Dokuz Eylül Üniversitesi-Mühendislik Fakültesi
Fen ve Mühendislik Dergisi
Cilt 20, Sayı 59, Mayıs, 2018

Dokuz Eylul University-Faculty of Engineering
Journal of Science and Engineering
Volume 20, Issue 59, May, 2018

DOI: 10.21205/deufmd. 2018205947

Kanser Alt-Türlerinin Sınıflandırılması İçin RNA-Sekanslama ve RPPA Verilerinin Karşılaştırılması

Zerrin IŞIK*[1]

[1]Dokuz Eylül Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Tınaztepe Kampüsü Buca 35160, İzmir (ORCID: 0000-0003-1779-1681)

(Alınış / Received: 10.10.2017, Kabul / Accepted: 26.01.2018,
Online Yayınlanma / Published Online: 15.05.2018)

Anahtar Kelimeler
Kanser Sınıflandırma, RNA-Sekanslama, Protein Seviyesi, Makine Öğrenmesi Algoritmaları

Özet: Hastaların kanser alt-türlerini henüz ameliyat olmadan kesin doğrulukla tespit edilebilmek, tanı ve tedavi masraflarının azaltılmasını sağlayacaktır. Bu çalışmanın amacı, over ya da akciğer kanseri olduğu tespit edilen bir hastanın kanser alt-türünü tespit edebilecek belirli sayıdaki biyoişaretin makine öğrenmesi yöntemleriyle bulunmasıdır. Bu amaçla, mRNA gen ekspresyon ve protein seviyesi bilgileri kullanılarak kısıtlı sayıda öz-nitelik seçilmiş, bu öz-niteliklerle gözetimli makine öğrenmesi metotları eğitilmiş, bu modeller ile yeni gelen bir hastanın kanser alt-türü tahmin edilmiştir. Destek vektör makineleri ve rastgele orman algoritmaları, yeni kanser hastalarının kanser alt-türlerini ortalama %87 ile %95 arasında değişen doğruluk dereceleriyle sınıflandırabilmiştir. mRNA gen ekspresyon verisi protein seviyesi verisine göre, her iki kanser türünde de daha başarılı sınıflandırma sonuçları sağlamıştır.

Comparison of RNA-Sequencing and RPPA Data for Classification of Cancer Subtypes

Keywords
Cancer Classification, RNA-Sequencing, Protein Level, Machine Learning Algorithms

Abstract: Accurate identification of cancer subtypes of patients before their surgery will decrease diagnostic and treatment costs. The goal of this study is the identification of a limited number of biomarkers, which can predict subtypes of cancer patients who were diagnosed as lung or ovarian cancer, by using machine learning methods. For this purpose, a limited number of features were selected by using gene expression and protein level data, then supervised machine learning methods were trained with selected features, cancer subtype of a new patient was predicted by using these models. Support vector machines and random forest algorithms can classify cancer subtypes of new patients with the average accuracies of 87% - 95%. mRNA gene expression data provided the better classification results compared to the protein level data.

*Sorumlu yazar: zerrin@cs.deu.edu.tr

1. Giriş

Günümüzde kanser tanısı konulan hastalara, sahip oldukları kanser alt-türüne ve safhasına göre doktorlar tarafından uluslararası kılavuzlara göre seçilen tedaviler uygulanmaktadır. Ameliyat sonrası patologlar tarafından hastaya teşhis konulmakta fakat hastanın sahip olduğu kanser alt-kategorisi her zaman kolaylıkla tespit edilememektedir. Hastayı ameliyat etmeden önce, kanser alt-kategorisinin tespiti veya olası yaşam süresinin tahmin edilmesi gibi durumların belirlenmesi, şu anda klinik düzeyde yaygın olarak uygulanabilir değildir. Bu nedenle geliştirilecek olan işlemsel bir yöntem, patoloğa ön bilgi vererek daha doğru tanı koyma olanağı sağlayabilir. Kanser hastasına özel bilgiler, hasta ameliyat edilmeden önce basit ve kolaylıkla uygulanabilir kan testleri ile tespit edilebilirse, hem hastanın yaşam kalitesini düşürmeden gerekli tedavi programları o hastaya özel olarak dizayn edilebilir; hem de ameliyat, patolojik inceleme gibi işlemlerin maliyetleri düşürülebilir. Bu amaçlarla son yıllarda geliştirilen bazı kanser tanı testleri (örneğin MammaPrint, OncoType vb.) klinik düzeyde kullanılmaya başlanmıştır [1,2]. Fakat birden fazla kanser biyoişaretinin (biomarker) kanserli dokudaki seviyelerini ölçerek uygulanan bu testler hala 100% oranında güvenirliğe sahip değildir. Bu alanda bulunacak yeni biyoişaretler, kanser tanısı koymada, hastalığın gidişatını ve hastanın olası yaşam süresini tahmin etmekte çok büyük katkı sağlayacaklardır.

Bu alandaki işlemsel analiz ihtiyacı göz önüne alınarak bu çalışmada, hastalık tanısı için kliniğe gelen bir hastanın kanser alt-kategorisini tayin edebilecek belirli sayıda biyoişaretin makine öğrenmesi yöntemleriyle bulunması hedeflenmiştir. Yakın zamanda bu amaçla genel kullanıma açılan birçok kanser türüne ait çok sayıda hasta verisi bulunmaktadır. Örneğin, hastaların genom bilgilerine bakılarak belirli gen bölgelerindeki mutasyon durumuna göre kanser türleri otomatik olarak sınıflandırılabilmektedir. Daha detaylı analizler için mutasyon, transkriptom düzeyindeki mRNA ve protein seviyesi, DNA kopya sayısı ve metilasyonu gibi veriler kullanılabilir. Yakın zamanda yapılan bir çalışma DNA metilasyonu, protein, mikro-RNA ve gen ekspresyon verilerini kullanarak over kanseri alt-türlerini tanımlamış; bu alt-türlerin, hastanın yaşam süresi, kanserin tekrarlaması, alınan tedavi türü gibi klinik faktörlerle olan ilişkisini araştırmıştır [3]. Yine birden fazla veri türünü kullanan kapsamlı bir böbrek kanseri çalışmasında, yeni alt-türleri birbirinden ayırt edebilen biyoişaretler ve sinyal yolakları keşfedilmiştir [4]. Başka bir çalışma ise, göğüs kanserinin histolojik evresini gen ekspresyon verisi kullanarak tespit etmeyi amaçlamış; I. ve III. evreleri 0.97 AUC ile ayırt etmeyi başarmıştır [5]. Benzer bir çalışma, protein ve antikor seviyesi verilerini kullanılarak agresif kolon kanserlerindeki biyoişaretleri tayin etmeye çalışmıştır [6]. Diğer bir çalışmada ise, mutasyon ve protein verileri kullanılarak akciğer kanseri alt-gruplarına ayrılmıştır [7]. Yakın zamanda gerçekleştirilen bir çalışmada etkin hale geçmiş yolakların tespiti için mRNA ekspresyon ve protein etkileşim ağları kullanılmış; nöroepitelyal dokulara ait kanser türlerini ayırt edebilmek için destek vektör makinaları kullanılarak %68'lik bir doğruluk derecesi elde edilmiştir [8]. Başka bir çalışma RNA-sekanslama ve RPPA verilerini kullanarak farklı (kanserli ve sağlıklı) doku örneklerini karşılaştırmış; sağlıklı ve hastalıklı dokularda ekspresyonları korelasyon gösteren gen ve proteinleri analiz etmiştir [9]. Yakın zamanda yapılan bir çalışmada rahim ağzı kanseri hastalarından alınan DNA, RNA ve protein örneklerinin genom analizi

yapılarak, bu kanserin moleküler alt-gruplarının profilleri çıkarılmıştır [10]. Prostat kanserinin alt-gruplarını tespit etmek amacıyla RNA-sekanslama verisi üzerinde gözetimsiz kümeleme yöntemleri uygulamış yakın tarihli bir çalışma da bulunmaktadır [11]. Sağlıklı bireyleri meme kanseri teşhisi konulan kişilerden ayırt etmek için, derin öğrenme tekniklerini RNA-sekanslama verisi üzerinde uygulayan yakın tarihli başka bir çalışma mevcuttur [12]. Yakın zamanda yapılan tüm bu çalışmalardan anlaşıldığı üzere, genom düzeyindeki farklı veri türleri kullanılarak hastalık evresi, tekrar etme durumu, alt-türleri, hastanın yaşam süresi gibi birçok klinik faktör, işlemsel yöntemlerle tahmin edilmeye çalışılmaktadır.

