Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Parçacık Sürü Optimizasyonu Kullanılarak Boyutu Azaltılmış Mikrodizi Verileri Üzerinde Makine Öğrenmesi Yöntemleri İle Prostat Kanseri Teşhisi

Serhat KILIÇARSLAN [a,*], Kemal ADEM [a], Onur CÖMERT [b]

[a] *Gaziosmanpaşa Üniversitesi Enformatik Bölümü Taşlıçiftlik Yerleşkesi, 60250 Tokat / Türkiye*
[b] *Gaziosmanpaşa Üniversitesi Tokat Teknik Bilimler MYO Taşlıçiftlik Yerleşkesi, 60250 Tokat / Türkiye*
[*] *Sorumlu yazar, e-posta: serhat.kilicarslan@gop.edu.tr*

ÖZET

Mikrodizi verilerine dayanan veri madenciliği analizi, hastalık teşhisi ve farmakoloji alanlarında kullanılmaktadır. Analiz aşamasında yaşanan en önemli zorluk, mikrodizilerin yüksek boyutlu olması ve çok sayıda gereksiz öznitelik içermesidir. Bu nedenle çalışmada kullandığımız prostat kanseri mikrodizi veri kümesi üzerinde öznitelik boyut azaltılması amacıyla Temel bileşenler analizi (TBA) ve Parçacık sürü optimizasyonu (PSO) kullanılmıştır. Bu sayede hastalıkları etkileyen genler tespit edilmektedir. Boyutu azaltılmış veri kümeleri Destek Vektör Makinesi ve k-En Yakın Komşuluk sınıflayıcı yöntemlerine giriş olarak verilmiş ve sınıflandırma başarı sonuçları değerlendirilmiştir. Sonuç olarak PSO boyut azaltma yöntemi ile prostat kanserinde etkin genler belirlenmiş ve 50 öznitelik ile %95.77 başarı elde edilmiştir.

Anahtar Kelimeler: *Mikrodizi, Prostat Kanseri, Temel bileşen analizi, Parçacık sürü optimizasyonu.*

Prostate Cancer Diagnosis With Machine Learning Methods On Microarray Data Reduced In Dimension Using Particle Swarm Optimization

ABSTRACT

Data mining analysis based on microarray data is used in disease diagnosis and pharmacology. The major challenge in the analysis phase is the high dimension of microarrays and the large number of unnecessary features. For this reason, Principle Component Analysis (PCA) and Particle Swarm Optimization (PSO) were used to reduce the feature dimension on the prostate cancer microarray dataset used in the study. In this way, genes that affect diseases are determined. Dimension reduced data sets are given as input to Support Vector Machine and k-Nearest neighbor classification methods and classification success results are evaluated. Finally, active genes in prostate cancer were identified by PSO dimension reduction method and 95.77% success was achieved with 50 attributes.

Keywords: *Microarray, Prostate cancer, Principal Component Analysis, Particle Swarm Optimization.*

Geliş: 26/09/2018, Düzeltme: 02/01/2019, Kabul: 21/01/2019

I. GİRİŞ

Teknoloji ve araştırmalardaki yenilikler sayesinde DNA mikrodizi teknolojisi ile binlerce genin ifade analizi gerçekleştirilmektedir [1]. DNA mikrodizisi, binlerce genin ifade analizini mümkün kılan yüksek yoğunluklu gen diziliminden oluşmaktadır [2]. Bu teknoloji, çeşitli hastalık türlerinin tanısı ve teşhisine olanak sağlar. Bu yöntem başta tıp ve ilaç sektörü olmak üzere birçok alanda kullanılmaktadır. Aynı dokuları elde eden hasta ve sağlıklı hücrelerdeki genlerin analizi DNA mikrodizi teknolojisi ile karşılaştırılabilir. Bu sayede hastalıkla ilişkili genlerin bulunması sağlanmaktadır. DNA mikrodizi veri analizi özellikle kanser gibi hastalıklarla ilişkili genlerin tanımlanmasında önemli bir rol oynamaktadır [3]. Hastalıkla ilişkili genleri tanımlayarak herhangi bir kişinin hasta veya sağlıklı olma olasılığı hesaplanmaktadır. Bu nedenle mikrodizi verileri gibi büyük ölçekli verilerin analizi için boyut azaltma teknikleri ve yüksek performanslı sınıflandırma yöntemleri kullanılmaktadır [4].

