğerlerinden daha az belirleyici olduğunu düşündürmektedir (Bkz. Tablo 4.6.) (Bkz. Tablo 4.7.). LUAD ve LUSC alt türleri arasında performans açısından farklı sonuçlar elde edilmiştir. Aynı omik veri tiplerinin sağkalım tahminine etkisi iki alt türde farklı gözlenmiştir. İki kanser alt türünün moleküler farklılıkları sağkalım tahmininde etkili olmuş olabilir (99).

Meme kanserinde yapılan farklı veri kombinasyonlarında çoklu omik verinin kullanıldığı modellerin tekli omik kullanılan modellerden daha yüksek tahmin performansı sağladığı düşünülmektedir. Benzer şekilde Huang ve ark. (82)'nın 2019'da yaptıkları çalışma, birden çok omik veri tipinin entegre edilmesinin meme kanserinde sağkalım tahmin performansını artıracağı sonucuna ulaşmışlardır. Yine 2019 yılında Sun ve ark. (77)'nın meme kanserinde çok boyutlu veri entegresinin daha yüksek performans verdiğini belirttikleri çalışmaları, meme kanserinde bulunan sonuçlarımızı desteklemektedir. Bizim çalışmamızda meme kanserinde bu

çalışmalarda kullanılmayan mutasyon verisiyle de çalışılmıştır ve mutasyonun bulunduğu modellerde tahmin performansında artış olduğu belirlenmiştir (Bkz. Tablo 4.13.).

Karaciğer kanserinde gen ifadesi, tek başına değerlendirildiği modele kıyasla diğer omik veri tiplerinin gen ifadesi ile değerlendirildiği modellerde performans artışı gözlenmiştir. Çoklu omik veri tipinin kullanmanın tek omik veri tipinin kullanımına kıyasla daha yüksek performansa sebep olduğu düşünülmektedir. Karaciğer sağkalım tahmininde elde edilen bu sonuç Chaudhary ve ark. (78)'nın karaciğer kanseri sağkalım tahmini için yaptıkları çalışmanın sonuçları ile örtüşmektedir (Bkz. Tablo 4.10.).

Böbrek kanserinde klinik ve ilaç verisi omik veriyle birlikte sağkalım tahmin performansını artırmaktadır. Böbrekte tüm veri tiplerinin kullanıldığı model, tek omik veri tipinin kullanıldığı modellere kıyasla daha yüksek performans gözlemlendi. Böbrek kanserinde çoklu omik verinin sağkalım tahmin performansını artırdığı düşünülmektedir (Bkz. Tablo 4.8.). Cilt kanserinde klinik verinin yanına omik veri tiplerinin eklenmesinin sadece klinik veri ile eğitilen modele göre daha yüksek performansa sebep olduğu düşünülmektedir (Bkz. Tablo 4.9). Kemik iliği kanserinde elde edilen sonuçlara bakıldığında, omik veri tipleri ile kullanılan klinik verinin model performansına katkısı olduğunu düşünüyoruz (Bkz. Tablo 4.11.). Kolorektal kanserde yalnızca mutasyon ile eğitilen model performansı çok düşükken, mutasyon ile kullanılan klinik ve ilaç verinin modelin performansını artırdığı düşünülmektedir (Bkz. Tablo 4.12.). Mide kanserinde gen ifadesi verisiyle diğer veri tiplerinin birlikte kullanılması, tek veri türü olarak kullanıldığı modelle kıyasla model performansını artığı düşünülmektedir (Bkz. Tablo 4.14.). Pankreas kanserinde genel olarak düşük performanslar gözlense de tek omik veri tipinin kullanıldığı diğer modellere kıyasla gen ifadesinin kullanıldığı modelde daha yüksek performans gözlenmiştir (Bkz. Tablo 4.15.). Rahim ağzı kanserinde sağkalım tahmini için gen ifadesi verisinin diğer veri tiplerinden daha az belirleyici olduğu düşünülmektedir (Bkz. Tablo 4.16.). Rahim kanserinde omik veri tipilerinin birlikte kullanılmasının rahim kanserinde sağkalım tahmin performansını artırdığı düşünülmektedir (Bkz. Tablo 4.17.). Yumurtalık kanserinde mutasyon verisinin, gen ifadesi verisinden daha yüksek performans

gösterdiği düşünülmektedir. Diğer dokularda yüksek performans gösteren klinik veri ile eğitilen model, yumurtalık kanserinde en düşük performansı göstermiştir (Bkz. Tablo 4.18.). Genel olarak tüm kanser türlerinde çoklu omik verinin kullanıldığı modellerde tek omik veri tipinin kullanılan modellere göre daha yüksek performans gözlenmiştir. Sonuçların, geniş çaplı kanser verisinde, model girdi verisinin çeşitlendirilerek hastanın sağkalımını daha yüksek performansla tahmin edileceği fikrimizi desteklediği düşünülmektedir.

Kanser türleri arasında tüm omik veri tipleri, klinik ve ilaç veri tiplerinin kullanıldığı model, en yüksek performansı böbrek ve kemik iliği kanserinde göstermiştir. Farklı kanserlerin ayrı ayrı eğitilen modellerinin sonuçlarına bakılınca, her kanser türü model performansının farklı olduğu gözlenmiştir. Kansere özgü farklı genetik işaretlerden kaynaklanan omik verisindeki değişimlerin, sağkalım tahmin performansı etkilediği düşünülmektedir.

