DÜMF Mühendislik Dergisi

web: http://dergipark.gov.tr/dumf

Evrişimsel sinir ağları kullanılarak normal ve göğüs kanseri hücreleri içeren genomların sınıflandırılması

Suat TORAMAN*
Fırat Üniversitesi, Enformatik Bölümü, Elazığ
storaman@firat.edu.tr ORCID: 0000-0002-7568-4131, Tel: (424) 247 00 00 (3152)

Bihter DAŞ
Fırat Üniversitesi, Yazılım Mühendisliği, Elazığ
bihterdas@gmail.com.tr ORCID: 0000-0002-2498-3297

Geliş: 26.08.2019, Revizyon: 01.10.2019, Kabul Tarihi: 29.11.2019

Öz

Geniş veri setlerinden anlamlı ve doğru bilgilerin çıkarılması biyoinformatik çalışmalarında önemli bir unsurdur. Karşılaşılan en önemli zorluklardan biri, kanser ile ilişkili olan genomik işaretçilerin tespitidir. Bu problemin çözümü için kullanılan genom dizilimlerinin sayısallaştırılması ve dizilimlerden öznitelik çıkarımı, sorunun çözümünde oldukça etkilidir. DNA dizilimlerinin sayısallaştırılması için literatürde var olan çeşitli yöntemler kullanılmaktadır. Öznitelik çıkarımında da, önceki çalışmalarda, belirli istatistiksel parametreler hesaplanmakta ve bu parametreler üzerinden bir ayrım gerçekleştirilmektedir. Ayrıca, hesaplanan parametreler uzmanın tecrübesine dayalı olarak seçilmektedir. Bu çalışmada önerilen yaklaşımda ise, yeni bir haritalama yöntemi olan Entropi tabanlı sayısal haritalama ile DNA dizilimleri sayısal sinyallere dönüştürülmüş ve daha sonra sayısallaştırılan DNA dizilimlerinden Evrişimsel Sinir Ağları (ESA) kullanılarak öznitelik çıkarımı yapılmıştır. ESA modelleri kullanarak yapılan öznitelik çıkarma işleminde sistem, veriden kendisi öznitelik çıkarmaktadır. Daha sonra ESA modellerinden elde edilen öznitelikler Destek Vektör Makinesi (DVM) ve k-En yakın komşu algoritması (k-NN) ile sınıflandırılmıştır. Bu çalışmada, yukarıda bahsedilen her iki yaklaşım kullanılarak DNA dizilerinden göğüs kanseri ve sağlıklı gen dizilimi gruplarının sınıflandırması için yeni bir yöntem önerilmektedir. Önerilen yöntem ile ulaşılan sınıflandırma doğruluğu %85.97'dir. Elde edilen sonuçlar, derin öğrenmenin genom analizinde genlerin sınıflandırılması, yeni genlerin bulunması gibi uygulamalarda etkili bir yöntem olabileceğini göstermektedir.

Anahtar Kelimeler *—DNA, Genom analizi, Kanser, Evrişimsel sinir ağları, Sınıflandırma*

* Yazışmaların yapılacağı yazar

DOI: 10.24012/dumf.610879

Giriş

Kanser, DNA'nın kimyasal ve yapısal değişimini içeren genetik ve epigenetik bir hastalık olarak tanımlanır. Bu hastalık tüm dünyanın karşı karşıya kaldığı en tehlikeli hastalıklardan biridir ve son zamanlarda kanser yüzünden ölüm oranları sürekli artmaktadır (Cheon vd., 2017). Bu nedenle bu hastalığın erken aşamada tespit edilmesi çok önemlidir. DNA'da meydana gelen mutasyonlar, belirli tür kanserlerin gelişme riskini önemli ölçüde arttırabilir. Protein ve DNA sekansları analiz edilerek, mevcut ve gelecekteki kanser olasılığı tahmin edilebilir. Biyoinformatik çalışmalarda karşılaşılan zorluklardan biri, kanserle ilişkili olduğu düşünülen genomik işaretçilerin tespit edilmesidir. Genomun kanser gelişimindeki rolünü tanımlayan erken dönem çalışmaları 19. yüzyılın sonlarına ve 20. yüzyılın başlarına dayanmaktadır. David Von Hansemann ve Theodor Boveri, kanser hücrelerinin mikroskop altında bölünmesini inceleyip, tuhaf kromozomal sapmaların varlığını gözlemlemişlerdir (Meng vd., 2013), (Stratton vd., 2009). Bu bulgular, kanserin kromozomların yapısında neden olduğu değişikliklerle ilgili olduğu tespit edilmiştir. Kwong-Sak Leung ve diğ., karaciğer kanseri olan hastalar ile sağlıklı bireyleri karşılaştırarak karaciğer kanseri ile ilişkili olan Hepatit B virüsündeki (HBV) genomik işaretçilerin tespitine yönelik çalışmışlardır. Bu çalışmada 200'den fazla hastadan Genotip B ve C grubu HBV DNA sekansları toplanmıştır. Evrimsel algoritmaya dayanan kural öğrenme adlı algoritma ile anlamlı kurallar çıkarılmış ve doğrusal olmayan integral tarafından yeni bir sınıflandırma yöntemi geliştirilmiştir. Bu yöntemin performansının karaciğer kanseri teşhisinde veri kümesi için %70'ten fazla doğruluğa ve %80 duyarlılığa sahip olduğu görülmüştür (Leung vd., 2011). Dayana binti Saiful Nurdin A ve arkadaşının yaptığı çalışmada, 11 ekzon DNA sekansı kullanılarak Yeni Nesil Dizilimi (YND) yöntemiyle göğüs kanseri tespiti için bir yaklaşım önerilmiştir. Çalışmada BRCA1 geninden 11 ekzon DNA sekansı kullanılmıştır. Mutasyona uğramış 11 ekzon sekansı MATLAB ve SSEARCH35 yazılımları kullanılarak hizalanmıştır. Çalışmada YND ile 11 ekzon sekansının göğüs kanserinin tespitinde kullanılması sonucu %99.4 doğruluk elde edilmiştir (Bordoloi vd., 2018). Naeem ve ark yaptıkları çalışmada Genomik sinyal işleme tekniklerinden olan ayrık Fourier Dönüşümü, güç spektral yoğunluğu ve Welch'in ortalama periodogram yöntemi kullanılarak normal ve kanserli hücreler arasındaki farklılaşma için tatmin edici sonuçlar elde edilmiştir. Algoritma NCBI gen bankasından elde edilen altı sağlıklı ve altı kanserli meme hücresi geninde test edilmiştir (Naeem vd., 2017).