Literatürde mikrodizi kanser verilerinin boyut azaltma işleminin ardından makine öğrenmesi yöntemleri kullanılarak sınıflandırma başarılarını karşılaştıran çok sayıda çalışma bulunmaktadır. Mikrodizi verilerinin sınıflandırılmasında makine öğrenmesi yöntemleri kullanılmaktadır. Ancak mikrodizi verilerinin boyutunun çok büyük olması ve performans düşüklüğüne sebebiyet vermesinden dolayı farklı boyut azaltma yöntemleri kullanılmıştır. Mikrodizi verisiyle ilgili olarak yapılmış çalışmalar incelendiğinde boyut azaltma amacıyla Temel bileşenler analizi (TBA) ve yapay zeka temelli optimizasyon algoritmalarının yanında Destek vektör makinesi (DVM), k-En yakın komşuluk (kNN), Karar Ağaçları gibi makine öğrenmesi yöntemleri kullanıldığı görülmüştür.

Mikrodizi verisinde boyut azaltma işlemi için Parçacık sürü optimizasyonu (PSO) kullanan çalışmaların ilki lenfoma, lösemi, kolon ve göğüs kanseri mikrodizi verisi üzerine yapılmıştır. PSO boyut azaltma yöntemi kullanılarak gerçekleştirilen sınıflandırmada %80 başarı oranı [5], ikincisinde ise üç adet mikrodizi verisi üzerinde boyut azaltma yapılarak kNN sınıflandırma algoritmasıyla % 92.54 başarı oranı elde edilmiştir [6]. PSO ile hamming uzaklığı birlikte kullanılarak lenfoma, lösemi ve kolon veri seti üzerinde boyut azaltma işlemi uygulanmış olup, kNN algoritması ile % 93.55 başarı elde edilmiştir [7]. Yumurtalık kanseri mikrodizi veri seti üzerine yapılmış çalışmada, boyut azaltma işlemi için hibrit bir yöntem olarak PSO ve Genetik algoritmaları kullanılmış ve DVM ile yapılan sınıflandırmada % 98 başarı [8], son olarak ta on adet farklı mikro dizi veri setine boyut azaltma işlemi uygulanarak C4.5 algoritmasıyla % 88.15 sınıflandırma başarısı gözlemlenmiştir [9]. Göğüs kanseri üzerine yapılmış TBA algoritması kullanılan çalışmaların ilkinde Apriori algoritması ve YSA uygulayarak % 98.29 başarı oranı [10], diğer çalışmada ise Naive Bayes sınıflandırma algoritması ile % 93.42 başarı oranına ulaşılmıştır [11].

Bu çalışmada, prostat mikrodizi veri kümesine TBA ve PSO kullanılarak boyut azaltılmasına gidilmiştir. Boyutu azaltılmış veri kümeleri DVM ve kNN sınıflayıcıya giriş olarak verilmiş ve sınıflandırma başarı sonuçları değerlendirilmiştir.

II. MATERYAL VE YÖNTEM

Yapılan çalışmada prostat kanseri mikrodizi veri seti kullanılmıştır [12]. Çalışmada kullanılan 10509 öznitelik bilgisine sahip veri seti 102 hastaya ait mikrodizi verisinden oluşmaktadır. Bu veri setinde 102 hastanın 50 sinde kanser vakası ile karşılaşılırken 52 hastada kanser vakası ile karşılaşılmamıştır. Buradaki özniteliklerin her birisi bir geni temsil etmektedir. Mikrodizi gen verisine http://www.gems-

system.org/ adresinden ulaşılabilmektedir. Çalışmada boyut azaltma işlemi için PSO ve TBA, sınıflandırma işlemi için de DVM ve kNN yöntemleri kullanılmıştır.

A. BOYUT AZALTMA

Boyut azaltma işlemi, büyük veri kümeleri içerisinden ilgisiz verileri silerek optimum öznitelik alt kümesi bulunması için kullanılan bir yöntemdir [13, 14]. Bu işlem tüm mikrodizi veri setleri için genellikle kullanılmaktadır. Çalışmada kullanılan mikrodizi veri boyutunun büyük olmasından dolayı boyut azaltma işlemi uygulanarak doğru sınıflandırma performansının arttırılması amaçlanmaktadır.

