Lojistik Regresyonun Özellik Azaltma Teknikleri İle Gen Dizilimlerinin Sınıflandırılmasındaki Başarısı

(The Success Of Logistic Regression With Feature Reduction Techniques On Microarray Gene Classification)

Yeliz YENGİ[1] Sevinç İLHAN OMURCA[2]
[1]Kocaeli Üniversitesi yelizyengi@gmail.com
[2]Kocaeli Üniversitesi silhan@kocaeli.edu.tr

Öz

Gen dizilimlerinin sınıflandırılması, hastalıkların ön görülebilmesi veya teşhis edilebilmesinde çok önemli rol oynamaktadır. Bütün gen dizilimi üzerinde etkili bir sınıflandırma yapabilmek mümkün olmadığından sağlıklı bir sınıflandırma yapılabilmesi için gerekli bilgiyi içeren genlerin özellik azaltma algoritmaları ile ayıklanması önem taşımaktadır. Bu çalışmada, özellikleri azaltmak için sezgisel arama teknikleri, özellik azaltma yaklaşımları (filter, wrapper, vb.) gibi farklı yöntemler analiz edilerek ön işleme adımının daha etkin bir şekilde gerçekleştirilmesi; bunun sonucunda elde edilen veri kümelerinin LR (Lojistik Regresyon) ve SVM (Destek Vektör Makineleri) gibi güçlü sınıflandırma araçları ile daha etkin şekilde sınıflandırılması hedeflenmiştir. Makine öğrenmesinde güçlü bir sınıflandırıcı olarak kabul edilen LR sınıflandırıcısı, özellik eksiltme yöntemleri ile gen dizilimlerinin sınıflandırılmasında SVM kadar geçerli ve etkin sınıflama aracı haline gelmiştir.

Anahtar Sözcükler: *Gen analizi, makine öğrenmesi, lojistik regresyon, özellik azaltma, SVM.*

Abstract

DNA microarray classification is important to discovery of differentially expressed genes between normal and diseased patients are a central research problem in bioinformatics. All the genes used in the expression profile are not informative. Further, many of them are redundant. A pre-processing step in order to reduce the number of genes by feature selection and still retaining best class prediction accuracy for the cla[1]ssifier is crucial for precise tumor classification. In this study comparison between class prediction accuracy of two different classifiers, LR (Logistic Regression) and SVM (Support Vector Machines), was carried out using the best genes select by wrapper and filter technique to use heuristic search methods. We conclude that LR together with heuristic search based feature selection is the as efficient as SVM to the microarray gene prediction techniques.

Keywords: *Microarray analysis, binary classification, machine learning, logistic regression, feature reduction, tumor analysis, SVM.*

1. Giriş

Günümüzde hızla gelişen mikro dizin teknolojisi yardımıyla büyük ölçekli gen dizilimleri üretilmektedir. Bu gen dizilimlerinin yorumlanması sayesinde kanser tanı ve tedavisine destek olunmaktadır. Kanser araştırmalarında normal ve kanserli hücrelerin aktivitelerini izlemenin, analiz etmenin en iyi yolu mikro dizin analizi olarak kabul edilmektedir. Binlerce gen; farklı hücrelerden, aynı hücrede farklı noktalardan ve farklı dokulardan alınan gen dizilimlerini ihtiva etmektedir [1]. Makine öğrenmesi algoritmaları binlerce genden oluşan dizinleri, filtrelemekte, benzerliklerine göre bir örüntü yakalayarak gruplamakta ve yorumlanmalarına katkı sunmaktadır [2]. Bu algoritmalar sayesinde genlerin bir birleri ile işlevsel ilişkilerinin keşfedilmesi; normal ve hasta dokular arasındaki farkların gözlemlenmesi olanaklı hale gelmektedir [3]. Bu yaklaşımlar kanserli hücrelerin analizi dışında, diğer klinik konuların yorumlanmasında da etkili olmaktadır [4].