Bu çalışmada ise göğüs kanseri DNA gen sekansları ve sağlıklı bireylere ait DNA gen sekansları derin öğrenme yöntemlerinden biri olan transfer öğrenme tekniği ile sınıflandırılmıştır. Literatürdeki çalışmalarda sayısallaştırılan gen dizilimlerini makine öğrenme teknikleri kullanılarak sınıflandırılmıştır. Yapılan literatür incelenmesinde Göğüs kanser geninin derin öğrenme yöntemleri ile sınıflandırıldığı herhangi bir çalışmaya rastlanılmamıştır. Göğüs kanseri ve normal DNA gen sekansları yeni bir sayısal haritalama yöntemi olan Entropi tabanlı sayısallaştırma tekniği ile sayısallaştırılmıştır. Sayısallaştırılan gen dizilimleri spektrogram görüntüsüne dönüştürülmüştür. Böylece 2 boyutlu CNN mimarilerine giriş verisi olarak verilmiştir. Her bir spektrogram görüntüsünün VGG16 ve VGG19 modelleri ile öznitelik vektörleri çıkarılmıştır. Elde edilen bu veriler DVM ile sınıflandırılmıştır.

Çalışmanın geri kalanı aşağıdaki şekilde düzenlenmiştir. Materyal ve metot bölümünde, veri kümesi, ESA modelleri, sınıflandırıcı ve performans değerlendirme ölçütlerinden bahsedilmiştir. Deneysel sonuçlar bölümünde önerilen yöntem ile ilgili elde edilen bulgular verilmiş ve tartışılmıştır. Sonuç bölümünde çalışmanın genel katkıları sunulmuştur.

Materyal ve metot

Veri kümesi

Deneysel çalışmada kullanılan veri kümesi NCBI veri tabanından elde edilmiştir (NCBI Gen

bankası, 2019). Uygulama için altı sağlıklı altı göğüs kanseri genine ait DNA sekansları kullanılmıştır. Gen bankasından alınan veriler Tablo 1' de gösterilmektedir.

Tablo 1. Kullanılan veri kümesi

İndeks	Erişim Numarası
Sağlıklı Sekanslar	
1	NM_001300741.2
2	NC_000001.11
3	NC_000002.12
4	NC_000003.12
5	NC_000010.11
6	NC_000011.10
Göğüs Kanserli Sekanslar	
1	NM_139163
2	NM_001127391
3	NM_001289993
4	NR_110620
5	NC_000023
6	CH471152

Sayısal haritalama yöntemi

DNA dizilimlerinin genomik sinyal işleme uygulamalarında kullanılabilmesi için sayısal sinyallere dönüştürülmesi gerekir. Bu çalışmada daha önceden birçok genomik çalışmada kullanılan yüksek performans elde edilmiş olan Entropi Tabanlı Sayısal Haritalama Tekniği kullanılmıştır. Bu teknikte sekanslar, Fraksiyonel Shannon Entropisi kullanılarak DNA dizilimlerindeki kodonların dağılım frekansına göre sayısallaştırılmıştır (Daş, vd 2018), (Daş, 2018), (Karcı, 2016). Shannon, bir dizilimde meydana gelen kodon olasılıkları *p1, p2, p3,…… pn,* olan ve bu olasılıklardan hangisinin önce meydana geldiği bilinmeyen bir olaylar kümesini ele almıştır. *S* belirsizlik ölçüsü olarak tanımlanırsa bu belirsizlik ölçüsü *Sf* olarak Denklem 1'de tanımlanmaktadır.

$$Sf = -\sum_{i}[(-p(x_i))^{\propto} p(x_i)\log\,(p(x_i))] \qquad (1)$$

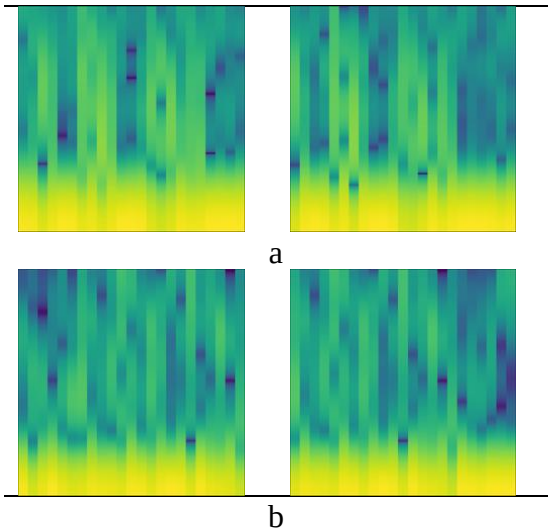
Denklem 1'deki $p(x_i)$ değeri verilen DNA dizilimindeki her bir kodonun tekrarlama sıklığını temsil eder. Alfa (α) değeri ise genellikle deneme yanılma yoluyla verilmektedir. Ancak Entropi tabanlı teknikte alfa değeri DNA verisinden çıkarılmaktadır. Alfa değeri için yeni tanımlanan formül Denklem 2'de gösterilmektedir. Alfa değeri, $p(x_i)$ değerinin logaritmasının 1'e bölünmesi olarak tanımlanmıştır.