B. TEMEL BİLEŞEN ANALİZİ (TBA)

Makine öğrenmesinde kullanılan ve öznitelik sayısının azaltılması için uygulanan istatistiksel bir yöntemdir [15]. Öğrenme sürecinin hızlanması amaçlanmaktadır. Yöntemin çalışma adımları aşağıdaki gibidir.

Adım 1. Bütün veri kümesini al.
Adım 2. Bütün boyutların ortalama değerlerini hesapla.
Adım 3. Tüm veri kümesinin dağılım matrisini (diğer bir deyişle, kovaryans matrisi) hesapla.
Adım 4. Öz vektörleri (e_1, e_2, ..., e_d) ve bunlara karşılık gelen öz değerleri (λ_1, λ_2, ..., λ_d) hesapla.
Adım 5. Oluşan tabloda öz değerleri büyükten küçüğe doğru sıralayarak $d \times k$ boyutlu bir matris olan W 'yi (burada her sütun bir öz vektörü temsil eder) oluşturmak için en büyük öz değere sahip k öz vektörü seç.
Adım 6. Bu $d \times k$ öz vektör matrisini kullanarak örnekleri yeni alt uzay üzerine dönüştür. Bu durum matematiksel denklemle şöyle özetlenebilir: $y = W^T \times x$ (x, bir örneği temsil eden $d \times 1$ boyutlu bir vektör ve y, yeni alt uzayda dönüştürülmüş $k \times 1$ boyutlu bir örnektir.)

Varyans, verinin yayılımı ile ilgili bir başka ölçüm bilgisi veren kavramdır. PCA da, yüksek boyutlu verideki maksimum varyansları bulmayı ve orjinal verininkine eşit veya ondan daha az boyuta sahip yeni bir alt uzay üzerine yansıtmayı sağlar. Bu nedenle, alt uzayın varyansının maksimum ve boyutunun minimum olması istenir. Bu çalışmada TBA yönteminin boyut azaltma amacıyla tercih edilmesinin nedeni literatürde çok sık kullanılması ve başarılı sonuçlar üretmiş olmasıdır.

C. PARÇACIK SÜRÜ OPTİMİZASYONU (PSO)

Kuş ve balık sürülerinin sosyal davranışlarından esinlenerek geliştirilmiştir. PSO da her bir parçacık bir ajan olarak kabul edilir ve her parçacık bir çözüm sunar. Bir uygunluk fonksiyonu ile her bir parçacığın uygunluk değeri hesaplanır. Bu parçacıklar bir hız ve konum bilgisine sahiptir [16].

Standart PSO algoritmasında bir parçacığın $pbest$ değerinin güncellenmesi ancak şu anki pozisyonunun $pbest$ değerinden daha iyi olması ile gerçekleşmektedir. Fakat bu durum PSO'nun boyut azaltmada kullanılması durumunda bir sınırlandırma getirebilmektedir. Aşağıda verilen algoritmada görüldüğü gibi kullandığımız yaklaşımda Standart PSO algoritmasından farklı olarak uygunluk fonksiyonu ($Fitness$) parçacığın sınıflandırma performansı (*ErrorRate*) aynı iken öznitelik sayısı (*#Features*) daha az olduğunda da $pbest$ güncellenecektir [17].

Kullanılan PSO Algoritması:

Adım 1. Başlangıçta, her bir parçacık, rastgele bir şekilde, öznitelik kümesinin bir alt kümesi (d) olarak ayarlanır.

Adım 2. Her bit parçacık için, eğitim verisinin ilgili parçacığa ait öznitelik alt kümesi üzerinde sınıflandırma yapılarak, $ErrorRate$ değerleri hesaplanır. Daha sonra, her bir parçacığın uygunluk değerleri Eşitlik 1 ve Eşitlik 2 kullanılarak hesaplanır.

$$ErrorRate(d_i) = (FP + FN) / (TP + TN + FP + FN) \quad (1)$$

$$Fitness_i = ErrorRate_i + \alpha \times \#Features \quad (2)$$

Eşitlik 2'de bulunan $ErrorRate$, sınıflandırma performansını, d_i, ilgili parçacığa ait öznitelik alt kümesini, $\#Features$, parçacığa ait öznitelik sayısını ve α, $\#Features$ değerinin önemini belirleyen çok küçük değerli bir katsayıyı ifade etmektedir.