Böbrek ve kemik iliğinde, tüm genler ve L1000 genleri ile yapılan model performansları karşılaştırıldığında L1000 veri setinde daha yüksek sonuçlar elde edilmiştir (Bkz. Tablo 4.24.) (Bkz. Tablo 4.25.). L1000 genlerinin kullanılması fark yaratmıştır çünkü önceki çalışmalarda da ortaya çıkarıldığı üzere hücrede meydana gelen değişimlerin birçoğu bu genlerin ifadelerindeki değişimlere yansımaktadır (89). Tüm genlerin kullanılması beraberinde verinin büyük boyutundan kaynaklı olarak sisteme yüksek miktarda gürültüyü de dahil etmemize neden olmaktadır. Öğrenmeye katkı sunamayan bu gereksiz bilgiler ayrıca algoritmanın verimliliğini azaltarak hatalar oluşmasına sebebiyet vermektedir. Verideki gürültü temizlenmezse yanlış analizler yapılabilir (100). Bu açıdan L1000 genlerinin model eğitimde kullanılması, algoritmanın veriden daha yüksek çıkarımlarda bulunmasını sağlar. Bu şekilde elde edilen model performansları daha isabetli analizler elde etmemize yardımcı olmuştur.

Hem böbrek kanserinde hem de kemik iliği kanserinde iki veri setinin sonuçları öznitelik seçimi öncesi performansa benzerdir. Bu sonuçlardan yola çıkarak söyleyebiliriz ki, öznitelik seçimi performans açısından önemli bir fark yaratmamıştır. İki veri seti karşılaştırıldığında öznitelik seçiminin L1000 veri setinde diğer veri setine göre daha yüksek performans gösterdiği gözlenmiştir. Ancak L1000 veri setinde öznitelik seçimi yapılmayan model performansı ile benzer sonuçlar elde edilmiştir. Bu

durum mutasyon, KSV ve gen ifadesinde seçilen bu genlerin güçlü sinyaller taşımasıdır. Bu nedenle aralarından bazılarının çıkarılması bir yandan verideki gürültüyü azaltırken diğer yandan da bilgi kaybına neden olmaktadır. Bundan dolayı performans öznitelik seçimi yapılmayan modelin performansı ile aynı aralıkta kalmaktadır.

Klinik veriyle eğitilen modelde, meme (doğruluk = 0,83), kolorektal (doğruluk =0,70) ve cilt (doğruluk = 0,66) kanserinde, dokudaki diğer modellerden yüksek performans gözlenmiştir (Bkz. Tablo 4.9.) (Bkz. Tablo 4.12.) (Bkz. Tablo 4.13.). Çoğu kanserde klinik veri eklenerek eğitilen modeller daha yüksek performans göstermiştir. Buna örnek olarak, meme kanserinde mutasyon ve ilaç verisiyle eğitilen modelin performansının klinik verinin eklenmesi ile %73'ten %83'e çıkması gösterilebilir. Bulgularımıza göre, klinik değişkenlerin omik veri tipleri ile kullanılması çoğu kanserde daha yüksek tahmin sonuçlarına yol açmıştır. Genomik ve klinik verinin birlikte kullanılmasını araştıran bazı çalışmalarda, klinik bilgileri ve genomik verisinin birleştirilmesinin, bu tür veriyi ayrı ayrı kullanmaktan daha yüksek tahminlere yol açabileceğini öne sürülmüştür (101-103). Klasik yöntemlerde sadece klinik veri ile yapılan sağkalım tahminleri yerine omik veri tipleriyle birlikte kullanılarak daha isabetli tahminler elde edilmesi sağlanabilir.

Hastanın kullandığı ilaç bilgisinin sağkalım tahmini üzerine etkisi olduğu düşünülmektedir. İlaç verisiyle eğitilen modelde yumurtalık (doğruluk = 0,73) ve pankreas (doğruluk = 0,62) kanserlerinde yüksek performans gözlenmiştir (Bkz. Tablo 4.15., Tablo 4.18.). Bu kanserler dışında diğer birçok kanserde ilaç verisi tipinin omik veri tipleri ile kullanılmasının performansı artırdığı gözlenmiştir. Ayrıca ilaç ve klinik verinin birlikte kullanılmasının sağkalım tahmin performansını artırdığı düşünülmektedir.

Mutasyon, klinik ve ilaç verisiyle eğitilen model akciğer kanseri LUAD veri setinde ve tüm kanserlerin birleştirildiği veriyle eğitilen modellerde yüksek performans göstermiştir. Yine aynı veri kombinasyonu ile meme, kolorektal, rahim, yumurtalık, akciğer LUSC kanserlerinde daha yüksek performans elde edilmiştir. Bazı kanser türünde, mutasyon verisinin hastanın klinik ve ilaç veriyle birlikte kullanılmasının, kanser sağkalım tahmininde, daha yüksek performansa sahip

olduğunu göstermiştir. Gen ifadesi, klinik ve ilaç verisiyle eğitilen model cilt ve mide kanserinde çoğu veri kombinasyonlarından daha yüksek sonuç vermiştir. Bu da bazı kanser türlerinde mutasyon yerine gen ifadesi verisinin hastanın klinik ve ilaç bilgileri ile değerlendirilmesinin kanser sağkalım tahmininde daha etkili olabileceğini göstermiştir. Bu sonuçlar kanser türleri arasında omik veri tiplerinin sağkalım tahmininde farklı etkilere sahip olduğunu ortaya koymaktadır.