Kanser Genom Atlası Projesi (The Cancer Genome Atlas Project - TCGA), 32 farklı kanser türünden hastalara ait, demografik, mutasyon, metilasyon, mRNA ve protein ekspresyon vb. gibi farklı biyolojik verileri içeren bir araştırma projesidir [13]. Bu değerli veri tabanında bazı kanserler için hastanın kanser alt-türleri verilirken, bazı kanser türleri için ise hastaların yaşam süresi gibi daha detaylı bilgiler de paylaşılmıştır. TCGA projesinin genel amacı geç teşhis edilen veya uygun tedavi bulunamadığı için hastanın kaybıyla sonuçlanan kanser türlerinin birbiriyle olan benzerliklerini geniş bir hasta topluluğu üzerinde araştırabilmektir. Sağlıklı bireylerin ölçümlerinin kanser hastalarıyla karşılaştırılması çoğunlukla DNA-dizi analizleriyle yapılırken; mRNA ve protein ekspresyon gibi daha maliyetli deneyler çoğunlukla sadece kanser hastalarının örnekleri için uygulanmıştır. Akciğer kanseri hem dünyada hem de Türkiye'de erkekler arasında en sık görülen kanser türüyken; over kanseri kadınlar arasında görülme sıklığı Türkiye'de dördüncü sırada yer alan ve teşhisi oldukça geç yapılabilen bir kanser türüdür [14].

Bu makalenin amacı, TCGA veri tabanında kanser alt-türü daha önceden tanımlanmış, mRNA gen ve protein ekspresyon verileri paylaşılmış, akciğer ve over kanseri hastalarının kanser alt-türünü tahmin edebilecek yeni akıllı biyoişaretleri bulmaktır [7,15]. Bu amaçla, transkriptom düzeyindeki mRNA ifadesi (RNA-sekanslama) ve protein seviyesi (Reverse Phase Protein Array - RPPA) bilgileri üzerinden kısıtlı sayıda öz-nitelik (mRNA, protein) seçimi yapılacak, bu öz-niteliklerle eğitilen gözetimli makine öğrenmesi yöntemiyle, yeni gelen bir hastanın kanser alt-kategorisi tahmin edilmeye çalışılacaktır.

Literatürde RNA-sekanslama verisini farklı amaçlar doğrultusunda analiz eden birçok çalışma olmasına rağmen, aynı hasta grubundan alınan RNA-sekanslama ve RPPA verilerini karşılaştırmalı olarak analiz ederek kanser alt-türlerini tespit edebilecek biyoişaret çalışmaları bilgimiz dahilinde bulunmamaktadır. Bu çalışmanın özgün değeri, hastalardan alınan mRNA ve protein seviyesi bilgilerini karşılaştırarak, hangi kanser alt-türüne sahip olduğunu tespit etme işleminde daha iyi sonuç verecek biyoişaretlerin bulunmasıdır. Bulunacak biyoişaretlerin sayısının kısıtlı olması, geliştirilecek klinik tanı testlerinde kullanılabilmeleri için maliyetlerin daha düşük olmasını sağlayarak, medikal alanda özgün bir katkı sunabilecektir.

2. Materyal ve Metot

Bu çalışmada kullanılan biyolojik verilerin analizi ve uygulanan makine öğrenmesi algoritmalarının detayları bu bölümde açıklanacaktır.

2.1. Veri ön-işleme

Çalışmamızda kullanılacak olan over (OV) ve akciğer (LUSC) kanseri hastalarına ait mRNA gen ekspresyonu ve protein seviyesi bilgileri, "TCGA Assembler" yazılımı ve "R-Bioconductor" yazılım ortamı kullanılarak lokal

sunucuya kaydedilmiştir [16]. TCGA projesi RNA-sekanslama verisi için gen ekspresyon analizini iki aşamada gerçekleştirmiştir. Öncelikle her bir mRNA sekansının tüm genom üzerinde toplam kaç defa yapıştığını hesaplamak için RPKM ("reads per kilobase per million") yöntemi uygulanmıştır [17]. Sonra bulunan toplam yapışma sayıları üzerinden her bir genin ekspresyonunu kestirebilmek için, RSEM yöntemi uygulanmıştır [18]. Bu çalışmamızda genom bazında kestirimi yapılmış bu gen ekspresyonları lokal sunucuya kayıt edilmiştir. Bu çalışmada kullanılacak olan her iki veri türünde (RNA-sekanslama ve RPPA) ortak olarak örneklemi olan hastalar seçildiğinde, sadece kanser hastalarına ait örneklere ve bu hastaların kanser alt-türlerine ait etiketlemelere ulaşılabilmiştir. Ham verileri, gözetimli bir öğrenme algoritmasında kullanabilmek için, temel veri temizliği ve normalizasyonu yapılmıştır. Öncelikle, tüm hasta örneklerinde tamamen eksik olan (sıfır değerine sahip) mRNA ölçümleri silinerek, ham veri içinde ölçümü yapılamayan mRNA verileri temizlenmiştir. İki kanser türünde bu eksik mRNA ölçümleri, toplam mRNA sayısının %2'si kadardır. Sonraki adımda, mRNA ifadesi ölçümleri logaritmik düzeye ($\log_2$) çevrilmiştir. Sonrasında satır ve sütun bazında normalizasyon yapılarak, tüm mRNA ifadeleri z-puanı ("z-score") olarak temsil edilmiştir.

Böylelikle farklı hastalara ait mRNA ifade örneklerinin hepsi, 0-ortalama ve 1-standart sapma ile gösterilebilmiştir. Protein seviyesi ölçümleri için benzer bir yöntem uygulanarak, ham veriler z-puanı olarak temsil edilmiştir. Gözetimli öğrenme modeli bu z-puanı verileri kullanılarak eğitilecektir. Öğrenme modelini anlamlı bir şekilde eğitebilmek için hem mRNA ifadesi hem de protein seviyesi bilgisi olan hasta örnekleri seçilmiştir. Bu nedenle hasta örneklerinin toplam sayısında bir azalma meydana gelse bile, öğrenme modeli için istatiksel olarak yeterli sayıda hasta örneği bulunmaktadır. Oluşturulan bu veri setinde, akciğer kanseri için toplam 112, over kanseri için ise 202 hasta verisi bulunmaktadır. Tüm bu sınıflandırma ve veriyi temizleme işlemlerinden sonra kullanılacak hasta verisinin büyüklüğü ile ilgili bilgiler Tablo 1'de verilmiştir. Akciğer kanserinin "Classical" alt-türüne (TCGA'da %36'lık örneklem) sahip hastaların, çoğunlukla yüksek miktarda sigara içen erkek hastalar oldukları bilinmektedir. "Primitive" alt-türüne sahip hastalar (TCGA'da %15'lik örneklem) hızlı hücre çoğalmalarına ve kısa yaşam sürelerine sahiptirler. "Basal" alt-türüne sahip hastalar (TCGA'da %25'lik örneklem) fenotip olarak kolay ayırt edilebilirler. "Secretory" alt-türüne sahip hastalarda (TCGA'da %24'lük örneklem) bağışıklık sisteminde görev alan genlerin değişikliğe uğradığı gözlenmiştir.