Adım 3. $pbest$ ve $gbest$ değerlerinin güncellenmesi aşağıdaki şekilde yapılır.
Eğer $Fitness_i < pbest_i$ ise $pbest_i = Fitness_i$
Eğer $Fitness_i = pbest_i$ ve $\#Features_i < \#Features$ ise $pbest_i = Fitness_i$
Eğer $Fitness_i > pbest_i$ ise hiçbir şey yapma.

Tüm parçacıkların hesaplanan $pbest$ değerlerinin en iyisi $gbest$ değeri Eşitlik 3'teki gibi ayarlanır.

$$gbest = \min(pbest) \quad (3)$$

Adım 4. Her bir parçacık için parçacık hızı (V_i) hesaplanır ve parçacık konumu (X_i), V_i 'ye bağlı olarak güncellenir. Güncelleme işlemleri için Eşitlik 4 ve 5 kullanılır.

$$V_i^{k+1} = V_i^k + c_1 rand_1\left(pbest_i - X_i^k\right) + c_2 rand_2(gbest - X_i^k) \quad (4)$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (5)$$

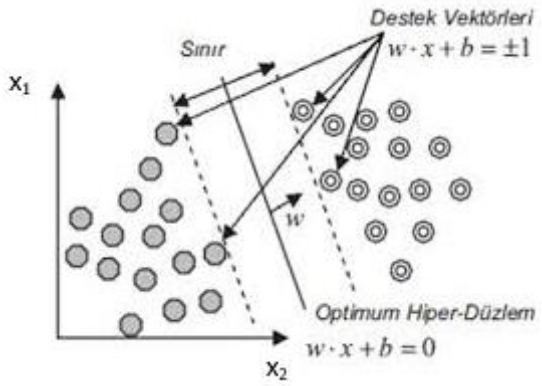
Burada;

- X_i : parçacığın mevcut pozisyonu (konumu),
- V_i : Parçacığın mevcut hızı (konum değişikliği),
- $rand$: (0,1) aralığında rastgele bir değer,
- k: iterasyon,
- $pbest$: her parçacığın kendi en iyi çözümü (Personal Best),
- $gbest$: tüm parçacıkların en iyi çözümü (Global Best),
- c_1, c_2: (0-4) aralığında seçilen öğrenme faktörüdür. Parçacıkları $pbest$ ve $gbest$ konumlarına doğru yönlendirir.

Adım 5. Maksimum iterasyon değerine kadar 2-4 adımları tekrarlanır.

D. DESTEK VEKTÖR MAKİNESİ (DVM)

DVM, örüntü tanıma ve sınıflandırma problemlerinin çözümü için Vapnik tarafından geliştirilmiş, dağılımdan bağımsız bir algoritmadır [18]. Veriyi sınıflandırırken sınıfların birbirlerine en yakın olan örneklerini bularak bu örneklerin iki sınıfı ayıracak olan ayırıcı yüzeye dik uzaklıklarını maksimize etmeyi amaçlar. Ayırıcı yüzeyin veri kümesi üzerindeki başarısı değişmeden birçok farklı kombinasyonu olabilir. DVM sayesinde ayırıcı yüzey her iki sınıfa da aynı mesafede ve maksimum uzaklıkta olur. Bu durum Şekil 1'de görülmektedir.