$$\alpha = \frac{1}{\log\,(p(x_i))} \qquad (2)$$

Çalışmada kullanılan DNA sekansları Denklem 1 ve Denklem 2'e göre sayısallaştırılmıştır.

Spektrogram görüntülerin oluşturulması

Entropi tabanlı sayısal haritalama tekniği kullanılarak farklı uzunluklardaki altı göğüs kanseri gen sekansı art arda eklenmiştir. Toplam uzunluğu 6944 olan dizilim 5 birim kaydırmalı 100 birimlik bir kayan pencere yöntemi ile bölümlendiğinde 1368 adet sinyal alt kümesi elde edilmiştir. Spektrogram görüntüleri bu alt kümedeki sinyal bölümleri kullanılarak oluşturulmuştur. Spektrogram görüntüleri elde edilirken, pencere genişliği 12 ms olan Hamming pencereleme, çakışma değeri 8 ms ve Fourier dönüşümü sayısı 512 olarak belirlenmiştir. Spektrogram görüntüleri viridis renk haritası kullanılarak elde edilmiştir. Şekil 1'de Entropi tabanlı sayısallaştırma yöntemi ile elde edilen göğüs kanseri genleri ile sağlıklı genlere ait spektogram görüntüleri gösterilmektedir.



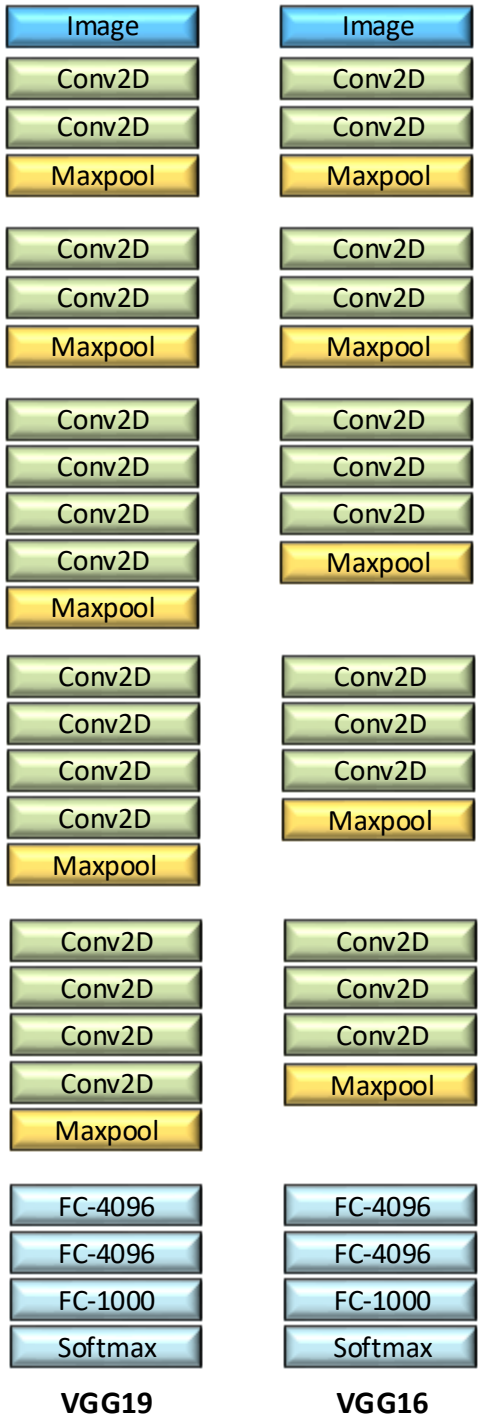
Şekil 1. a) Göğüs kanserli b) sağlıklı genlere ait spektrogram görüntüleri

Evrişimsel sinir ağları (ESA)

ESA, nesne tanıma, nesne takibi, sinyal/görüntü işleme ve sınıflandırma alanlarında başarılı sonuçlar vermesinden dolayı günümüzde en çok kullanılan derin öğrenme yöntemlerindendir (Toraman vd., 2018), (Yıldırım, 2019). Bu çalışmada Entropi tabanlı sayısal haritalama yöntemi ile oluşturulan göğüs kanseri genleri ile sağlıklı genlere ait verilerinin sınıflandırılması için kullanılmıştır. Çalışmada için kullanılan ESA modelleri VGG16 ve VGG19'dur. Bir ESA modelinin eğitilmesi için büyük miktarda veri ve aynı zamanda iyi bir donanıma sahip alt yapıya ihtiyaç duyulmaktadır. Sahip olunan donanım özeliklerine göre verilerin işlenmesi için, saatler hatta günlerce beklemek gerekebilmektedir. Eğer bir modelin eğitilmesi için gerekli olan veri bulunmuyorsa, bu durumda önceden eğitilmiş modeller kullanılabilir. Bu durum transfer öğrenme olarak adlandırılır. Transfer öğrenme ile önceden büyük veri kümeleri kullanılarak eğitilmiş bir model ile farklı bir alandaki veri kümesinden öznitelik çıkarımı yapılabilmektedir (Hasan vd., 2019). Bu çalışmada da, önceden eğitilmiş VGG16 ve VGG19 modellerinden faydalanılmıştır.