Gönderim ve kabul tarihi : 30.3.2015-20.05.2015

Gen profillerinde birçok gürültü ve gereksiz veri bulunmaktadır bu nedenle, ön işlemler uygulamadan gen dizileri üzerinde sınıflandırma yapmak mümkün olamamaktadır [5]. Çeşitli özellik azaltma algoritmaları kendi özellik değerlendirme ölçütlerine ve tekniklerine (tek tek veya alt küme olarak) göre gruplama yapmaktadır ve bu algoritmalar genellikle filter veya wrapper teknikleri olarak sınıflandırılmaktadır [6]. Filter tekniği sınıfında yer alan algoritmalar makine öğrenmesi algoritmalarını içermeden gen alt kümelerini verinin yapısına bakarak seçmektedir. Wrapper tekniğin de ise makine öğrenmesi algoritmaları kullanılmaktadır [7]. Performansı arttırmak ve en iyi veri alt kümesini bulabilmek için tekrarlamalı olarak seçilen sınıflandırma algoritması çağırılarak veri alt kümelerini değerlendirmektedir. Yüksek boyutlu veri kümelerinde filter tekniğine göre daha pahalı bir hesaplama gerçekleştirilmekte olsa da öğrenme algoritmalarının hibrit bir şekilde uygulanmasını sağlamakta olduğundan farklı yaklaşımların uygulanmasına temel oluşturmaktadır [8].

Literatürde makine öğrenmesi yöntemleri kullanılarak gen dizilimleri üzerinden kanser teşhisi için gerçekleştirilen çalışmalar yer almaktadır. Inza vd. 2003' de wrapper tekniğini uyguladıkları çalışmada, kolon ve kan kanseri gen dizilimi veri kümeleri kullanmışlardır. Sınıflandırıcı olarak IB1, C4.5, CN2 ve Naive Bayes kullandıkları çalışmada %95,16 ve %100 doğruluk oranlarını elde etmişlerdir [12]. Ruiz vd. 2005'de filter ve wrapper tekniklerini Naive Bayes algoritması ile birlikte kolon, lenfoma ve kan kanseri gen dizilimi veri kümeleri üzerinde uygulamışlardır. Elde ettikleri sonuçlara göre filter tekniği ile kolon, lenfoma, kan kanseri veri kümeleri için sırasıyla %85, %76, %93; wrapper tekniğinde %85, %82, %93; herhangi bir özellik azaltım yöntemi uygulanmadan %53, %75, %98 doğruluk oranlarında sınıflandırma yapmışlardır [6].

Gerekli bilgiyi içeren genlerin seçilebilmesi için uygulanan filter tekniği ile özellikler ağırlıklandırılırken, wrapper tekniği ile olası bütün alt kümeleri değerlendirilmektedir. Bu teknikler "özellik seçerek" gerçeklenmektedir fakat diğer bir yaklaşım olan "özelliklerin azaltılarak" alt kümenin seçilmesi yaklaşımında singular value decomposition (SVD) [9], principal component analysis (PCA) [10] ve independent component analysis (ICA) [11] gibi özellik azaltma algoritmaları sıkça kullanılmaktadır. Bu algoritmalar Neural Network (NN), Random Forest (RF) ve Support Vector Machine (SVM) gibi sınıflandırıcıların daha doğru tahminde bulunmasını ve daha etkin çalışmasını sağlamaktadır.

Bartenhagen vd. gerçekleştirdikleri çalışmada PCA ile Kernel PCA (KPCA), Isomap (IM), Maximum Variance Unfolding (MVU), Diffusion Maps (DM), Locally Linear Embedding (LLE) ve Laplacian Eigenmaps (LEM) gibi özellik azaltma yöntemlerini karşılaştırmışlardır. Göğüs kanseri için %87,4, kan kanseri için %97,2 başarı oranları yakalamışlardır. Elde ettikleri sonuçlara göre PCA algoritmasının iki sınıflı veri kümeleri için diğer yöntemler ile benzer sonuçlar ürettiği fakat çoklu sınıf içeren veri kümeleri için diğer yöntemlere göre daha başarısız olduğu gösterilmiştir [21].Lenfoma kanseri ve kolon kanseri gen dizilimi veri kümeleri üzerinde gerçekleştirilmiş güncel çalışmalardan elde edilmiş olan sonuçların özeti Çizelge 1 ve Çizelge 2' de gösterilmiştir. Özetlenmiş olan çalışmalarda kullanılmış olan lenfoma ve kolon kanseri veri kümeleri farklı kaynaklardan alınmıştır.