Over kanserinde ise "Immunoreactive" alt-türüne sahip hastaların (TCGA'da %29'luk örneklem) bağışıklık sistemi zayıf tepkiler verirken, T-hücreleri kemokin ligandları aktif haldedir. "Proliferative" alt-türüne sahip hastalarda (TCGA'da %28'lik örneklem) hücre çoğalması ile ilgili genlerde değişiklik gözlenir. "Differentiation" alt-türüne sahip hastalarda (TCGA'da %25'lik örneklem) yumurtalık tüplerinin salgılamasıyla ilgili genlerde daha olgun düzeye erişmiş bir değişiklik gözlenir. "Mesenchymal" alt-türüne sahip hastalarda (TCGA'da %18'lik örneklem) stroma dokularda oluşan yapılanmalar ile ilgili değişiklikler gözlenir.

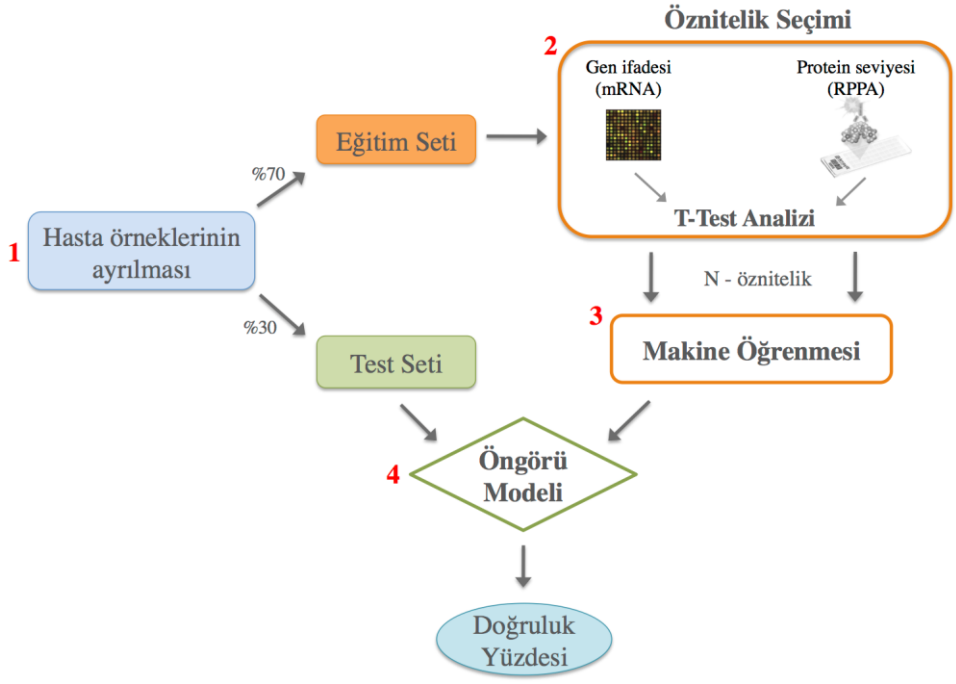
2.2. Gözetimli öğrenme metodu

Transkriptom düzeyindeki mRNA verisi ve protein seviyesi bilgileri üzerinden ayrı olarak en anlamlı öz-nitelikler (biyoişaret) seçilerek, gözetimli öğrenme modeline sınırlı sayıdaki bu öz-nitelikler girdi olarak verilecektir. Bu bilişsel

yöntem dört temel aşamadan oluşur (Şekil 1). Detayları aşağıdaki bölümlerde açıklanan öğrenme ve sınama yöntemi rastgele seçilen hasta örnekleri için 500 defa (500-katlı çapraz doğrulama) tekrarlanır. Sonuç olarak, hesaplanan ortalama doğruluk yüzdesi, her bir algoritmanın genel öngörü performansı olarak sunulmuştur. Yapılan testler sonucu en yüksek doğruluk yüzdesini elde eden öğrenme algoritması ve optimal öznitelik sayısı (N), genel öngörü modeli olarak seçilecektir. Seçilen bu N-öznitelik, tanımlanan sınıflandırma problemi için tahmin edilen biyoişaret setini oluşturur.

Tablo 1. TCGA veri tabanından indirilen ve veri temizliği, normalizasyonu yapılan kanser hastalarına ait verilerin sayısal dökümü

Kanser tipi	Alt-Tür	Toplam hasta sayısı	Toplam mRNA ifadesi	Toplam protein seviyesi
Akciğer	Basal	28	20218 (gen)	186 (protein)
Akciğer	Classical	38	20218 (gen)	186 (protein)
Akciğer	Primitive	18	20218 (gen)	186 (protein)
Akciğer	Secretory	28	20218 (gen)	186 (protein)
Over	Differentiated	50	20131 (gen)	175 (protein)
Over	Immunoreactive	59	20131 (gen)	175 (protein)
Over	Mesenchymal	36	20131 (gen)	175 (protein)
Over	Proliferative	57	20131 (gen)	175 (protein)



Şekil 1. İşlemsel yöntemin temel çalışma aşamaları

2.2.1. Hasta örneklerinin ayrılması

Hastanın kanser alt-türünü tahmin etmeyi sağlayan model gözetimli makine öğrenmesi yöntemiyle oluşturulacağı için, var olan hasta örnekleri (alt-türleri içinde dengeli bir oranda) eğitim (%70) ve test (%30) amaçlı olarak rastgele iki bölüme ayrılır. Seçilen her bir eğitim veri seti için aşağıda açıklanan iki aşamalı öğrenme tekniği uygulanır.

2.2.2. Öz-nitelik seçimi

İnsan genomundaki yaklaşık 20000 genin, kanser hastalarındaki mRNA seviyeleri TCGA veri tabanında bulunmaktadır. Makine öğrenmesi yöntemine en belirleyici olan öznitelikler girdi olarak verilmelidir. Aksi takdirde bu kadar büyük boyutlu bir veri uzayında, bir modelin eğitilebilmesi için gereken hasta örneklem sayısı yeterli olmayacaktır. İşlemsel yöntemin ilk aşaması, en önemli "N" tane özniteliğin seçilmesidir. Bu amaçla istatiksel bir yöntem olan *t*-testi yöntemi uygulanmıştır. Bu yöntem birbirinden bağımsız olarak örneklenen iki grubunun ortalamalarının birbirinden farklı olup olmadığını analiz etmektedir. Örneklem ortalaması birbirinden en farklı olan "N" tane öznitelik, en anlamlı olarak seçilecektir. Bizim kanser hastalarımız kendi içinde dört ayrı alt-türe ayrılmıştır. Her bir hasta grubunu (alt-tür) diğer üç grupla karşılaştırabilmek için; örneğin "Basal" grubundaki hastaları bir sınıf olarak alırken, "Classical", "Primitive" ve "Secretory" gruplarındaki tüm hastalar da diğer sınıf olarak alınır. R-Bioconductor içindeki "genefilter" kütüphanesinin "rowttests" fonksiyonu kullanılarak *t*-test analizi gerçekleştirilmiştir. Analiz sonucunda *t*-istatistik puanına göre belirleyici olan "N" tane mRNA ya da protein seçilmiştir. Buradaki N değeri 500-katlı çapraz doğrulama deneyleriyle (5 sefer tekrarlanarak) 10 ile 100 arasında değişen bir aralıkta belirlenmiştir. Çünkü bu öznitelik sayısı kısıtlı tutularak, daha sonra geliştirilecek olan klinik tanı testlerinde az sayıdaki mRNA ve proteinin kontrol edilmesi sağlanacağından, tanı kitinin maliyeti de düşük tutulacaktır. Seçilen bu N-özniteliğin mRNA gen ifadesi ya da protein seviyesi değerleri (z-puanı) bir sonraki aşamaya girdi olarak verilir.