VGG16 ve VGG19

ESA'ların popüler hale gelmesinde etkili olan model AlexNet'tir. Daha sonra Oxford Üniversitesi Visual Geometry Group (VGG) tarafından VGG16 modeli geliştirilmiştir. VGG16 da konvolüsyon katmanlarında küçük filtreler (3x3) kullanılmıştır. VGG16, 13 konvolüsyon katmanı ve 3 tam bağlı katmandan oluşmaktadır. 2x2 boyutlu 5 adet havuzlama (max pooling) katmanı bulunmaktadır. Son katmanda ise softmax bulunmaktadır. Softmax katmanı ile gelen giriş verisi sınıflandırılmaktadır (Ullah vd., 2018). Aktivasyon fonksiyonu olarak ReLu kullanılmıştır. VGG19, 16 konvolüsyon katmanı ve 3 tam bağlı katmandan oluşmaktadır. VGG19'da, VGG16 gibi 5 havuzlama ve son katman olarak softmaxtan oluşmaktadır. VGG16, 138 milyon parametre içerirken, VGG19 yaklaşık 144 milyon parametre içermektedir. (Gopalakrishnan vd., 2017), (Simonyan vd., 2015) Şekil 2'de VGG16 ve VGG19'un mimarisi yapısı gösterilmektedir.



Şekil 2. VGG16 ve VGG19 ESA mimarileri

Sınıflandırma

Verilerin sınıflandırılması için DVM ve k-NN kullanılmıştır. Aşağıda her iki sınıflandırıcıya ait bilgiler ayrıntılı şekilde verilmiştir.

Destek vektör makineleri

DVM, sınıflandırma problemlerinde kullanılan ve yapısal risk azaltma ilkesine göre tasarlanmış bir yöntemdir. DVM, iki veya daha fazla sayıda sınıfı birbirinden ayıracak en uygun hiper düzlemi bulmayı amaçlamaktadır. İki sınıflı bir veri kümesinde $\{x_s, y_s\}, s = 1,2,3, \ldots n\ x_s \in R^d$ d-boyutlu uzayda eğitim verileri olsun. $y_s \in \{-1, +1\}$ ise sınıfları temsil eden etiketler olsun (Toraman vd., 2019), (Khazaee vd., 2010). En uygun hiper düzlem tanımlayan eşitsizlikler şöyle tanımlanır;

$$\begin{aligned} w.x_s + b &\geq +1, \quad y = +1 \\ w.x_s + b &\leq -1, \quad y = -1 \end{aligned} \tag{3}$$

Sonuç olarak, doğrusal olarak ayrılabilen iki sınıflı bir veri kümesi için elde edilen karar fonksiyonu şöyledir;

$$f(x) = sign\left(\sum_{s=1}^{n} y_s \alpha_s (x.x_s) + b\right) \tag{4}$$

Burada α, Lagrange çarpanı, b bias ve x_s destek vektörleri temsil etmektedir (Toraman vd., 2019), (Khazaee vd., 2010). Çalışmada kullanılan çekirdek fonksiyonları; Radial Basis Fonksiyon (RBF), Polinomial ve Lineer'dir.

k-En yakın komşu algoritması

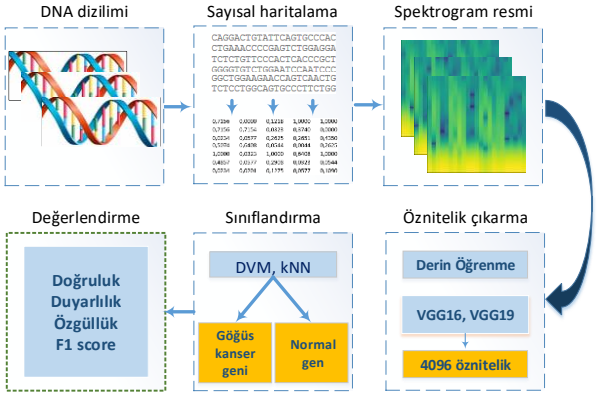
Verilerin sınıflandırması için kullanılan diğer bir sınıflandırma yöntemi k-en yakın komşu algoritması (k-NN) dir. Parametrik olmayan bir yöntem olan k-NN'de, k değeri sınıflandırılacak yeni verinin karşılaştırılacağı komşu sayısını temsil etmektedir. k-NN yönteminde, gelen yeni veri hangi gruba yakınsa o gruba dahil edilmektedir. Yakınlık ölçüsü olarak genellikle Öklid mesafesi kullanılmaktadır. Böylece, gelen yeni veri en yakın sınıfa atanır (Duda vd., 2000), (Tuncer vd., 2019)