Çizelge-1: Lenfoma Literatür özeti

Çalışma yazar/yılı	Gen Seçimi	Sınıflandırıcı	Tahmin
Ruiz vd. /2005 [6]	Wrapper Tekniği	Naive Bayes	%82
Nguyen vd./2002 [23]	PLS	QDA - LD	%96 – 98,1
Kim vd. /2006 [22]	MLP	SVM	%92 – 100
Dağlıyan vd. /2011 [26]	CBF	Hyper-box enclosure method (HBE)	%92,25 – 93,71
Vimaladevi vd./2014 [27]	Hybrid Fast PSO-BPN	Gas + SVM	%82
Pauly vd./ 2014 [28]	PCA	SVM	%94 – 100
Bizim çalışmamız	CFS + Sezgisel	LR	%100

Geleneksel sınıflandırma yöntemlerinden olan lojistik regresyon (LR) birçok tıbbi sınıflandırma problemlerinde kullanılmasına rağmen [13], gen analizi probleminde yeterince değerlendirilmemiştir. Son yıllarda gerçekleştirilen bilimsel araştırmalarda genelde tıbbi sınıflandırma problemleri için NN, SVM, Classification Trees (CT) ve RF gibi algoritmaların uygulandığı gözlenmektedir [14].

Çizelge-2: Kolon Literatür özeti

Çalışma yazar/yılı	Gen Seçimi	Sınıflandırıcı	Tahmin
Ruiz vd. /2005 [6]	Wrapper Tekniği	Naive Bayes	%85
Inza vd. /2003 [12]	Wrapper Tekniği	Naive Bayes - C4.5	%87-95,16
Kim vd. /2006 [22]	MLP	SVM	%83,9 - 93,5
Yan vd./2012[29]	Real-time PCR	Random Forest	%91,94
Thorsteinsson vd./2012 [30]	-	PAM	% >90
Bennet vd. / 2014 [31]	Discrete Wavelet Transform(DWT)	SVM	%92,5
Bizim çalışmamız	CFS + Sezgisel	LR	%90,9

Gerçekleştirdiğimiz çalışmanın amacı, LR sınıflandırıcısının daha önce uygulanmadığı gen dizilimleri analizi alanında SVM gibi güçlü sınıflandırıcılar kadar etkin olduğunun gösterilmesi ve LR sınıflandırıcısının performansının birçok farklı özellik azaltma tekniği ile arttırılmasıdır. Günümüzde makine öğrenmesi uygulamalarında SVM sınıflandırıcının ikili sınıflandırma noktasında en iyi algoritma olarak kabul görmesi nedeniyle bu algoritmayı LR ile karşılaştırmak algoritmanın başarısını kıyaslamamızı sağlamaktadır.

2. Kuramsal Altyapı

Makalede kullanılan yöntemler gen seçimi ve sınıflandırma alt başlıklarında incelenecektir.

2.1. Gen Seçimi

2.1.1. Temel Bileşenler Analizi

Temel bileşenler analizi (Principal Component Analysis-PCA), özellik azaltmada çok yaygın ve etkin şekilde uygulanan lineer bir yöntemdir. PCA özellikler arasındaki bağımlılığı yok etmeye çalışarak veri kümesinin boyutunu azaltmaktadır. Varsayalım ki n boyutlu bir $X = (x_1, x_2, ..., x_n)^T$ vektörü verileri temsil etmektedir. PCA adımları aşağıda özetlenmektedir.

Veri kümesine ait her boyutun ortalama değerinin hesaplanır, $\bar{x} = E(x)$.

Veri kümesine ait kovaryans matrisi $\sum = E\left[\left(x - \bar{x}\right)\left(x - \bar{x}\right)^T\right]$ hesaplanır. Eğer verinin x_i ve x_j bileşenleri birbiri ile ilişkisiz ise kovaryansları 0 dır, $\sum_{ij} = \sum_{ji} = 0$.