Tüm kanserlerin birleştirildiği veriyle eğitilen model performansı (doğruluk = 0,75) birçok kanser bazlı modelden daha yüksek sonuç vermiştir (Bkz. Tablo 4.19.). Tüm kanserlerin birleşmesi modelin eğitildiği gözlemlerin sayısını artırmıştır. Algoritma daha fazla gözlemle öğrenmeyi gerçekleştirdiği için verinin özelliklerini öğrenmesi kolaylaşmıştır. Bu da model performansının artırmasına katkı sağlamıştır. Vale-Silva ve ark. (79)'nın 33 kanser türünde yaptıkları çalışmada, farklı omik veri tipleri ve klinik veri tipinin kullanılmasının tüm kanser türlerinde yüksek tahmin sağladığını belirtmişlerdir. Benzer şekilde model sonuçlarımız gösterir ki, çoklu kanser için farklı omik ve klinik verinin kullanıldığı modellerde daha yüksek tahmin performansı elde edilmiştir. Bu da geniş kanser verisinde moleküler ve diğer veri tiplerinin bir arada kullanılması ile daha yüksek sonuçlar elde edilmesini sağlar. Ayrıca bu tez çalışmasında, ilgili çalışmada olmayan omik veri tipleri KSV ve mutasyon bilgisi de kullanılmıştır.

Tüm omik veri tipleri kullanıldığı 10 kanser türünde ve tüm kanserlerde grupların sağkalımları arasında istatistiksel olarak anlamlı sonuçlar bulunmuştur (Bkz. Tablo 4.20.). 3 kanser türü dışında bulunan anlamlı sonuçlar çoklu omik verinin kullanılmasının sağkalım tahmin etmede başarılı olduğunu ortaya koymaktadır. Makine öğrenmesinde girdi verisi olarak kullanılan omik veri tiplerinin çeşitlendirilmesi sağkalımın doğru tahmin edilmesini sağlamaktadır. Literatürdeki çoklu omik verinin kullanılmasının sağkalımı tahminindeki başarısını vurgulayan çalışmalara benzer sonuçlara ulaştığımızı düşünmekteyiz (77-80).

Sağkalım tahmini modelimiz ile klasik bir yöntem olan COT modeli karşılaştırılmıştır. Her kanser türünde iki model için ayrı ayrı veri tipleri ve tüm veri tipleri ile eğitilen modeller değerlendirilmiştir. Modellerin C-endeks değerleri ile karşılaştırılmıştır. COT modelini eğitirken L1000 genleri seçilen veri setleri

kullanılmıştır. Ancak omik verinin yüksek boyutları COT için sorun yaratmıştır. Bu sorun COT algoritmasının öznitelikler arasında bulduğu yüksek doğrusallıktan kaynaklanmıştır. Bu durumun önüne geçmek için LASSO ile öznitelik seçimi yapılmıştır (Bkz. Tablo 4.5.). COT'da öznitelik seçimi yapılan omik veri tipleriyle eğitilen modellerinde aşırı öğrenme olduğu, bundan dolayı gerçek sonuçlar vermediğini düşünüyoruz. COT yönteminde omik veri tipleri dışında kalan klinik ve ilaç veri tipleri sonuçları, karşılaştırma için dikkate alınabilir. Klinik veride COT bütün kanser türlerinde daha yüksek sonuçlar vermiştir. Rastgele orman ile eğitilen modeller meme (0,80) ve rahim (0,66) kanserinde COT modeline en yakın sonuçları vermiştir. İlaç verisinde ise rahim ağzı (0,59), yumurtalık (0,73) ve rahim (0,62) kanserlerinde rastgele orman modeli daha yüksek sonuç vermiştir. İlaç verisinde bu 3 kanser türü dışında COT modelinin daha yüksek sonuç verdiği belirlenmiştir (Bkz. Tablo 4.21.) (Bkz. Tablo 4.22.).

Rastgele orman modelimizin C-endeks değerleri tüm veri tiplerinin kullanıldığı modelde 0,5'ten büyük olduğu gözlenmiştir (Bkz. Tablo 4.22.). C-endeks için 0,5'ten yüksek değerler genellikle anlamlı kabul edilir, bu da modellerin kabul edilebilir sonuçlar verdiği göstermektedir. Böbrek kanserinde ve tüm kanserlerin birlikte değerlendirildiği modelde en yüksek sonuçlar elde edilmiştir. Klinik verinin kullanıldığı modelde en yüksek 0,8 ile meme kanserinde, gen ifadesinde verisinde en yüksek 0,75 ve KSD verisinde en yüksek 0,64 ile tüm kanserlerin kullanıldığı model, mutasyon verisinde 0,66 ile rahim kanseri, miRNA verisinde böbrek ve tüm kanserler modeli, ilaç verisi için 0,73 ile en yüksek C-endeks değerini vermiştir. Bu modellerde 0,5'ten yüksek değere sahip olduğu için anlamlı kabul edilebilir. Sağkalım tahmininde kanser türleri arasında ve farklı veri çeşitlerinde model performansında farklılık gözlenmiştir.

Cheerla ve Gevaert (80)'nın 2019 yılında 20 kanser türü ile yaptıkları çalışmada TCGA projelerinin verisi kullanılmıştır. Tez çalışmamızda kullanılan 13 TCGA projesine de çalışmada yer verilmiştir. Model mimarisi olarak derin sinir ağlarını kullanmışlar ve sağkalım tahmini yapan bir model geliştirmişlerdir. Çalışmada miRNA, gen ifadesi omik veri tipleri ile histopatolojik mikroskop slaytlar ve klinik veri kullanılmıştır. Farklı veri tipi kombinasyonları denemişlerdir. Modelimizle