Performans değerlendirmesi

Önerilen yöntemin performansının değerlendirebilmesi için k-katlı çapraz doğrulama yöntemi kullanılmıştır. k değeri 5 olarak seçilmiştir. Böylece, veri kümesi 5 parçaya ayrılmıştır. 4 parça eğitim için kullanılırken, kalan bir parça ise test için kullanılmıştır. Bu işlem tüm parçalar için uygulanmıştır. Performans değerlendirmesi için beş değerin ortalaması hesaplanmıştır. Performansları karşılaştırması için kullanılan parametreler aşağıdaki gibi tanımlanır; Doğru Pozitif (TP), doğru şekilde tanımlanan göğüs kanseri gen sayısını, yanlış negatif (FN), yanlış şekilde tanımlanan göğüs kanseri gen sayısını, doğru negatif (TN), doğru şekilde tanımlanan sağlıklı gen sayısı, yanlış pozitif (FP), yanlış şekilde tanımlanan sağlıklı gen sayısını göstermektedir.

$$Duyarlılık = TP/(TP + FN)x100$$
$$\text{Özgüllük} = TN/(TN + FP)x100$$
$$F1Score = 2TP/(2TP + FP + FN)$$
$$Doğruluk = \frac{TP + TN}{(TP + FP + FN + TN)}x100$$

Çalışmanın akış diyagramı Şekil 3'de gösterilmiştir.



Şekil 3. Önerilen yöntemin akış diyagramı

Deneysel sonuçlar ve tartışma

Çalışmada 6944 baz uzunluğundan göğüs kanser geni ile yine aynı uzunlukta sağlıklı gen verisi kullanılmıştır. Her iki DNA dizilimi entropi tabanlı sayısal haritalama yöntemi ile sayısallaştırılmıştır. Sayısallaştırılan dizilimler kayan pencere yöntemi kullanılarak bölümlere ayrılmıştır. Ayrıştırma işlemi sırasında %95 örtüşme ile veriler elde edilmiştir. Her bir bölüm daha sonra spektrogram görüntüsüne dönüştürülmüştür. Spektrogram görüntüleri 875 × 656 piksel boyutundadır. Bu görüntüler,

VGG16 ve VGG19 modelleri için 224x224 piksel olacak şekilde yeniden boyutlandırılmış ve her bir spektrogram görüntüsünden derin öznitelikler çıkarılmıştır. VGG16 ve VGG19 modellerinde her bir resimden 4096 boyutunda öznitelik vektörü oluşmaktadır. Daha sonra çıkarılan bu öznitelik vektörleri DVM ve k-NN kullanılarak sınıflandırılmıştır. Yapılan çalışmada, göğüs kanser geni ile sağlıklı gen verileri, iki farklı yöntem kullanılarak öznitelik çıkarımı yapılmıştır. İki farklı ESA modelinden elde edilen öznitelikler, iki farklı sınıflandırıcı yardımıyla sınıflandırılmıştır. Sınıflandırmanın daha nesnel olması için *k*-katlı çapraz doğrulama yöntemi kullanılmıştır. *k* değeri 5 olarak belirlenmiştir. Sınıflandırma işlemlerinin sonuçları Tablo 2'de gösterilmiştir.

Tablo 2. Entropi tabanlı sayısallaştırma yöntemlerinin VGG16 ve VGG19 modelleri ile gerçekleştirilen doğruluk değerleri verilmiştir. 5-katlı çapraz doğrulama uygulanmış a) DVM ve b) k-NN sınıflandırma sonuçları gösterilmektedir.

ESA modelleri	DVM	k-NN
VGG16	**85.97 ± 0.026**	**80.48 ± 0.009**
VGG19	84.03 ± 0.029	76.90 ± 0.013

Tablo 2 incelendiğinde, Entropi tabanlı sayısallaştırma yöntemi ile sayısallaştırılan verilerin her iki farklı ESA modelinde birbirlerine oldukça yakın sınıflandırma sonuçları ürettikleri görülmektedir. DVM ve k-NN sınıflandırıcıların başarımı karşılaştırıldığında ise DVM daha iyi bir sınıflandırma doğruluğuna ulaşırken, k-NN sınıflandırıcısı da ise daha düşük bir doğruluk elde edilmiştir. Çünkü DVM, genetik verilerin denetimli sınıflandırılmasında (DNA dizilimleri, protein yapı verileri, mikroarray gen ekspresyonu, vb.) en popüler araçlardan biridir. DVM, biyoinformatikteki sınıflandırma ve regresyon görevleri için k-NN ve diğer makine öğrenmesi modellerinden çok daha etkili yöntemdir. Uygulamadaki gibi kanserli ve sağlıklı veri kümesi için ikili bir sınıflandırma problemi göz önüne alındığında DVM, pozitif olanları negatif olanlardan ayıran maksimum marjlı bir hiper düzlem oluşturur. Eğer DVM iyi bir şekilde eğitilmişse, mükemmel sonuçlar üretebilir. Bununla birlikte, eğitimin kalitesi, DNA verisinin özelliğini en iyi yansıtan bir haritalama tekniğiyle sayısallaştırılmasına, verinin özellik alanına eşleştirilmesine ve problem için en uygun çekirdek fonksiyonunun seçilmesine bağlıdır.