***Şekil 1**. Genel bir destek vektör makinesi modeli [18]*

Şekil 1'de görüldüğü gibi, x_1 ve x_2 düzleminde bulunan veri kümesinin iki sınıfın ayırıcı yüzeye uzaklığı olan $\|\vec{w}\|$ değerini minimize etmek iki sınıf arasında ayrılmayı maksimize eder. $\|\vec{w}\|$ değerini minimize etmek amacıyla Lagrange optimizasyonu kullanılır. Lagrange optimizasyonu bir fonksiyonun en küçük ve en büyük değerlerini bir kısıda bağlı olarak bulmak gerektiğinde uygulanan en temel yöntemdir. Bir *g(x)* kısıt fonksiyonunun *α* katsayısı oranında *f(x)* amaç fonksiyonuna eklenmesi ile elde edilir. Bu işlem Eşitlik 6 ile ifade edilmektedir.

$$L(x, \alpha) = f(x) + \alpha\, g(x) \tag{6}$$

Doğrusal olarak ayrılabilen bir veri kümesinde DVM algoritması aşağıdaki gibi çalışır:

Adım 1. İki farklı sınıfın birbirine en yakın iki noktası (destek vektörleri) seçilir ve bu noktalardan geçen ve birbirine paralel iki hiper düzlem çizilir. Bu iki hiper düzlem sınırını (H_1, H_2) oluşturur. Bu düzlemler arasında hiçbir örnek olmamalıdır.

Adım 2. Kullanılacak hiper düzlemin normali aynı zamanda ağırlık vektörü ($\vec{w}$) nün normu $\|\vec{w}\|$ Eşitlik 7 ile hesaplanır. Optimum hiper düzlem, sınır düzlemlerinin tam ortasında bulunan düzlemdir.

$$\|\vec{w}\| = \|x_1 - x_2\| \tag{7}$$

Adım 3. Örnekleri en iyi şekilde ayıran hiper düzlem aşağıdaki eşitliği minimize eden düzlemdir.

$$\theta(w) = \frac{1}{2}\|\vec{w}\|^2 \tag{8}$$

Bu nedenle veri setindeki tüm örneklerin aşağıdaki eşitliği sağlaması gerekir.

$$\forall i \; için \; y_i(\langle w, x_i \rangle + w_0) - 1 \geq 0 \tag{9}$$

Amaç $\|\vec{w}\|$ yi minimize ederek hiper düzlemin örnekleri en iyi şekilde ayırmasını sağlamaktır. Bu optimizasyon işlemi Lagrange Optimizasyon algoritması kullanılarak yapılır.

Adım 4. Önceki adımlarda hesaplanan değerler aşağıdaki eşitlikte yerlerine konularak ayırıcı hiper düzlemin denklemi elde edilir.

$$g(w) = \vec{w}^{\mathrm{T}}\vec{x} + w_0 \tag{10}$$

Adım 5. Bütün noktaların koordinatları Eşitlik 10'a verilerek $g(\vec{x})$ değerleri hesaplanır ve sınıflandırma Eşitlik 11'deki gibi yapılır.

$$g(\vec{x}_i) >= 1 \, , y_i = 1 \; (class_1)$$
$$g(\vec{x}_i) <= -1 \, , y_i = -1 \; (class_2) \tag{11}$$

Doğrusal olmayan DVM yönteminde, çekirdek fonksiyonları (sigmoid, RBF, polinomsal) kullanılarak doğrusal olmayan veri setleri başarılı bir şekilde ayrılarak sınıflandırılmaktadır. Çalışmamızda doğrusal olmayan DVM için polinomsal çekirdek fonksiyonu kullanılmıştır. Polinomsal çekirdek fonksiyonu Eşitlik 12'deki gibi ayarlanır ve fonksiyon içerisinde kullanılan parametrelerden *a* alfa eğimini, *c* sabit terimi, *d* polinom derecesini temsil etmektedir. Bu parametreler deneysel çalışmalar sonucu elde edilmektedir [19].

$$g(x, y) = (ax^T + c)^d, a > 0 \tag{12}$$

E. k EN YAKIN KOMŞULUK (kNN)

kNN, eğiticili ve örnek tabanlı bir sınıflandırma algoritmasıdır. kNN sınıflandırma algoritması veri madenciliği, biyoenformatik, tıp, saldırı tespit ve örüntü tanıma sistemleri gibi birçok alanda kullanılmaktadır. Bu algoritmada sınıflandırma, belirlenecek *k* değerine göre hangi komşusu daha yakın ise veriler o sınıfa ait olacak şekilde yapılmaktadır [20]. Uzaklık mesafelerinin hesaplanabilmesi için Euclidean, Manhattan, Minkowski ve Chebyschev uzaklık ölçütleri kullanılmaktadır. Çalışmada Euclidean uzaklık parametresi kullanılmış ve *k* değerini 5 alarak sınıflandırma işlemi gerçekleştirilmiştir.