karşılaştırmak için kullandıkları gen ifadesi, miRNA ve klinik veriyle eğitilen model sonuçları ele alınmıştır. Modelimizde tüm veri tiplerini kullanarak eğitilen modelle karşılaştırılmıştır. İlgili çalışmadaki kullanılmayan mutasyon, KSV ve ilaç bilgisinin kullanılmasının model performansına etkisi araştırılmıştır. Tablo 4.27.'de iki modelde ortak çalışılan TCGA projelerinin model performansları verilmiştir. Kemik iliği kanseri için kullanılan LAML veri setinde rastgele orman modelimiz daha yüksek performans göstermiştir. Bu tez çalışmasında böbrek kanseri için kullanılan KIRP veri setinde %73 ile diğer modele göre %12 daha yüksek performans gösterildiği ortaya konmuştur. Rastgele orman modeli, cilt ve rahim kanseri için kullanılan SKCM ve UCEC veri setlerinde diğer modelle benzer sonuçlar ortaya koymuştur. Diğer kanser türleri için derin sinir ağlarının kullanıldığı diğer model daha yüksek performans göstermiştir. Böbrek kanseri için ortaya çıkan önemli fark daha fazla omik veri tipi ve ilaç bilgisi ile daha yüksek bir performans elde edildiğini gösterilmektedir. Ancak diğer çoğu dokuda derin öğrenme bazlı modelin daha yüksek sonuç vermesi veriden daha yüksek çıkarımlar yaparak sağkalımı tahmin edebildiklerini göstermektedir. Verideki eksik bilgileri olan hastaların kaybını önlemek için denetimsiz öğrenme kullanılmıştır. Rastgele orman modeli için veri düzenlerken eksik bilgileri olan hastaları çalışmadan çıkarmıştık. Bu yaklaşımla bizim çalışmamızdan farklılaşmaktadır. Hasta sayısında veride artış olması derin öğrenme modelinin daha yüksek performansına neden olmuş olabilir. Derin öğrenme bazlı bu çalışmada farklı omik veri tiplerinin entegrasyonunun daha yüksek sonuçlar doğurabileceği belirtilmiştir. Böbrek kanserinde daha yüksek performans gösteren modelimizin bu fikri desteklediğini düşünüyoruz.

Çalışmada kanser hastalarının tedavi süresince kullanmış oldukları ilaç bilgisi bazı modellerde girdi verisinin bir parçası olarak kullanılmıştır. Daha önceki benzer kanserde sağkalım tahmini çalışmalarında hastanın ilaç kullanım bilgisine yer verilmemiştir. Bu çalışmada ilaç bilgisinin girdi verisi olarak kullanmasıyla ilaç bilgisinin sağkalım tahminine etkisini ele aldık. İlaç bilgisinin çoğu kanser türünde sağkalım tahmin performansını iyileştirdiği gözlenmiştir. Ancak bu durumu değerlendirirken bazı faktörleri göz önüne almaktayız. Birtakım biyobelirteçlerin hastada var olup olmaması prognozu etkilemektedir. Doktor hastanın prognozuna yönelik beklenti doğrultusunda hastaya ilaç reçete etmektedir. Modelin girdi

aşamasında ilaç kullanan hastaların prognozuyla ilgili doktor görüşünü de dahil etmekteyiz. Bu yüzden ilaç verisinin kullanılması modelde veri sızıntısına neden olabilir. Veri sızıntısı modelin daha iyi tahminde bulunmasına sebep olarak yanıltıcı sonuçlar ortaya çıkarabilir. Bu durum çalışmamızın potansiyel sınırlarını ortaya koymaktadır.

6. SONUÇ VE ÖNERİLER

6.1. Sonuç

Çalışmamız, modellemede kullanılan girdi veri tipleri çeşitlendirilerek daha başarılı sağkalım tahmin modelleri geliştirmek üzerine tasarlanmıştır. Kişiselleştirilmiş çoklu omik veri tipleri kullanılarak, hasta sağkalımını tahmin eden yeni bir model önerilmiştir.

Birden çok sayıda omik veri tipinin kullanıldığı modeller, tek bir omik veri tipinin kullanıldığı modellere kıyasla hasta sağkalımını belirlemede daha yüksek performans sergilemiştir. Çoklu omik verinin kullanımında literatürdeki çalışmalara benzer sonuçlar elde edilmiştir. Tüm kanser türlerinin bir arada değerlendirildiği modelde tek kanser bazlı modellere göre daha yüksek performans bulunmuştur.

Sonuçlar, çoklu omik verinin kanser hastalarının sağkalımını belirlemede başarılı olacağı fikrini doğruladı. Omik verinin kullanılmasının kanserde sağkalım tahmininde daha başarılı tahminler ortaya koyacağı sonucuna varılmıştır. Kanser türlerinde farklı omik veri tiplerinin model performansına etkisinin farklı olduğu görülmüştür.

Sağkalım tahmin için omik veri tiplerine klinik ve ilaç veri tipleri dahil edildiğinde model performansı artmaktadır. Bu sonucun klinik ve ilaç veri tiplerinin sağkalım için önemli olduğunu göstermektedir.

Meme kanserinde omik veri tiplerine ek olarak kullanılan ilaç ve klinik veri ile eğitilen modelin tahmin ettiği sağkalımları arasında istatistiksel olarak anlamlı sonuçlar elde edilmiştir. L1000 genleri tüm genlere kıyasla daha yüksek sonuç vererek, tüm gen bilgisine gerek olmadan daha önemli bir grup genle daha yüksek sağkalım tahmini yapılabileceğini göstermektedir.

Daha önceki hesaplamalı çalışmalarda sağkalım tahmini için hastanın ilaç kullanım bilgilerinden yararlanılmamıştır. Çalışmamızda sağkalım tahmini için ilaç bilgisi kullanılmasıyla diğer çalışmalardan farklılaşmaktadır. Sağkalım için ilaç bilgisi çoğu kanser türünde önemli bir bilgi olarak sağkalım tahmin performansını arttırmıştır.