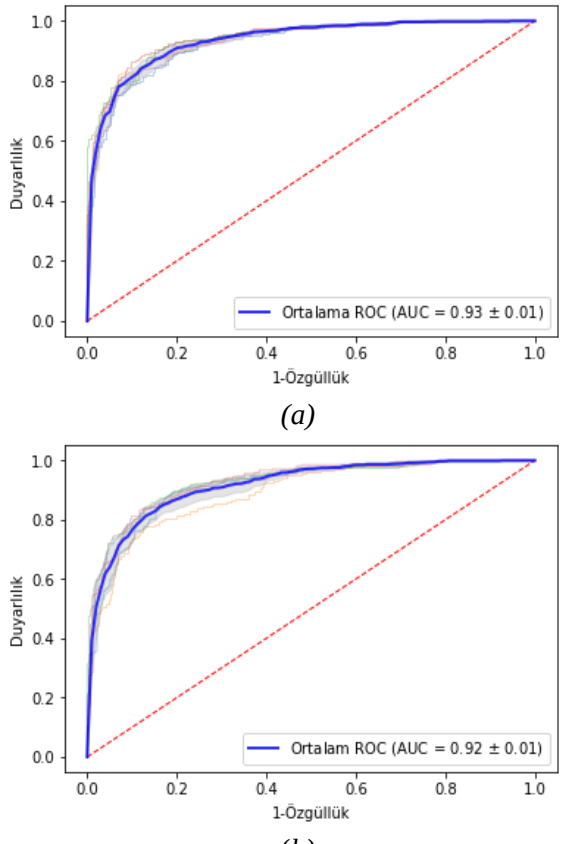
En iyi sınıflandırma başarımı %85.97 ile DVM sınıflandırıcı ve VGG16 modeli ile elde edilmiştir. VGG16 modeli VGG19 ya göre daha sığ bir modeldir. Sonuçlar incelendiğinde, modellerin derinliğin artmasının her zaman daha iyi bir sonuç vermeyeceği görülmektedir. VGG16 modelinin VGG19'a göre daha iyi bir sonuç vermesinin farklı nedenleri olabilir. Örneğin sonuçları daha genel bir pencereden irdelersek, hangi derin öğrenme modelinin hangi veri kümesinde daha iyi bir sınıflandırma sonucu vereceği verinin büyüklüğüne, verini ayırt edicilik özelliğine, verinin kalitesine, modelin parametrelerine ve modelin derinliği gibi birçok etkene bağlı olduğu söylenebilir. Ayrıca veri sayısı arttırıldığında sonuçlarda farklılaşmanın olabileceği de göz ardı edilmemelidir.

Yöntemde DNA verilerinin sayısallaştırılması için kullanılan Entropi tabanlı haritalama tekniğinin özellikle tercih edilme sebebi, bu haritalama tekniğinin DNA diziliminin karmaşık yapısını daha iyi yansıtması ve sayısallaştırmayı kodonların tekrarlama sıklığına göre yapmasındandır. Ayrıca bu haritalama tekniği gen dizisi üzerinde kodonların tekrarlama sıklık değerlerine göre geniş bir korelasyon bilgisi aralığı sunmaktadır. Spektrogram görüntülerini görsel olarak göğüs kanseri geni ve normal gen dizileri olarak ayırt etmek zordur. Önceden eğitilmiş VGG16 ve VGG19 gibi ESA modelleri, görüntülerin özelliklerini çıkararak bu görevi başarıyla gerçekleştirebilir. Entropi tabanlı haritalama tekniğiyle sayısallaştırılmış kanserli ve sağlıklı genlerin spektrogram görüntülerinin karakteristikleri daha belirgindir. Bu durum spektrogram görüntülerinin etkili özellik çıkarımında önemlidir. Sınıflandırıcının performansı da etkili özellik kümesiyle paraleldir.

Literatürde DNA dizilimlerinden karaciğer (Leung vd., 2011), göğüs (Nurdin v., 2016) ve diğer kanser türlerinin taranmasına yönelik çalışmalar mevcuttur. Bu çalışmalarda genellikle

Ayrık Fourier Dönüşümü (Chakraborty vd., 2016), Hidden Markov Modeli (Mayilvaganan vd., 2014) gibi sinyal işleme yöntemleri kullanılmaktadır ve hastalıklı gen sinyallerinde anormallikler şekilsel olarak veya baz mutasyonlarını dikkate alarak ifade edilmiştir. Fakat önerilen yöntemde segment bazlı bir sınıflandırma yapılmıştır.

Entropi tabanlı sayısal haritalama ile sayısallaştırılan verilerin VGG16 ve VGG19 modelleri kullanılarak yapılan sınıflandırma işlemi sonucu elde edilen ROC eğrileri Şekil 4'te gösterilmiştir. Testlerin doğru karar vermede gücünü göstermek için ROC eğrisi kullanılmaktadır. ROC eğrisinin altında kalan alan AUC değerini göstermektedir.



Şekil 4. Entropi tabanlı sayısal haritalama tekniği ile elde edilen verilerin 5 katlı çapraz doğrulama ile gerçekleştirilen sınıflandırma işlemine ait ROC eğrileri. a) VGG16 b) VGG19

AUC değerinin sol üst köşeye yakın olması istenen durumdur. Böylece, testin verileri ayırt etme gücünün yüksek olduğu söylenebilmektedir. AUC değeri en yüksek %93 ile VGG16 modelinde elde edilmiştir. Geliştirilen yöntem ile elde edilen sınıflandırma sonuçları bu ön çalışma için umut vaat edicidir. Geliştirilen yöntemin performans değerlendirmesi için kullanılan diğer göstergeler ise doğruluk, duyarlılık, özgüllük ve F1 skor (F1 score)'dur. Tablo 3'de hesaplanan diğer parametreler verilmiştir. Elde edilen sonuçlara göre VGG16 daha iyi bir sınıflandırma gerçekleştirilmiştir. Yapılan sınıflandırmada en yüksek doğruluğa DVM'nin RBF çekirdek fonksiyonu ile ulaşılmıştır. DVM parametrelerinden C [10^{-3},…, 10^{+3}] aralığında incelenmiş ve C=1000 olarak seçilmiştir.