III. BULGULAR VE TARTIŞMA

Çalışmada kullanılan mikrodizi veri kümesine doğrudan ve boyut azaltma yöntemleri (TBA ve PSO) ile DVM ve kNN sınıflayıcıları uygulanarak gerçekleştirilen deneysel çalışmalar Tablo 1,2 ve 3'te gösterilmektedir.

Tablo 1. *kNN ve DVM Sınıflandırma Başarısı*

kNN			DVM		
Duy.	**Özg.**	**Doğ.**	**Duy.**	**Özg.**	**Doğ.**
78.6	81.2	80.4	89.4	90.9	90.2

Tablo 2. *TBA-DVM ve TBA-kNN Sınıflandırma Başarısı*

Varyans Oranı *(%)*	**kNN**			**DVM**		
	Duy.	**Özg.**	**Doğ.**	**Duy.**	**Özg.**	**Doğ.**
75	77.8	79.2	78.4	82.5	84.2	83.1
80	73.9	75.1	74.5	80.4	82.6	81.7
85	74.8	76.2	75.5	80.4	82.6	81.7
90	**88.6**	**89.8**	**89.2**	**93.8**	**95.2**	**94.4**
95	89.4	90.9	90.2	87.9	89.6	88.7

Tablo 3. *PSO-DVM ve PSO-kNN Sınıflandırma Başarısı*

Parçacık Sayısı	**kNN**			**DVM**		
	Duy.	**Özg.**	**Doğ.**	**Duy.**	**Özg.**	**Doğ.**
10	67.8	69.3	68.6	72.7	74.1	73.2
20	57.9	59.5	58.8	69.2	71.2	70.4
30	77.6	79.2	78.4	86.4	87.9	87.3
40	82.5	83.9	83.3	76.9	78.6	77.5
50	**72.2**	**74.2**	**73.5**	**95.1**	**96.4**	**95.8**
60	71.8	73.1	72.5	80.8	82.4	81.7
70	76.7	78.2	77.5	89.5	91.2	90.1
80	79.5	80.9	80.4	86.7	88.1	87.3
90	77.6	79.2	78.4	87.2	89.2	88.7
100	77.6	79.2	78.4	87.2	89.2	88.7

Tablo 1'de boyut azaltma işlemi yapmadan DVM ve kNN sınıflayıcısı uygulanarak sırayla %90.2 ve %80.4'lük başarı oranı elde edildiği gözlemlenmiştir. Tablo 2'de görüldüğü gibi % 90 varyans kullanılan TBA boyut azaltma yöntemi ve DVM sınıflayıcısı uygulanarak % 94.4 doğru sınıflandırma oranı yakalanmıştır. Tablo 3'te ise 50 parçacık seçilerek PSO ile boyut azaltma yöntemi ve DVM sınıflayıcısı uygulanarak doğru sınıflandırma oranı % 95.77'e çıkarılmıştır. Gerçekleştirilen deneysel çalışmalarda, DVM'de polinomsal çekirdek fonksiyonu kullanılmıştır. Bu fonksiyonda kullanılan *a* değeri 0.5, *c* değeri 1, *d* ise 2 olarak deneysel çalışmalar sonucunda belirlenmiştir. PSO'da maksimum iterasyon 50 olarak alınmıştır. kNN sınıflayıcısında ise k=5 ve Euclidean uzaklık ölçütü kullanılmıştır. Deneysel çalışmalar sonucunda boyut azaltma işlemi gerçekleştirerek prostat kanseri mikrodizi veri kümesi üzerinde DVM'nin kNN'ye göre daha başarılı olduğu gözlemlenmiştir.