Bir kare matris olan kovaryans matrisine ait özvektör ve özdeğerleri değerlerinin hesaplanır. Bu değerler veri hakkında yararlı bilgiyi temsil etmeleri açısından önemlidir. Özvektörler $e_i (i = 1,...,k \le n)$ ve özdeğerleri λ_i, $\sum_i e_i = \lambda_i e_i$ denklemin çözümleridir.

En yüksek özdeğerine sahip olan özvektör, veri kümesinin temel bileşenidir. Kovaryans matrisine bağlı özvektörler hesaplandığında bu değerler büyükten küçüğe sıralanırlar. Bu sıralama bileşenlerin büyükten küçüğe önem sıralamasını vermektedir. Başlangıçta "n" boyutlu bir veri kümeniz var ise "n" adet özvektör ve özdeğerler hesaplanmaktadır. Bu durumda k<n olmak üzere ilk "k" adet özvektör veri kümesini temsil etmek üzere seçilebilir. Böylece başlangıçta "n" boyutu olan veri kümesi "k" boyut ile temsil edilebilir. Son olarak verinin özellik vektörü seçilen özvektörler ile kurulmaktadır.

Özellik vektörü = $(\ddot{o}zv_1, \ddot{o}zv_2, \ldots, \ddot{o}zv_n)$

Bir önceki adımda kurulan özelik vektörünün transpozesi alınarak özvektörlerin, en önemli olanlar en üstte olmak üzere matris satırlarında yer alması sağlanır. Veri kümesini temsil eden son matris veri yapısının satırları veri nesnelerini ve sütunları özellikleri temsil etmektedir.

Diyelim ki A matrisi, kovaryans matrisinin özvektörlerinin satır vektörlerini oluşturduğu bir geçiş matrisi olsun. Gen dizilimi veri kümesini bir vektöre dönüştürerek özellik azaltma işleminin y çıkışını elde ederiz [25].

$$y = A(x - \bar{x}) \qquad (1)$$

Orijinal x veri vektörünü y vektöründen yeniden elde edebiliriz.

$$x = A^T + \bar{x} \quad (2)$$

Eğer A matrisini ilk K adet özvektörler, A_K^T, ile kurmuşsak yukarıdakine benzer bir dönüşümü aşağıdaki denklemler ile elde edebiliriz.

$$y = A_K(x - \bar{x}) \quad (3)$$

$$x = A_K^T y + \bar{x} \quad (4)$$

2.1.2. CFS

Özellik uzayında ilişki tabanlı seçim yöntemiyle alt kümelerin değerlendirilmesi yöntemidir [15]. Her özelliğin için diğer özelliklerle ilişkisi derecelendirilmekte ve özellikler arasındaki sınıf değişkeni değeri ile simetrik belirsizlik hesaplanmaktadır. Bu değer herhangi iki nominal X, Y değişkeni için denklem 5'de gösterilmektedir

$$Simetrik\ Belirsizlik = 2 \times \left(\frac{H(Y) - H(Y|X)}{H(Y) + H(X)} \right) \quad (5)$$

burada, H(y) y'nin entropisidir,

$$H(Y) = -\sum_{y \in Y} \mathrm{P(y)} \log_2 \mathrm{p(y)} \quad (6)$$

$$H(Y|X) = -\sum_{x \in X} \mathrm{P(x)} \sum_{y \in Y} \mathrm{P(y|x)} \log_2 \mathrm{p(y|x)} \quad (7)$$

H(Y|X), X'e bağlı Y'nin koşullu entropisidir [16]. Hesaplamanın doğruluğu alt veri kümelerinin bulunması ile ilişkilidir. Alt veri kümeleri sezgisel yöntemlerle aranmaktadır.

2.1.3. Wrapper Tekniği

Wrapper tekniğinde seçilen makine öğrenmesi algoritması ile aday alt gen dizilimleri performans testi yolu ile aranmaktadır. Bu yöntemde kullanılan öğrenme algoritmasının sürekli olarak çağırılması, yöntemin filter tekniğine göre daha yavaş çalışmasına neden olmaktadır. Wrapper tekniği çok yüksek boyutlu veri kümeleri için çok fazla bilgisayar zamanına ihtiyaç duymaktadır. Dolayısıyla yöntemin uygulanmasında hızlı çalışan basit öğrenme algoritmalarının seçilmesi önem kazanmaktadır. Bu çalışmada C4.5, Naive Bayes ve SVM algoritmaları seçilmiştir.