6.2. Öneriler

Bazı dokularda, mutasyon veri tipi kullanılarak diğer dokulara kıyasla daha yüksek tahmin performansı elde edilmiştir. Her dokuda farklı sayıda genlerin mutasyon bilgisi kullanılmıştır. Mutasyon veri tipinin kullanıldığı modellerde doku bazında performansı değişmektedir. İleriki çalışmalarda bir kanserde iyi sonuç veren bir grup gendeki mutasyon bilgisi diğer kanser türleri için de değerlendirilebilir. Aynı şekilde gen ifadesi verisindeki benzer durumlar da diğer kanser türleri üzerinde değerlendirilebilir.

Hastaya ait mutasyon bilgisinin bazı kanser türlerinde sağkalım tahmin performansını iyileştirdiği görülmüştür. Veride mutasyon hakkında daha bilgilendirici bir ifade şekli model performansını önemli ölçüde arttırabilir. Büyük boyutlu yapısal mutasyonlar farklı tipleri üzerinden ele alınarak modellerde kullanılabilir. Gelecekte farklı tipteki mutasyonları kategorize ederek özniteliklerin oluşturulmasının tahmin performansına etkisi araştırılacaktır.

Çalışmamızda sadece tek ilaç kullanımları dikkate alınmıştır ancak kanserde yaygın olarak kullanılan protokoller mevcuttur. İlaç kombinasyonları, kanser ilaç direnciyle mücadelede artırılmış etkinlik sağlayabilir ve bu nedenle hastalar için daha sürdürülebilir tedavi seçenekleri sağlayabilir Standart tedavilere dirençli hale gelen birçok kanser hastasının kanserli hücrelerini etkili şekilde inhibe edebilen, sağlıklı hücrelerine daha az zarar veren ve ilaç direncinin ortaya çıkmasını engelleyen ilaç kombinasyonları kullanılabilir (104). Kanserinde kullanılan ilaç kombinasyonları hastalara nüks, direnç ve toksik etkilerini en aza indirirken tedaviden maksimum fayda elde etme fırsatı sunar. Örneğin metastatik meme kanserinde dosetaksel, doksorubisin gibi ajanlarla kombinasyon halinde kullanılmıştır ve tedavide çok etkili olduğu raporlanmıştır (105). Yine benzer şekilde paklitaksel/sisplatin kombinasyonu ve paklitakselin, siklofosfamid, 5-FU ve mitoksantron gibi diğer ilaç kombinasyonları da meme kanserin tedavisinde etkilidir (106). İleriki çalışmalarda meme kanserinde örnek olarak verildiği gibi farklı kanser türleri için uygulanan ilaç kombinasyonu protokolleri de değerlendirilecektir.

Çalışma sonuçları, çoklu omik verinin modelleme aşamasında kullanılmasının kanser hastalarının sağkalımını belirlemede başarılı olacağı fikri doğrulanmıştır. Gelecekteki yapılacak çalışmalar, proteomik, lipidomik, glikomik ve benzeri veri türlerini de modellemeye dahil etmeyi düşünüyoruz. Ayrıca, özellikle çok modlu öğrenme çerçevesinde, yeni makine ve derin öğrenme algoritmalarını kullanmayı planlamaktayız.

7. KAYNAKLAR

1. Mariotto AB, Noone AM, Howlader N, Cho H, Keel GE, Garshell J, et al. Cancer Survival: An Overview of Measures, Uses, and Interpretation. JNCI Monographs. 2014;2014(49):145-86.

2. Courtnay R, Ngo DC, Malik N, Ververis K, Tortorella SM, Karagiannis TC. Cancer metabolism and the Warburg effect: the role of HIF-1 and PI3K. Molecular biology reports. 2015;42(4):841-51.

3. Roy P, Saikia B. Cancer and cure: A critical analysis. Indian journal of cancer. 2016;53(3):441.

4. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. New England journal of medicine. 2000;343(2):78-85.

5. Hanahan D, Weinberg RA. The hallmarks of cancer. cell. 2000;100(1):57-70.

6. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet. 2015;385(9963):117-71.

7. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021;71(3):209-49.

8. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nature Medicine. 2004;10(8):789-99.

9. Jones PA, Baylin SB. The Epigenomics of Cancer. Cell. 2007;128(4):683-92.

10. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31(1):27-36.

11. Sobin LH, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours: John Wiley & Sons; 2011.

12. T.C. Sağlık Bakanlığı, Kanser Tedavisi Nelerdir [Internet]. 2017 [Erişim Tarihi 5 Haziran 2022]. Erişim adresi: https://hsgm.saglik.gov.tr/tr/kanser-tedavisi#:~:text=Kanserde%20yayg%C4%B1n%20olarak%20kullan%C4%B1lan%20tedavi,tek%20ba%C5%9F%C4%B1na%20veya%20birlikte%20uygulanmaktad%C4%B1r. [

13. T.C. Sağlık Bakanlığı, Cerrahi Tedavi [Internet]. 2017 [Erişim Tarihi 5 Haziran 2022]. Erişim adresi: https://hsgm.saglik.gov.tr/tr/kanser-tedavisi/kanser-tedavisi-nelerdir/cerrahi-tedavi.html#:~:text=Cerrahi%2C%20kanserli%20dokunun%20v%C3%BCcuttan%20%C3%A7%C4%B1kart%C4%B1lmas%C4%B1d%C4%B1r,azalt%C4%B1lmas%C4%B1nda%20kullan%C4%B1lan%20bir%20tedavi%20y%C3%B6ntemidir. [

14. T.C. Sağlık Bakanlığı, Radyasyon Tedavisi [Internet]. 2017 [Erişim Tarihi 5 Haziran 2022]. Erişim adresi: https://hsgm.saglik.gov.tr/tr/kanser-tedavisi/kanser-tedavisi-nelerdir/kanser-tedavisinde-radyasyon.html [

15. T.C. Sağlık Bakanlığı, Kemoterapi [Internet]. 2017 [Erişim Tarihi 5 Haziran 2022]. Erişim adresi: https://hsgm.saglik.gov.tr/tr/kanser-tedavisi/kanser-tedavisi-nelerdir/kemoterapi.html [

16. Wilson B, Nicholls SG. The Human Genome Project, and recent advances in personalized genomics. Risk Management and Healthcare Policy. 2015:9.