2.2.3. Makine öğrenmesi metotları

Seçilen N-öznitelik kullanılarak bir öğrenme modeli oluşturulur. Amaç, bir sonraki aşamada gelecek olan yeni hasta verisinin, kanser alt-kategorisini başarılı bir şekilde tahmin edebilmektir. Öğrenme işlemi gözetimli bir model olarak yapılacağı için, en az iki hasta sınıfı tanımlanmalıdır. Hastanın kanser alt-türünü bulmak için toplam dört sınıf tanımlanır ve her bir hasta sınıfını diğer üç sınıftan ayırt edebilecek bir makine öğrenmesi modeli eğitilir. Veri türüne (mRNA / protein) göre seçilen N-özniteliğin, eğitim veri setindeki z-puanları makine öğrenmesi algoritmasına girdi olarak verilir.

- Destek vektör makineleri

Destek Vektör Makineleri (DVM) sınıflandırma ve örüntü tanıma problemlerini çözebilmek için geliştirilmiş istatiksel öğrenme modeline dayanan bir makine öğrenmesi yöntemidir [19]. İki sınıftan oluşan bir problemde, DVM'ler iki sınıfı birbirinden ayırabilecek optimal düzlemi bulmayı hedefler. Bu amaçla iki sınıfın birbirine en yakın örnekleri destek vektörleri için sınır olarak belirlenir, sonrasında ise bu destek vektörleri arasındaki uzaklığı olabildiğince artıracak tüm olası düzlemler hesaplanır. Eğer iki sınıfın örnekleri uzayda doğrusal vektörler ile ayırt edilemeyecek durumdaysa, n-boyutlu girdi vektörü çekirdek fonksiyonları kullanılarak, N-boyutlu özellik vektörüne dönüştürülür. Böylece

daha yüksek bir boyuta dönüştürülen girdi vektörü, artık doğrusal düzlemler ile kolaylıkla ayrılabilir bir hale gelir. Bu çalışmada, R-Bioconductor içinde yer alan "e1071" kütüphanesinin DVM fonksiyonları (svm, predict) kullanılmıştır. DVM parametreleri varsayılan değerler olarak bırakılmıştır.

- Rastgele orman algoritması

Rastgele Orman (RO) algoritması ağaç tabanlı olan ve rastgelelik özelliği eklenmiş bir torbalama yönteminin gelişmiş halidir [20]. Temel olarak, ağaç üzerinde bir düğüm ayrılacağı zaman, tüm değişkenlerin arasından rastgele olarak m-tane değişken seçilerek bu ayrım uygulanır. Her bir ağaç gerçek bir veri setinden değiştirmeli alt-örnekler alınarak oluşturulur ve oluşturulan ağaçlarda budama uygulanmaz. RO algoritmasının uygulanması için iki temel parametre vardır: öznitelik sayısı (m), toplam ağaç sayısı (ntree). Bu çalışmada, R-Bioconductor içinde yer alan "randomForest" kütüphanesinin fonksiyonları (randomForest, predict) kullanılmıştır. Algoritmanın temel parametrelerinden sadece toplam ağaç parametresi 2000 olarak artırılmıştır, *m* ise toplam sınıf sayısı (m=2) olarak varsayılan değeriyle kullanılmıştır.

Yukarıda özetlenen iki makine öğrenmesi algoritması, verilen eğitim hasta verisi üzerinde eğitilerek kendilerine ait sınıflandırma modellerini oluştururlar. Bu aşamanın çıktısı, farklı algoritmalar ile oluşturulmuş iki ayrı sınıflandırma modelidir.

2.2.4. Öngörü modeli

Eğitilen her bir modelin tahmin yeteneği, daha önce modelin eğitilmesi aşamasında hiç kullanılmamış olan, test veri kümesinde sınanır ve doğruluk yüzdesi hesaplanır. İki sınıftan oluşan sınıflandırma probleminin başarım sonuçları Tablo 2'deki gibi bir karışıklık matris oluşturularak hesaplanır.

Tablo 2. Karışıklık matrisi

		Öngörülen Sınıf	
		Sınıf 1	Sınıf 2
Doğru Sınıf	Sınıf 1	Gerçek Pozitif	Yanlış Negatif
	Sınıf 2	Yanlış Pozitif	Gerçek Negatif

Doğruluk yüzdesi DVM, RO yöntemleri ve mRNA, protein verileri üzerinde ayrı olarak hesaplanmıştır.

$$DY = \frac{GP + GN}{GP + YP + YN + GN} \quad (1)$$

Sonuçların değerlendirmesinde kullanılan diğer bir metrik de "ROC" eğrisi ve altında kalan alanı temsil eden "AUC" değeridir [21]. ROC eğrisi bir öğrenme modelinin farklı eşik değerleri için duyarlılık (hassasiyet) ile seçicilik (özgüllük) arasındaki ilişkisini temsil etmektedir. ROC eğrisinin y-ekseninde DPO (Doğru Pozitif Oran), x-ekseninde ise YPO (Yanlış Pozitif Oran) yer almaktadır (Denklem 2).

$$DPO = \frac{GP}{GP + YN}$$
$$YPO = \frac{YP}{YP + GN} \quad (2)$$

ROC eğrisinin altında kalan toplam alan ise AUC olarak hesaplanır. Örneğin iki sınıftan oluşan bir sınıflandırma probleminde AUC değeri 0.5 değerinden düşük bir değere sahipse, o modelin yararsız olduğu kabul edilir. Modelin AUC değeri 1.0 değerine yaklaştıkça, ayırt etme performansı da mükemmele yaklaşmaktadır. Bu çalışmada, ROC eğrisi çizilmesi ve AUC hesabı için R-Bioconductor içinde yer alan "ROCR" ve "cvAUC" kütüphanelerinin çeşitli

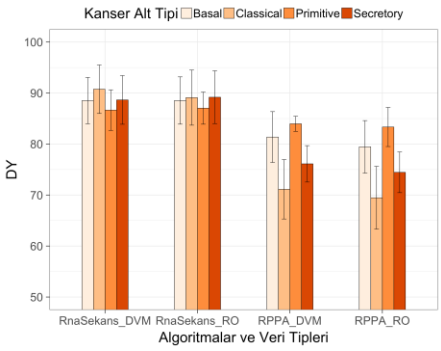
fonksiyonları (prediction, performance, cvAUC) kullanılmıştır.

Kanser hastalarına ait toplam dört farklı alt-tür (sınıf) bulunduğundan, birden fazla ikili sınıflandırıcı model oluşturulması gerekir. Bu nedenle çalışmada "one against all" (OAA) yöntemi uygulanmıştır. Örneğin S1 sınıfındaki hastaları diğer üç sınıftan ayırmak için, yine iki sınıftan oluşan bir model oluşturulur. Bu model S1 sınıfındaki hastaları pozitif örnek olarak alırken; S2, S3 ve S4 sınıfından gelen tüm hastaları da negatif örnek olarak kabul eder. Dolayısıyla dört sınıftan oluşan bir veri setinde, bu şekilde dört ayrı ikili sınıflandırma modeli oluşturulur. Tüm bu doğruluk yüzdeleri ve AUC hesaplamaları 500-katlı çapraz doğrulama ile tekrarlanır; sonuç olarak ortalama doğruluk yüzdesi, ortalama AUC değeri ve ROC eğrileri rapor edilir.

3. Bulgular

Akciğer kanseri için alt-türü bilinen toplam 112 hasta bulunmaktadır. Bu hastalara ait toplam 20218 mRNA gen ifadesi ve 186 protein seviye verisi vardır. Over kanserinde alt-türü bilinen toplam 202 hasta bulunmaktadır. Bu hastalara ait toplam 20131 mRNA gen ifadesi ve 175 protein seviye verisi vardır. Sınıflandırma algoritmaları dört kanser alt-türünü birbirinden ayırt edebilmek için birbirinden bağımsız olarak eğitilmiştir. Burada iki sınıftan fazla sayıda sınıf için öngörü yapılacağı için, "one against all" (OAA) yöntemi uygulanarak ortalama doğruluk yüzdeleri ve ortalama AUC değerleri hesaplanmıştır.