K-NN sınıflandırıcı için k parametresi ise [1, .., 5] aralığında incelenmiştir. Tablo 2'de verilen k-NN sonuçları k=1 değeri için en yüksek sonuçları vermiştir. Spektrogram verileri MATLAB ortamında elde edilmiştir. Diğer işlemler Python ortamında keras kütüphanesi kullanılarak gerçekleştirilmiştir.

Tablo 3. Entropi tabanlı sayısal haritalama yöntemi kullanılarak gerçekleştirilen sınıflandırma işleminin DVM'ye göre VGG16 ve VGG19 modellerinin doğruluk, duyarlılık, özgüllük ve F1 skor değerleri (Doğ:Doğruluk, Duy: Duyarlılık, Özg: Özgüllük, F1: F1 skor)

ESA modelleri	Doğ (%)	Duy (%)	Özg (%)	F_1 (%)
VGG16	**85.97**	**84.13**	**87.79**	**85.70**
VGG19	84.03	83.04	85.01	83.86

Sonuç

Bu çalışmada göğüs kanseri genine ait DNA sekansları ile sağlıklı bireylere ait DNA sekanslarının sınıflandırmasına yönelik yeni bir yaklaşım önerilmektedir. Önerilen yöntem ile DNA sekanslarının sınıflandırma problemine farklı bir bakış açısı getirilmesi hedeflenmiştir. Bu yeni bakış açısı, DNA sekanslarının entropi tabanlı yeni bir yöntem ile sayısallaştırılması ve sayısallaştırılan DNA dizilimlerinin 100 birimlik parçalara ayrılarak spektrogram resimlerine dönüştürülerek sınıflandırılması ile desteklenmektedir. Önerilen yöntem ile spektrogram resimleri kullanılarak derin ağlar

yardımıyla etkili bir öznitelik çıkarımı yapılmıştır. Çıkarılan özniteliklerin sınıflandırma başarımı, yapılan ön çalışma için tatmin edici olduğu görülmüştür. Sınıflandırma başarımı için derin ağların daha büyük veri setleri ile eğitilmesi sistemin geçerliliği ve güvenirliğinin belirlenmesinde etkili bir adımdır. Bu nedenle, ileri dönük yapılacak olan çalışmalarda, derin ağların daha fazla veri kümesi ile eğitilmesi ve veri kümesinin artırılması planlanmaktadır. Ayrıca farklı derin öğrenme mimarilerinin de karşılaştırılması düşünülmektedir.

Kaynaklar

Bordoloi, H., Roy, D., Nirmala, S.R. (2018). A Framework for Codon Based Analysis to detect abnormalities responsible for Esophagus Cancer using Soft Computing Tool, 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), 170-174, Noida, India.

Chakraborty S., Gupta V. (2016). DWT based cancer identification using EIIP, 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), 718-723, Ghaziabad, India.

Cheon H., Son J-H. (2016) Terahertz molecular resonance of cancer DNA , Scientific Reports, vol:6, Article number:37103.

Das, B., ve Turkoglu, I. (2018). A novel numerical mapping method based on entropy for digitizing DNA sequences, Neural Comput. Appl. 29,8, 207-215.

Daş, B. (2018). Development of New Approaches Based On Signal Processing For Disease Diagnosis From Dna Sequences, PhD Thesis, Fırat University, Graduate School of Natural and Applied Sciences, Elazig, Turkey.

Duda, R.O., Hart, P.E., Stork, D.G. (2000). Pattern Classification, Second, Wiley-Interscience New York, NY, USA.

Gopalakrishnan, K., Khaitan, S.K., Choudhary, A., Agrawal, A. (2017). Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection, Constr. Build. Mater. 157, 322–330.

Hasan, M.J., Islam, M.M.M. , Kim, J.-M. (2019). Acoustic spectral imaging and transfer learning for reliable bearing fault diagnosis under variable speed conditions, Measurement, 138, 620–631.

Khazaee, A., Ebrahimzadeh, A., (2010). Classification of electrocardiogram signals with support vector machines and genetic algorithms using power spectral features, Biomed. Signal Process. Control. 5(4), 252-263.

Karci, A. (2016). Fractional order entropy: New perspectives, Optik, 127,20, 9172-9177.

Leung KS, Lee KH, Wang JF, Ng EY, Chan HL, Tsui SK, Mok TS, Tse PC, Sung JJ. (2011). Data Mining on DNA Sequences of Hepatitis B Virus, IEEE/ACM Transactions on Computational Biology And Bioinformatics, 8,2, 428-440.

Mayilvaganan M., Rajamani R. (2014), Analysis of nucleotide sequence with normal and affected cancer liver cells using Hidden Markov model, 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India.

Meng T, Soliman AT, Shyu ML, Yang Y, Chen SC, Iyengar SS, Yordy JS, Iyengar P., (2013). Wavelet Analysis in Current Cancer Genome Research: A Survey, IEEE/ACM Transactions on Computational Biology And Bioinformatics, 10, 6, 1442-1459.

Naeem SM., Mabrouk, MS., Eldosoky, MA. (2017) Detecting genetic variants of Breast cancer using different power spectrum methods, 13th International Computer Engineering Conference (ICENCO), 147-153, Cairo, Egypt.

NCBI Genbankası: https://www.ncbi.nlm.nih.gov (01 Haziran 2019).

Nurdin A. D. binti S. and Isa M. N. bin M. (2016). Development and validation of BRCA1 for Next Generation Sequencing (NGS), 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), 702-706, Kuala Lumpur, Malaysia.