17. Zugazagoitia J, Guedes C, Ponce S, Ferrer I, Molina-Pinelo S, Paz-Ares L. Current Challenges in Cancer Treatment. Clinical Therapeutics. 2016;38(7):1551-66.

18. Offit K. Personalized medicine: new genomics, old lessons. Human Genetics. 2011;130(1):3-14.

19. Verma M. Personalized Medicine and Cancer. Journal of Personalized Medicine. 2012;2(1):1-14.

20. Peck RW. The right dose for every patient: a key step for precision medicine. Nature Reviews Drug Discovery. 2016;15(3):145-6.

21. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer. 2003;89(2):232-8.

22. Stevenson M, EpiCentre I. An introduction to survival analysis. EpiCentre, IVABS, Massey University. 2009.

23. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association. 1958;53(282):457-81.

24. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. British Journal of Cancer. 1977;35(1):1-39.

25. Lee ET, Wang J. Statistical methods for survival data analysis: John Wiley & Sons; 2003.

26. Cutler SJ, Ederer F. Maximum utilization of the life table method in analyzing survival. Journal of chronic diseases. 1958;8(6):699-712.

27. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187-202.

28. Therneau TM, Grambsch PM. The cox model. Modeling survival data: extending the Cox model: Springer; 2000. p. 39-77.

29. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer. 2003;89(3):431-6.

30. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning: MIT press; 2018.

31. Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? Genome Biology. 2013;14(5):205.

32. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321-32.

33. Pang-Ning T, Michael S, Vipin K. Introduction to data mining, 2006. EQ.1:337-41.

34. De Cock M, Dowsley R, Nascimento ACA, Railsback D, Shen J, Todoki A. High performance logistic regression for privacy-preserving genome analysis. BMC Medical Genomics. 2021;14(1).

35. Lavalley MP. Logistic Regression. Circulation. 2008;117(18):2395-9.

36. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Briefings in Bioinformatics. 2006;7(1):86-112.

37. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and Information Systems. 2008;14(1):1-37.

38. Guenther N, Schonlau M. Support Vector Machines. The Stata Journal: Promoting communications on statistics and Stata. 2016;16(4):917-37.

39. Foody GM, Mathur A. Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sensing of Environment. 2004;93(1-2):107-17.

40. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. 2005;34(2):113-27.

41. Dongare A, Kharde R, Kachare AD. Introduction to artificial neural network. International Journal of Engineering and Innovative Technology (IJEIT). 2012;2(1):189-94.

42. Kingsford C, Salzberg SL. What are decision trees? Nature Biotechnology. 2008;26(9):1011-3.

43. Rokach L, Maimon O. Decision Trees. Springer-Verlag. p. 165-92.

44. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

45. Qi Y. Random Forest for Bioinformatics. Springer US; 2012. p. 307-23.

46. Breiman L. Bagging predictors. Machine Learning. 1996;24(2):123-40.

47. Stransky B, Galante P. Application of Bioinformatics in Cancer Research. Springer Netherlands; 2010. p. 211-33.

48. Wu D, Rice CM, Wang X. Cancer bioinformatics: A new approach to systems clinical medicine. BMC Bioinformatics. 2012;13(1):71.

49. Baldi P, Brunak S. Bioinformatics: the machine learning approach: MIT press; 2001.

50. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine. 2010;50(2):105-15.

51. Carter RJ, Dubchak I, Holbrook SR. A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Research. 2001;29(19):3928-38.

52. Bockhorst J, Craven M, Page D, Shavlik J, Glasner J. A Bayesian network approach to operon prediction. Bioinformatics. 2003;19(10):1227-35.

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

54. Zhang W, Li F, Nie L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology. 2010;156(2):287-301.

55. Ishii N, Tomita M. Multi-Omics Data-Driven Systems Biology of E. coli. Springer Netherlands; 2009. p. 41-57.

56. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Molecular and cellular biology. 1999;19(3):1720-30.

57. Ter Kuile BH, Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. FEBS Letters. 2001;500(3):169-71.

58. Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, et al. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. Cell. 2003;115(5):629-40.

59. Hegde PS, White IR, Debouck C. Interplay of transcriptomics and proteomics. Current opinion in biotechnology. 2003;14(6):647-51.

60. Debnath M, Prasad GBKS, Bisen PS. Omics Technology. Springer Netherlands; 2010. p. 11-31.

61. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: Technologies and Their Applications. Journal of Chromatographic Science. 2017;55(2):182-96.

62. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. The Lancet. 2002;359(9318):1686-9.

63. Dickman PW, Adami HO. Interpreting trends in cancer patient survival. Journal of Internal Medicine. 2006;260(2):103-17.

64. Cole BF, Gelber RD, Gelber S, Coates AS, Goldhirsch A. Polychemotherapy for early breast cancer: an overview of the randomised clinical trials with quality-adjusted survival analysis. The Lancet. 2001;358(9278):277-86.

65. Lewis DR, Seibel NL, Smith AW, Stedman MR. Adolescent and Young Adult Cancer Survival. JNCI Monographs. 2014;2014(49):228-35.

66. Group CCLoORS. Survival after laparoscopic surgery versus open surgery for colon cancer: long-term outcome of a randomised clinical trial. The lancet oncology. 2009;10(1):44-52.