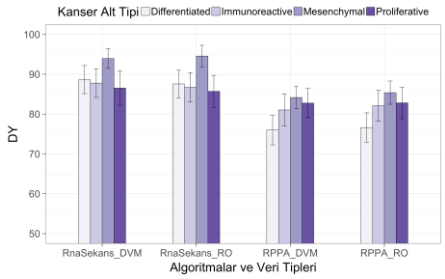
Akciğer kanseri hastalarının alt-türlerine göre sınıflandırılmasında, mRNA gen ekspresyon verisi içinden toplam 50-öznitelik seçilmiş, protein verisi içinden ise 40-öznitelik seçilerek algoritmalar eğitilmiştir. Şekil 2 algoritmaların en anlamlı RNA-sekanslama ve RPPA öz-nitelikleriyle eğitildiğinde ortaya çıkan doğruluk yüzdeleri sonuçlarını özetlemektedir. Buna göre, mRNA ekspresyon verisinden seçilen öz-niteliklerle makine öğrenmesi algoritmaları eğitildiğinde, protein seviyesi verilerinden seçilen öz-niteliklere göre, farklı alt-türleri %3,3 ile %19,7 arasında değişen bir aralıkta daha yüksek doğruluk yüzdeleri ile sınıflandırmıştır. Şekil A1 RNA-sekanslama verileri ile sınıflandırma yapıldığında, ROC eğrileri ve ortalama AUC değerleri ile elde edilen sonuçları göstermektedir. Benzer şekilde Şekil A2 ise RPPA verileriyle elde edilen ROC eğrileri ve ortalama AUC değerlerini göstermektedir. DVM ve RO algoritmalarının RNA-sekanslama verisinden seçilen öz-niteliklerle eğitildiğinde, RPPA'dan seçilenlere göre, farklı alt-türler için %12,5 ile %27 arasında değişen daha yüksek AUC değerleri ile sınıflandırma yaptıkları görülmüştür.



Şekil 2. DVM ve RO algoritmalarının akciğer kanserinin alt-türlerini sınıflandırma performansları

Bu sonuçlardan yola çıkarak, uygulanan öz-nitelik seçme yönteminin ve farklı veri kaynaklarının, akciğer kanserinin alt-türlerini sınıflandırma problemine değişik oranlarda katkı sağladığı görülmektedir. İki öğrenme algoritmasının (DVM ve RO) ortalama doğruluk yüzdeleri ve AUC değerleri birbirine oldukça yakındır.

Kullanılan veri tipleri incelendiğinde ise, RNA-sekanslama verisinin tüm akciğer kanseri alt-türlerini sınıflandırmada, RPPA verisine göre daha başarılı sonuçlar elde ettiği gözlenmiştir. Kanser alt-türlerine bakıldığında ise, DVM algoritması en yüksek doğruluk yüzdesini (%91) ve ortalama AUC değerini (%89) "Classical" alt-türü için RNA-sekanslama verisini kullanarak elde etmiştir. Öte yandan RO algoritması, en yüksek değerleri "Classical" (%89 doğruluk yüzdesi, %87 ortalama AUC değeri) ve "Secretory" (%89 doğruluk yüzdesi, %83 ortalama AUC değeri) alt-türleri için yine RNA-sekanslama verisini kullanarak elde etmiştir.



Şekil 3. DVM ve RO algoritmalarının over kanserinin alt-türlerini sınıflandırma performansları

Over kanseri hastalarının alt-türlerine göre sınıflandırılmasında, mRNA verisi içinden 50-öznitelik alınmış, protein verisi içinden ise 10-öznitelik seçilerek makine öğrenmesi algoritmaları eğitilmiştir. Şekil 3 algoritmaların en anlamlı RNA-sekanslama ve RPPA öznitelikleriyle eğitildiğinde over kanseri için elde ettikleri doğruluk yüzdeleri sonuçlarını göstermektedir. Algoritmaların RNA-sekanslama verisinden seçilen öz-niteliklerle eğitildiğinde, protein verilerinden seçilenlere göre, farklı alt-türleri %3,3 ile %11,8 arasında değişen bir aralıkta daha yüksek doğruluk yüzdeleri ile sınıflandırdıkları gözlenmiştir. Şekil A3 RNA-sekanslama verileri ile sınıflandırma yapıldığında, ROC eğrileri ve ortalama AUC değerleri ile elde edilen sonuçları göstermektedir. Şekil A4 ise RPPA verileriyle elde edilen ROC eğrileri ve ortalama AUC değerlerini göstermektedir. DVM ve RO algoritmalarının RNA-sekanslama verisinden seçilen öz-niteliklerle eğitildiğinde, RPPA'dan seçilenlere göre, farklı alt-türler için %7,5 ile %19,5 arasında değişen daha yüksek AUC değerleri ile sınıflandırma yaptıkları görülmüştür.

Over kanseri alt-türlerini sınıflandırma sonuçları incelediğinde, uygulanan öznitelik seçme yönteminin bu sınıflandırma problemine ek bir katkı sağladığı söylenebilir. İki öğrenme algoritmasının ortalama doğruluk yüzdeleri ve AUC değerleri birbirine yakındır. Kullanılan veri tipleri incelendiğinde, RNA-sekanslama verisinin over kanserinin tüm alt-türlerinin sınıflandırma probleminde protein seviyesi verisine göre daha başarılı bir performans gösterdiği söylenebilir. Kanser alt-türlerine bakıldığında ise, kullanılan makine öğrenmesi algoritmasından bağımsız olarak özellikle RNA-sekanslama verisi için "Mesenchymal" alt-türünün daha başarılı (ortalama %94 doğruluk yüzdesi, %87 AUC değeri) bir şekilde diğer alt-türlerden ayırt edildiği görülmüştür.

4. Tartışma ve Sonuç

Son yirmi yılda makina öğrenmesi yöntemleri birçok farklı alanda uygulanırken [22,23] son on yılda Destek Vektör Makinaları [24] ve Rastgele Orman [25] öğrenme algoritmaları biyolojik verilerin analizinde de sıklıkla kullanılmaya başlanmıştır.

Sunulan bu çalışmada ise, akciğer ve over kanserlerinin alt-türlerini birbirinden ayırt etmeyi sağlayacak ve klinik tanı kitlerinde kullanılabilecek

sınırlı sayıda biyoişaret olabilecek mRNA ya da protein tespit edilmeye çalışılmıştır. mRNA ekspresyon ölçümlerini sunan RNA-sekanslama ve RPPA verileri karşılaştırıldığında, RNA-sekanslama verisi her iki kanser türünde de daha başarılı sınıflandırma sonuçları sağlamıştır. DVM ve RO öğrenme algoritmalarının sınıflandırma performansları birbirine oldukça yakındır. Öğrenme algoritmaları mRNA ekspresyon verisinden elde edilen öz-nitelikleri kullandıklarında; akciğer hastalarının kanser alt-türleri ortalama %87 ile %91 arasında değişen bir doğruluk yüzdesiyle sınıflandırılırken, over kanseri içinse %87 ile %95 arasında değişen bir başarıya ulaşmışlardır. ROC analizi sonuçları detaylıca incelendiğinde, akciğer kanserinin alt-türlerinin tespiti için protein seviyesi verisi kullanıldığında özellikle "Primitive" ve "Secretory" alt-türleri için AUC değerlerinin, rastgele tahmin kabul edilen 0.5 değerine yakın sonuçlar verdiği görülmüştür. Özellikle bu alt-türleri daha başarılı sınıflandırabilmek için protein seviyesi yerine mRNA ekspresyon verisi tercih edilmesi, daha tutarlı sonuçlar elde edilmesini sağlayacaktır.