Simonyan, K., Zisserman, A. (2015). Very Deep Convolutional Networks For Large-Scale Image Recognition. arXiv:1409.1556v6.

Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009). The Cancer Genome. Nature, 458,7239, 719-724.

Toraman, S., Girgin, M., Üstündağ, B., Türkoğlu, İ. (2019). Classication of the Likelihood of Colon Cancer With Machine Learning Techniques Using FTIR Signals Obtained From Plasma, Turk J Elec Eng & Comp Sci, 27, 1765-1779.

Toraman, S. ve Türkoğlu İ., (2018). Kolon Kanseri Hastaları ve Sağlıklı Kişilerin FTIR Spektrogram Görüntülerinin Derin Öğrenme ile Sınıflandırılması, Fırat Üniv. Müh. Bil. Dergisi, 30(2), 115-120.

Tuncer, SA., Akılotu, B., Toraman, S. (2019). A deep learning-based decision support system for diagnosis of osas using ptt signals, Medical Hypotheses, 127, 15-22.

Ullah, I., Hussain, M., Qazi, E.-H., Aboalsamh, H. (2018). An automated system for epilepsy detection using EEG brain signals based on deep learning approach, Expert Syst. Appl. 107, 61–71.

Yildirim, Ö. (2019). ECG Beat Detection and Classification System Using Wavelet Transform and Online Sequential ELM, Journal of Mechanics in Medicine and Biology 19, 1, 1940008.

Classification of Genomes Containing Normal and Breast Cancer Cells Using Convolutional Neural Networks

Extended Abstract

Extracting meaningful and accurate information from large data sets is an important element in bioinformatics studies. One of the most important challenges is the detection of genomic markers associated with cancer. Digitization of genome sequences used to solve this problem and feature extraction from the sequences are very effective in solving the problem. Different methods in the literature are used for digitizing DNA sequences. In feature extraction, in the previous studies, certain statistical parameters are calculated, and a distinction is made on these parameters. Moreover, the calculated parameters are selected based on the expert's experience. Furthermore, the calculated parameters are selected based on the expert's experience.

DNA sequences should be converted to digital signals to use them in genomic signal processing applications. In this study, Entropy Based Digital Mapping Technique, which is used in many previous genomic studies and has a high performance, has been used. In this technique, sequences were digitized according to the frequency of distribution of codons in DNA sequences using Fractional Shannon Entropy. The data set used for the experimental study was obtained from the NCBI database. For the application, DNA sequences belonging to six breast cancer genes and six healthy genes were used.

Since the data digitized by Entropy Based Digital Mapping Technique will be extracted features by using 2-D Convolutional Neural Network (CNN), the data should be converted to 2-D spectrogram image. While spectrogram images were obtained, it was set as Hamming window width was 12ms, overlap value was 8ms and Fourier transform number was 512. VGG16 and VGG19 CNN models were used for feature extraction from the digitized DNA sequences. The model that makes CNNs popular is AlexNet. Then, the VGG16 and VGG19 models were developed by the Oxford University Visual Geometry Group (VGG). The VGG16 consists of 13 convolution layers and 3 fully connected layers, while the VGG19 consists of 16 convolution layers and 3 fully connected layers. Both models consist of 5 max pooling and softmax as the last layer (Gopalakrishnan et al., 2017), (Simonyan et al., 2015), (Ullah et al., 2018).

In the feature extraction process using CNN models, the system extracts features from the data itself. The properties obtained in CNN models were then classified with the Support Vector Machine (SVM) and the k-Nearest neighbor algorithm (k-NN). The flow diagram of the study is shown Figure 1.

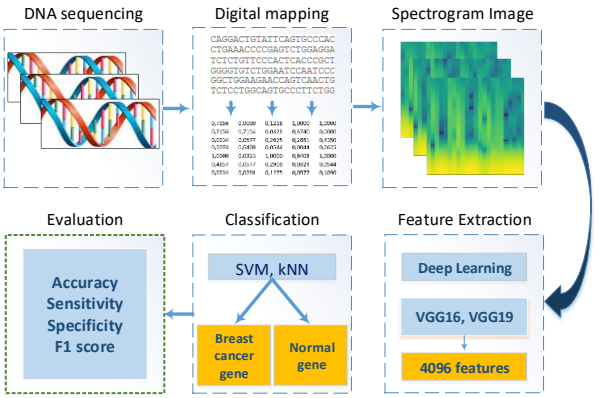


Figure 1. The flow diagram of the proposed method

In this study, the feature extraction was performed by using two different methods for breast cancer gene and healthy gene data. The properties obtained from two different CNN models were classified with the help of two different classifiers. In order to make the classification more objective, k-fold cross validation method was used. k value was determined as 5. The results of the classification procedures are shown in Table 1.

Table 1. The accuracy values of the entropy based numerical technique performed with VGG16 and VGG19 models are given. 5-fold cross-validation was performed. a) SVM, b) k-NN classification results are shown.

CNN Models	*SVM*	*K-NN*
VGG16	***85.97 ± 0.026***	***80.48 ± 0.009***
VGG19	*84.03 ± 0.029*	*76.90 ± 0.013*

In this study, a new approach is proposed for the classification of DNA sequences belonging to breast cancer gene and healthy individuals' gene. The proposed method aims to get a different perspective to the problem of classification of DNA sequences.

Keywords: *DNA, Genom Analysis, Convolutional Neural Network, Classification, Cancer*