67. Nagy Á, Munkácsy G, Győrffy B. Pancancer survival analysis of cancer hallmark genes. Scientific Reports. 2021;11(1).

68. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015;13:8-17.

69. Zupan B, Demšar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artificial Intelligence in Medicine. 2000;20(1):59-75.

70. Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics. 2006;2:117693510600200.

71. Chang S-W, Abdul-Kareem S, Merican AF, Zain RB. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinformatics. 2013;14(1):170.

72. Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Journal of the American Medical Informatics Association. 2013;20(4):613-8.

73. Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. Engineering Applications of Artificial Intelligence. 2013;26(9):2194-205.

74. Chen Y-C, Ke W-C, Chiu H-W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Computers in biology and medicine. 2014;48:1-7.

75. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology. 2018;18(1).

76. Lee C, Zame W, Yoon J, Van Der Schaar M. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1).

77. Sun D, Wang M, Li A. A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019;16(3):841-50.

78. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver CancerUsing Deep Learning to Predict Liver Cancer Prognosis. Clinical Cancer Research. 2018;24(6):1248-59.

79. Vale-Silva LA, Rohr K. MultiSurv: Long-term cancer survival prediction using multimodal deep learning. medRxiv. 2020.

80. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics. 2019;35(14):i446-i54.

81. Lee T-Y, Huang K-Y, Chuang C-H, Lee C-Y, Chang T-H. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. Computational Biology and Chemistry. 2020;87:107277.

82. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. Frontiers in genetics. 2019;10:166.

83. National Cancer Institute, Genomic Data Commons Data Portal (GDC) [Internet]. 2017 [Erişim Tarihi 10 Mayıs 2022]. Erişim adresi: https://portal.gdc.cancer.gov/ [

84. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Współczesna Onkologia. 2015;1A:68-77.

85. National Cancer Institute, The Cancer Genome Atlas Program [Internet]. 2019 [Erişim Tarihi 12 Mayıs 2022]. Erişim adresi: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga [

86. National Cancer Institute, NCI's Genome Characterization Pipeline [Internet]. 2017 [Erişim Tarihi 12 Mayıs 2022]. Erişim adresi: https://www.cancer.gov/about-nci/organization/ccg/research/genomic-pipeline.

87. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013;45(10):1113-20.

88. National Cancer Institute, Outcomes & Impact of The Cancer Genome Atlas [Internet]. 2017 [Erişim Tarihi 12 Mayıs 2022]. Erişim adresi: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history [

89. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Res. 2014;42(Web Server issue):W449-60.

90. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437-52 e17.

91. DrungBank [Internet]. 2006 [Erişim Tarihi 10 Mayıs 2022]. Erişim adresi: https://go.drugbank.com/ [

92. Webb GI, Sammut C, Perlich C, Horváth T, Wrobel S, Korb KB, et al. Leave-One-Out Cross-Validation. Springer US; 2011. p. 600-1.

93. Fonti V, Belitser E. Feature selection using lasso. VU Amsterdam research paper in business analytics. 2017;30:1-25.

94. Yimin W, Aidong Z, editors. Feature selection for classifying high-dimensional numerical data: IEEE.

95. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

96. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997;16(4):385-95.

97. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

98. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15(4):361-87.

99. Kawase A, Yoshida J, Ishii G, Nakao M, Aokage K, Hishida T, et al. Differences Between Squamous Cell Carcinoma and Adenocarcinoma of the Lung: Are Adenocarcinoma and Squamous Cell Carcinoma Prognostically Equal? Japanese Journal of Clinical Oncology. 2012;42(3):189-95.

100. Garćia LPF, De Carvalho ACPLF, Lorena AC. Noisy Data Set Identification. Springer Berlin Heidelberg; 2013. p. 629-38.

101. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Human Molecular Genetics. 2003;12(suppl 2):R153-R7.

102. Pittman J, Huang E, Dressman H, Horng C-F, Cheng SH, Tsou M-H, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proceedings of the National Academy of Sciences. 2004;101(22):8431-6.