Bu çalışmada kullanılan akciğer kanseri örnekleri skuamöz hücreli akciğer kanseri hastalarına ait olduğundan, literatürde bu tür ile ilgili yapılan çalışmalar incelenmiştir. Birçok çalışma akciğer kanserinin alt-türlerini (Basal, Classical, Secretory, Primitive) kendi oluşturdukları hasta grupları üzerinde farklı kümeleme yöntemleriyle tespit etmekte ve bu alt-türlerin hastaların yaşam sürelerine olan etkilerini araştırmaktadır [26-28]. Wilkerson ve arkadaşlarının yaptığı çalışmada "Consensus" kümeleme yöntemiyle 382 akciğer kanseri örneği içinden dört alt-türü ayırt edilebilen 208 genin mRNA ölçümlerinin oluşturduğu öznitelik vektörünün, tek çıkarımlı çapraz doğrulama ("one-leave-out cross validation") yöntemiyle ortalama %96 doğruluk derecesiyle sınıflandırma yapabildiği belirtilmiştir [29]. Makalede sunulan çalışma ile literatürdeki çalışmanın [29] sonuçları karşılaştırıldığında, DVM ve RO algoritmalarının tüm alt-türler üzerindeki ortalama performansının %89 doğruluk derecesine ulaştığı görülmektedir. İki çalışmada kullanılan hasta gruplarının, biyolojik deney yönteminin, makine öğrenmesi yöntemlerinin, çapraz doğrulama tekniğinin ve öznitelik sayılarının farklı olduğu düşünüldüğünde; sunulan çalışmanın akciğer kanseri için RNA-sekanslama verilerini kullanarak literatürdeki çalışmalarla benzer başarımlara ulaşabileceği söylenebilir.

Literatürdeki over kanseriyle ilgili çalışmalarda da dört alt-tür kümeleme yöntemleri ve istatiksel analizlerle tespit edilmiş; bu alt-türlerin hastaların yaşam sürelerine olan etkileri ve farklı çalışmalarda belirlenen alt-türlerin ne ölçüde örtüştüğü incelenmiştir [30-34]. Leong ve arkadaşları farklı deney platformlarını farklı hücre türleriyle (taze dondurulan / FFPE) kullanmışlardır [35]. Bu makalede sunulan dört farklı over alt-türünü, taze dondurulan hücrelerde (TCGA'da da bu hücre örnekleri kullanılmıştır) Leong ve arkadaşları %88-100 arasında değişen ortalama doğruluk dereceleriyle sınıflandırmıştır. Çalışmalarında 42 ile 48 arasında değişen sayıdaki genin ölçümlerini, ilk defa mikro-array tabanlı olmayan bir deney yöntemiyle ölçerek farklı bir deneysel karşılaştırma yapmışlardır. Makalede sunulan çalışma ile literatürdeki bu çalışmanın [35] sonuçları karşılaştırıldığında, bazı deney platformları (Taqman ve Fluidigm) için RNA-sekanlasma ile elde edilen sınıflandırma başarımlarının (DVM ve RO algoritmalarının ortalama %89 doğruluk derecesi) aynı düzeyde

olduğu görülmektedir. Illumina ve Nanostring teknolojileri kullanılarak %96 ile %100 arasında hesaplanan doğruluk derecelerinin ise, bu yöntemlerin sınanmasında kullanılan örnek sayısının oldukça az sayıda (17 ve 28 örnek) olmasından kaynaklandığı düşünülmektedir. İki çalışmada kullanılan hasta gruplarının, örnek sayılarının, deney platformlarının oldukça farklı olması nedeniyle, sunulan çalışmanın over kanseri için RNA-sekanslama verilerini kullanarak literatürdeki çalışmalardan daha yüksek başarımlara ulaşabilme potansiyeli olduğu düşülmektedir.

Bundan sonraki araştırmalarda, akciğer ve over kanserlerinin alt-türlerinin tedavisi için geliştirilecek olan yeni ilaçların, bu çalışmada bulunan yeni biyoişaretleri de dikkate alarak geliştirilmesi, uygulanacak tedavinin etkisinin artırılmasına yardımcı olabilir. Gelecekte bu çalışmanın devamı olarak, tespit edilen yeni biyoişaretlerin geliştirilecek kanser tanı kitlerine dahil edilebilmesi için; yeni alınan RNA-sekanslama hasta verileri üzerinde işlemsel analizlerinin yapılması, sonrasında ise hücre deneyleri ile tasdikinin sağlanması amaçlanmaktadır.

Teşekkür

Bu araştırma TÜBİTAK tarafından 115C012 nolu proje kapsamında desteklenmiştir.

Kaynakça

[1] van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H. 2002. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, Nature, Cilt. 415, s. 530-536. DOI: 10.1038/415530a

[2] Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T., Hiller, W., Fisher, E.R., Wickerham, D.L., Bryant, J., Wolmark, N. 2004. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer, N Engl J Med, Cilt. 351, s. 2817-2826. DOI: 10.1056/NEJMoa041588

[3] Zhang, Z., Huang, K., Gu, C., Zhao, L., Wang, N., Wang, X., Zhao, D., Zhang, C., Lu, Y., Meng, Y. 2016. Molecular Subtyping of Serous Ovarian Cancer Based on Multi-Omics Data, Sci Rep, Cilt. 6, s. 26001. DOI: 10.1038/srep26001

[4] Chen, F., Zhang, Y., Şenbabaoğlu, Y., Ciriello, G., Yang, L., Reznik, E., Shuch, B., Micevic, G., De Velasco, G., Shinbrot, E., Noble, M.S., Lu, Y., Covington, K.R., Xi, L., Drummond, J.A., Muzny, D., Kang, H., Lee, J., Tamboli, P., Reuter, V., Shelley, C.S., Kaipparettu, B.A., Bottaro, D.P., Godwin, A.K., Gibbs, R.A., Getz, G., Kucherlapati, R., Park, P.J., Sander, C., Henske, E.P., Zhou, J.H., Kwiatkowski, D.J., Ho, T.H., Choueiri, T.K., Hsieh, J.J., Akbani, R., Mills, G.B., Hakimi, A.A., Wheeler, D.A., Creighton, C.J. 2016. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma, Cell Rep, Cilt. 14, Sayı. 10, s. 2476–2489. DOI: 10.1016/j.celrep.2016.02.024

[5] Wang, M., Klevebring, D., Lindberg, J., Czene, K., Grönberg, H., Rantalainen, M. 2016. Determining Breast Cancer Histological Grade From RNA-Sequencing Data, Breast Cancer Res, Cilt. 18, Sayı. 1, s. 48. DOI: 10.1186/s13058-016-0710-8

[6] French, C.L., Ye, F., Revetta, F., Zhang, B., Coffey, R.J., Washington, M.K., Deane, N.G., Beauchamp, R.D., Weaver, A.M. 2015. Linking Patient Outcome to High Throughput Protein Expression Data Identifies

Novel Regulators of Colorectal Adenocarcinoma Aggressiveness, F1000Research, Cilt. 4, s. 99. DOI: 10.12688/f1000research.6388.1

[7] Cancer Genome Atlas Research Network. 2014. Comprehensive Molecular Profiling of Lung Adenocarcinoma, Nature, Cilt. 511, Sayı. 7511, s. 543–550. DOI: 10.1038/nature13385

[8] Hung, F.H., Chiu, H.W. 2017. Cancer subtype prediction from a pathway-level perspective by using a support vector machine based on integrated gene expression and protein network, Computer Methods and Programs in Biomedicine, Cilt. 141, Sayı. Supplement C, s. 27-34. DOI: 10.1006/j.cmpb.2017.01.006

[9] Kosti, I., Jain, N., Aran, D., Butte, A.J., Sirota, M. 2016. Cross-tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues, Sci Rep, Cilt. 6, Sayı. 24799, DOI: 10.1038/srep24799

[10] Cancer Genome Atlas Research Network. 2017. Integrated genomic and molecular characterization of cervical cancer, Nature, Cilt. 543, s. 378–384. DOI: 10.1038/nature21386

[11] Gao, S., Qiu, Z., Song, Y., Mo, C., Tan, W., Chen, Q., Liu, D., Chen, M., Zhou H. 2017. Unsupervised clustering reveals new prostate cancer subtypes, Translational Cancer Research, Cilt. 6, Sayı. 3, DOI: 10.21037/tcr.2017.05.15

[12] Danaee, P., Ghaein, R., Hendrix, D.A. 2017. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. Proceedings of the Pacific Symposium, 5-8 Ocak, Hawaii-USA, 219-229.