IV. SONUÇ

Mikrodizi veri kümelerinin analizi kanser hastalıklarıyla ilişkili genlerin tanımlanmasında önemli bir rol oynamaktadır. Bu nedenle mikrodizi verileri gibi büyük ölçekli verilerin analizi için boyut azaltma teknikleri ve yüksek performanslı sınıflandırma yöntemleri kullanılmaktadır. Bu çalışmada da mikrodizi veri kümesi üzerinde prostat kanserini sınıflandırmak için TBA ve PSO boyut azaltma yöntemlerinin yanında DVM ve kNN sınıflayıcıları kullanılmıştır. Yapılan sınıflandırma işlemi sonucunda prostat kanseri mikrodizi verisi üzerinde DVM'nin, kNN'e göre daha iyi bir sınıflandırma performansı gösterdiği görülmüştür. Sonuç olarak mikrodizi veri kümelerine boyut azaltma ve sınıflandırma algoritmaları uygulanırken kullanılan parametrelerin seçimleri ile öznitelik sayısının belirlenmesi doğru sınıflandırma oranında önemli bir etkendir. Doğru parametrelerin seçilmesi durumunda biyoinformatik alanında klinik çalışmalara da yol gösterecektir.

V. KAYNAKLAR

[1] H. Liu, I. Bebu and X. Li, "Microarray probes and probe sets," Frontiers in bioscience (Elite edition), vol. 2, pp. 325-338, 2010.

[2] H.U. Luleyap, "The Principles of Moleculer Genetics," İzmir, Türkiye: Nobel Yayınevi, 2008.

[3] K. Ipekdal, "Microarray Technology," (2018, 10 Aralık). [Online]. Available: http://yunus.hacettepe.edu.tr/ ~mergen/sunu/s_mikroarrayan decology.pdf.

[4] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning", *In Computer science'98 proceedings of the 21st Australasian computer science conference ACSC*, 1998, pp. 181-191.

[5] B. Sahu, and D. Mishra, "A novel feature selection algorithm using particle swarm optimization for cancer microarray data," Procedia Engineering, vol. 38, pp. 27-31, 2012.

[6] S. Kar, , K. D. Sharma and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique," Expert Systems with Applications, vol.42, no.1, pp. 612-627, 2015.

[7] H. Banka and S.A Dara, "Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, classification and validation," Pattern Recognition Letters, vol. 52, pp. 94-100, 2015.

[8] P. Yasodha and N. R. Ananthanarayanan, "Analysing big data to build knowledge based system for early detection of ovarian cancer," Indian Journal of Science and Technology, vol. 8, no. 14, 2015.

[9] K. H. Chen, K. J. Wang, M. L. Tsai, K. M. Wang, A. M. Adrian, W. C. Cheng, ... and K. S. Chang, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," BMC bioinformatics, vol. 15, no. 1, pp. 49, 2014.

[10] O. Inan, M. S. Uzer, and N. Yılmaz, "A new hybrid feature selection method based on association rules and PCA for detection of breast cancer," International Journal of Innovative Computing, Information and Control, vol. 9, no. 2, pp. 727-729, 2013.

[11] M. Kaya, H. Ş. Bilge ve O. Yıldız, "Gen ifadelerinde Öz Nitelik Seçimi ve Boyut İndirgeme, 21. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı SIU2013, Haspolat, Türkiye, 2013.

[12] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, ... and E. S. Lander, "Gene expression correlates of clinical prostate cancer behavior," Cancer cell, vol. 1, no. 2, pp. 203-209, 2002.

[13] H. Göker ve H. Tekedere, "Fatih Projesine Yönelik Görüşlerin Metin Madenciliği Yöntemleri İle Otomatik Değerlendirilmesi," *Bilişim Teknolojileri Dergisi*, c. 10, s. 3, ss. 291-299, 2017.

[14] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[15] I. T. Jolliffe, "Principal component analysis and factor analysis," In Principal component analysis, pp. 115-128, 1986.

[16] J. Kennedy, and R. Eberhart,"PSO optimization," *In Proc. IEEE Int. Conf. Neural Networks*, 1995, pp. 1941-1948

[17] B. Xue, M. Zhang, & W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," Applied Soft Computing, vol. 18, pp. 261-276, 2014.

[18] V. Vapnik, "The nature of statistical learning theory," Springer-Verlag: New York, 1995, pp. 75-100.

[19] E. Karacan, "Hastalıkların Uyarlanmış Destek Vektör Makinesiyle Teşhis Edilmesi, " Yüksek Lisans Tezi, Bilgisayar Mühendisliği, Ondokuz Mayıs Üniversitesi, Samsun, Türkiye, 2015.

[20] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 6, no. 4, pp. 325–327, 1976.