103. Van 'T Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530-6.

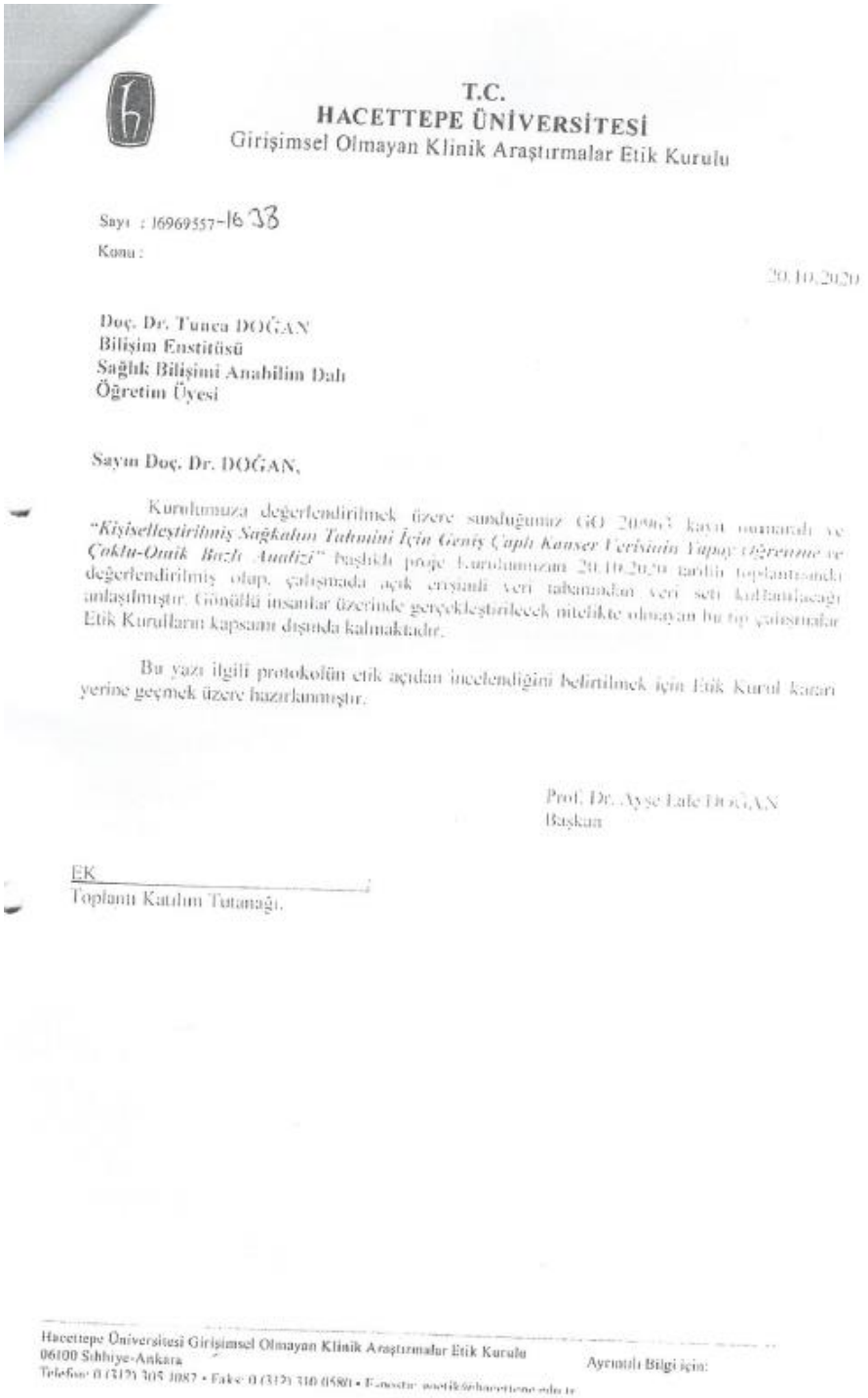
104. Hanahan D. Rethinking the war on cancer. The Lancet. 2014;383(9916):558-63.

105. Esteva FJ, Valero V, Pusztai L, Boehnke-Michaud L, Buzdar AU, Hortobagyi GN. Chemotherapy of metastatic breast cancer: what to expect in 2001 and beyond. The oncologist. 2001;6(2):133-46.

106. Fisusi FA, Akala EO. Drug Combinations in Breast Cancer Therapy. Pharmaceutical Nanotechnology. 2019;7(1):3-23.

8. EKLER

EK-1: Tez Çalışması ile İlgili Etik Kurul İzinleri



T.C.
HACETTEPE ÜNİVERSİTESİ
Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu

Sayı : 16969557-1633
Konu :

20.10.2020

Doç. Dr. Tunca DOĞAN
Bilişim Enstitüsü
Sağlık Bilişimi Anabilim Dalı
Öğretim Üyesi

Sayın Doç. Dr. DOĞAN,

Kurulumuza değerlendirilmek üzere sunduğunuz GO 20/863 kayıt numaralı ve *"Kişiselleştirilmiş Sağkalım Tahmini İçin Geniş Çaplı Kanser Verisinin Yapay Öğrenme ve Çoklu-Omik Bazlı Analizi"* başlıklı proje Kurulumuzun 20.10.2020 tarihli toplantısında değerlendirilmiş olup, çalışmada açık erişimli veri tabanından veri seti kullanılacağı anlaşılmıştır. Gönüllü insanlar üzerinde gerçekleştirilecek nitelikte olmayan bu tip çalışmalar Etik Kurulların kapsamı dışında kalmaktadır.

Bu yazı ilgili protokolün etik açıdan incelendiğini belirtilmek için Etik Kurul kararı yerine geçmek üzere hazırlanmıştır.

Prof. Dr. Ayşe Lale DOĞAN
Başkan

EK
Toplantı Katılım Tutanağı.

Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 1087 • Faks: 0 (312) 310 0580 • E-posta: [illegible]

Ayrıntılı Bilgi için:

EK-2: Tez Çalışması Orijinallik Raporu

turnitin

Dijital Makbuz

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

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Ayşe Nur Çoruh
Ödev başlığı:	Ayşe Nur Çoruh - MSc Tez Final
Gönderi Başlığı:	YL Tez Final
Dosya adı:	Ayse_Nur_Coruh_Tez_4_ekim_2022_1.pdf
Dosya boyutu:	2.4M
Sayfa sayısı:	112
Kelime sayısı:	27,201
Karakter sayısı:	157,235
Gönderim Tarihi:	04-Eki-2022 11:56ÖÖ (UTC+0300)
Gönderim Numarası:	1916292298

YL Tez Final

ORİJİNALLİK RAPORU

%7	%6	%1	%4
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	www.openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	%2
2	tr.ulfsciences.com İnternet Kaynağı	<%1
3	Submitted to Hacettepe University Öğrenci Ödevi	<%1
4	tr.intermediapub.com İnternet Kaynağı	<%1
5	acikbilim.yok.gov.tr İnternet Kaynağı	<%1
6	openaccess.hacettepe.edu.tr:8080 İnternet Kaynağı	<%1
7	Submitted to Bahcesehir University Öğrenci Ödevi	<%1
8	biyoinformatikdunyasi.blogspot.com İnternet Kaynağı	<%1
9	sunaemir.com İnternet Kaynağı	<%1

9. ÖZGEÇMİŞ