[13] The Cancer Genome Atlas. 2007. https://cancergenome.nih.gov/ (Erişim tarihi: 07.07.2016).

[14] Özkan, S., Keskinkılıç, B., Gültekin, M., Karaca, AS., Öztürk, C., Boztaş, G. 2014. T.C. Sağlık Bakanlığı Türkiye Halk Sağlığı Kurumu. Ulusal kanser kontrol planı 2013-2018. Sağlık Bakanlığı Yayınları.

[15] Cancer Genome Atlas Research Network. 2011. Integrated Genomic Analyses of Ovarian Carcinoma, Nature, Cilt. 474, Sayı. 7353, s. 609–615. DOI: 10.1038/nature10166

[16] Ihaka, R., Gentleman, R. 1996. R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics, Cilt. 5, Sayı 3, s. 299-314. DOI: 10.1080/10618600.1996.10474713

[17] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B. 2008. Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq, Nat Methods, Cilt. 5, Sayı. 7, s. 621-628. DOI: 10.1038/nmeth.1226

[18] Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., Dewey, C.N. 2010. RNA-Seq Gene Expression Estimation With Read Mapping Uncertainty, Bioinformatics, Cilt. 26, Sayı. 4, s. 493-500. DOI: 10.1093/bioinformatics/btp692

[19] Cortes, C. Vapnik, V. 1995. Support Vector Networks, Machine Learning, Cilt. 20, s. 1-25. DOI: 10.1023/A:1022627411411

[20] Ho, T.K. 1995. Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282.

[21] Bradley, A.P. 1997. The Use of The Area Under The ROC Curve in

The Evaluation of Machine Learning Algorithms, Pattern Recogn, Cilt. 30, Sayı. 7, s. 1145-1159. DOI: 10.1016/S0031-3203(96)00142-2

[22] Maglogiannis, I.G., Karpouzis, K., Wallace, B.A., Soldatos, J. 2007. Emerging Artificial Intelligence Applications in Computer Engineering : Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies, 1st edition, IOS Press, Washington, DC, 408s.

[23] Dietterich, T.G. 2000. Ensemble Methods in Machine Learning, The First International Workshop on Multiple Classifier Systems, 21-23 Haziran, İtalya, 1-15.

[24] Cyran, K.A., Kawulok, J., Kawulok, M., Stawarz, M., Michalak, M., Pietrowska, M., Polańska, J. 2013. Support Vector Machines in Biomedical and Biometrical Applications. s. 379–417. Ramanna, S., Jain, L.C., Howlett, R.J. ed. 2013. Emerging Paradigms in Machine Learning, Springer Berlin Heidelberg, Germany, 498s.

[25] Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I. R. 2012. Overview of Random Forest Methodology and Practical Guidance With Emphasis on Computational Biology and Bioinformatics, WIREs Data Mining Knowl Discov, Cilt. 2, s. 493–507. DOI: 10.1002/widm.1072

[26] Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J.M., Macdonald, J., Thomas, D., Moskaluk, C., Wang, Y., Beer, D.G. 2006. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res. Cilt. 66, Sayı. 15, s. 7466-72. DOI: 10.1158/0008-5472.CAN-06-1191

[27] Lee, H.J., Lee, J.J., Song, I.H., Park, I.H., Kang, J., Yu, J.H., Ahn, J.H., Gong, G. 2015. Prognostic and predictive value of NanoString-based immune-related gene signatures in a neoadjuvant setting of triple-negative breast cancer: relationship to tumor-infiltrating lymphocytes, Breast Cancer Res Treat, Cilt. 151, Sayı. 3. S. 619-627. DOI: 10.1007/s10549-015-3438-8

[28] Faruki, H., Mayhew, G.M., Serody, J.S., Hayes, D.N., Perou, C.M., Lai-Golgman, M. 2017. Lung Adenocarcinoma and Squamous Cell Carcinoma Gene Expression Subtypes Demonstrate Significant Differences in Tumor Immune Landscape, J of Thoracic Onc, Cilt. 12, Sayı. 6 s. 943-953. DOI: 10.1016/j.jtho.2017.03.010

[29] Wilkerson, M.D., Yin, X., Hoadley, K.A., Liu, Y., Hayward, M.C., Cabanski, C.R., Muldrew, K., Miller, C.R., Randell, S.H., Socinski, M.A., Parsons, A.M., Funkhouser, W.K., Lee, C.B., Roberts, P.J., Thorne, L., Bernard, P.S., Perou, C.M., Hayes, D.N. 2010. Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types, Clin Cancer Res. Cilt. 16, Sayı. 19, s. 4864-4875. DOI: 10.1158/1078-0432.CCR-10-0199

[30] Verhaak, R.G., Tamayo, P., Yang, J.Y., Hubbard, D., Zhang, H., Creighton, C.J., Fereday, S., Lawrence, M., Carter, S.L., Mermel, C.H., Kostic, A.D., Etemadmoghadam, D., Saksena, G., Cibulskis, K., Duraisamy, S., Levanon, K., Sougnez, C., Tsherniak, A., Gomez, S., Onofrio, R., Gabriel, S., Chin, L., Zhang, N., Spellman, P.T., Zhang, Y., Akbani, R., Hoadley, K.A., Kahn, A. Kobel, M., Huntsman, D., Soslow, R.A., Defazio, A., Birrer, M.J.,

Gray, J.W., Weinstein, J.N., Bowtell, D.D., Drapkin, R., Mesirov, J.P., Getz, G., Levine, D.A., Meyerson, M., Cancer Genome Atlas Research Network. 2013. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma, J. Clin. Invest., Cilt. 123, Sayı. 1, s. 517-525. DOI: 10.1172/JCI65833

[31] Tan, T. Z., Miow, Q. H., Huang, R. Y.-J., Wong, M. K., Ye, J., Lau, J. A., Wu, M.C., Hadi, L.H.B.A., Soong, R., Choolani, M., Davidson, B., Nesland J.M., Wang, L.Z., Matsumura, N., Mandai, M., Konishi, I., Goh, B.C., Chang, J.T., Thiery, J.P., Mori, S. 2013. Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. EMBO Molecular Medicine, Cilt. 5, Sayı. 7, s. 983–998. DOI: 10.1002/emmm.201201823

[32] Sfakianos, G.P., Iversen, E.S., Whitaker, R., Akushevich, L., Schildkraut, J.M., Murphy, S.K., Marks, J.R., Berchuck, A. 2013. Validation of ovarian cancer gene expression signatures for survival and subtype in formalin fixed paraffin embedded tissues, Gynecologic Oncology, Cilt. 129, Sayı. 1, s. 159-164. DOI: 10.1016/j.ygyno.2012.12.030

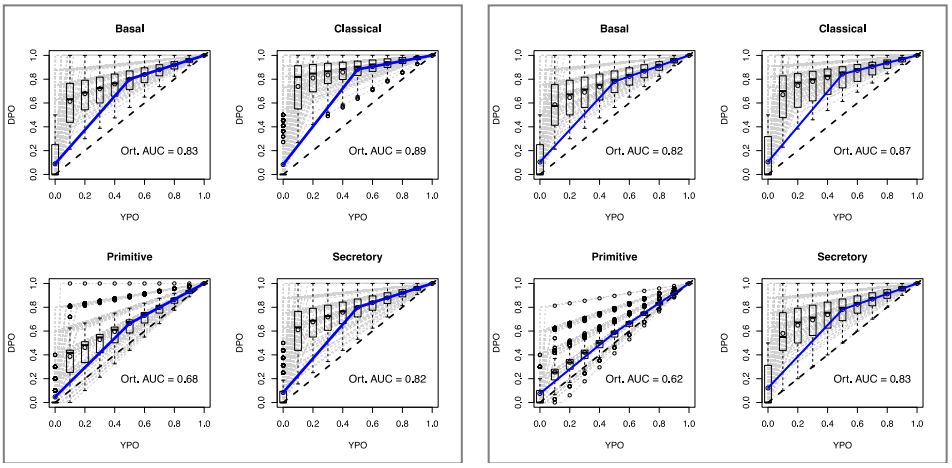
[33] Konecny, G.E., Wang, C., Hamidi, H., Winterhoff, B., Kalli, K.R., Dering, J., Ginther, C., Chen, H.W., Dowdy, S., Cliby, W., Gostout, B., Podratz, K.C., Keeney, G., Wang, H.J., Hartmann, L.C., Slamon, D.J., Goode, E.L. 2014. Prognostic and Therapeutic Relevance of Molecular Subtypes in High-Grade Serous Ovarian Cancer, Journal of the National Cancer Institute, Cilt. 106, Sayı. 10. DOI: 10.1093/jnci/dju249

[34] Way, G.P., Rudd, J., Wang, C., Hamidi, H., Fridley, B.L., Konecny, G.E., Goode E.L., Greene, C.S., Doherty, J.A. 2016. Comprehensive Cross-Population Analysis of High-Grade Serous Ovarian Cancer Supports No More Than Three Subtypes, G3: Genes, Genomes, Genetics, Cilt. 6, Sayı. 12, s. 4097-4103. DOI: 10.1534/g3.116.033514

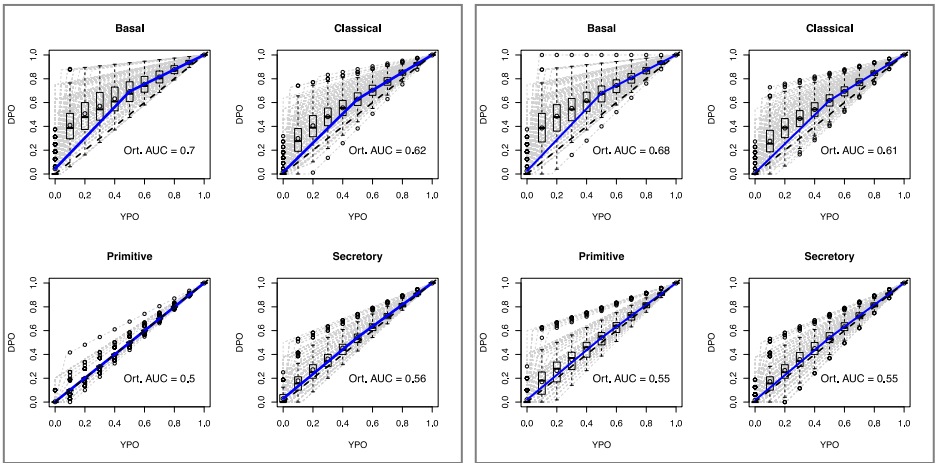
[35] Leong, H. S., Galletta, L., Etemadmoghadam, D., George, J., The Australian Ovarian Cancer Study, Köbel, M., Ramus, S. J., Bowtell, D. 2015. Efficient molecular subtype classification of high-grade serous ovarian cancer. J. Pathol., Cilt. 236, s. 272–277. DOI:10.1002/path.4536

Ekler

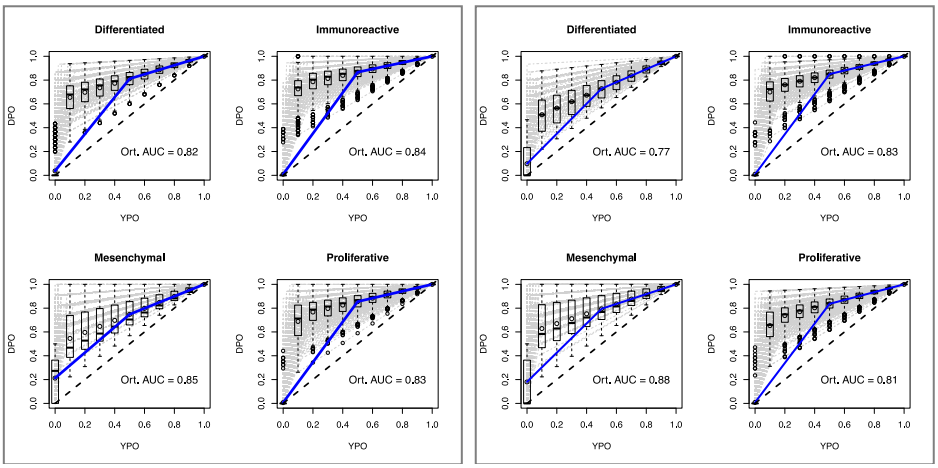
Ek A. ROC analizi sonuçları



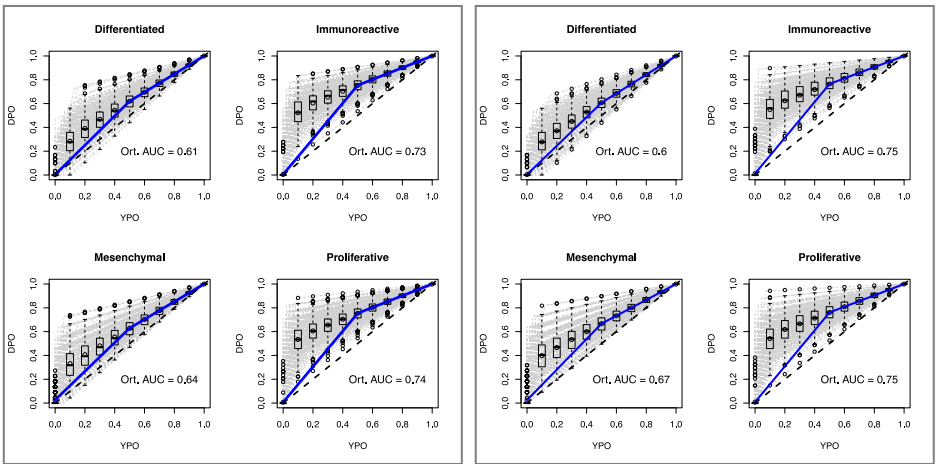
Şekil A1. DVM ve RO algoritmalarının akciğer kanserinin alt-türlerini sınıflandırma performanslarının RNA-sekanslama verisi üzerinde, ROC eğrileri ve ortalama AUC ile hesaplanması



Şekil A2. DVM ve RO algoritmalarının akciğer kanserinin alt-türlerini sınıflandırma performanslarının RPPA verisi üzerinde, ROC eğrileri ve ortalama AUC ile hesaplanması



Şekil A3. DVM ve RO algoritmalarının over kanserinin alt-türlerini sınıflandırma performanslarının RNA-sekanslama verisi üzerinde, ROC eğrileri ve ortalama AUC ile hesaplanması



Şekil A4. DVM ve RO algoritmalarının over kanserinin alt-türlerini sınıflandırma performanslarının RPPA verisi üzerinde, ROC eğrileri ve ortalama AUC ile hesaplanması