2.2. Lojistik Regresyon

Lojistik regresyon (LR) özellikler arasındaki ilişkiyi açıklayan, sınıf etiketi {0,1} şeklinde verilen ikili sınıflandırma problemlerinde kullanılan istatistiksel sınıflandırma yaklaşımıdır. Bu çalışmada LR sınıflandırıcısının kullanılan üç versiyonu aşağıda özetlenmektedir

2.2.1. Logistic Model Trees

Logistic Model Trees (LMTs), lineer lojistik regresyon ve ağaç sınıflandırıcılarının kombinasyonundan türetilmiştir. Lineer lojistik regresyon modeli veri uzayını yinelemeli bir şekilde bölümleyerek her bölüme farklı lojistik regresyon verir, bu lineer olmayan verinin dönüştürülmeden modellenmesini sağlamaktadır. LMTs yukarıdan aşağıya lojistik regresyon düğümlerinden oluşan ağaç meydana getirmektedir [17], verileri dağıtırken yığılma problemi oluşabileceği için verilerin ayrıştırılırken dengeli olmasını ki-kare testi kullanarak lineer etkilerinden ve lineer olmayanı ayırarak dengeler. İlk ağaca lojistik regresyonlar yerleştirildikten sonra CART (classification and regression trees) algoritmasına benzer bir yöntemle geri budanır [18].

2.2.2. Multinominal Logistic Regression (MLS)

LRMLS denklem 8 ile maksimum olabilirlik teknikleri tahmin eden lineer olmayan bir modeldir.

$$P(Y_i = j) = \frac{\exp(X_i \beta_j)}{1 + \sum_j^J \exp(X_i \beta_j)} \quad (8)$$

Burada her (i) bir seçenekler kümesi(j) ile karşılaştırılır.P(Y = j | x), j = 0, 1, …,J. β, d-boyutlu parametre vektörüdür. (X,Y) verisi için l tahmin modeli aşağıdaki gibi hesaplanmaktadır.

$$l(\beta) = \sum_i [Y_i \log p(X_i) + (1 - Y_i) \log \{1 - p(X_i)\}] \quad (9)$$

l(β) maksimizasyonu

$$l^{\lambda}(\beta) = l(\beta) - \lambda||\beta||^2 \quad (10)$$

l(β) sonsuza giderken denklem,

$$||\beta|| = (\sum \beta_j^2)^{1/2} \quad (11)$$

şeklinde ifade edilir; λ değeri azaltılarak *l(β)* değeri kontrol edilmektedir.

Muhtemel olan bir tane j değeri denklem 12 ile hesaplanmaktadır.

$$P(Y_i = 0) = \frac{1}{1 + \sum_j^J \exp(X_i\beta_j)} \quad (12)$$

burada, X potansiyel olarak belirtilen bağımsız değişkenler kümesidir.

2.2.3. Kernel Logistic Regression

Veri kümesi n adet örnek içerirken, (i = 1, ⋯, n), y_içıktısı 0 (hasta değil) veya 1 (hasta) sonucu üretir.x_i,qx1 boyutlu kovaryans vektörü, z_i, px1 boyutlu gen dizilimleri vektörüdür ve aşağıdaki gibi LR modeli ile ifade edilebilir.

$$logit(\mu_i) = x_i^T\beta + h(z_i) \quad (13)$$

μ_i = P (y_i = 1|x_i, z_i), β → q × 1 regresyon katsayılarıdır, h(z_i) bilinmeyen merkezli pürüzsüz bir fonksiyondur, h(z) =$\gamma_1 z_1 + \cdots + \gamma_p z_p$.

h(z) fonksiyonu Gauss çekirdek (kernel) metodu kullanılarak yazılırsa, K(z_1, z_2) = exp{-||$z_1 - z_2$||$^2/p^2$}, ρ bilinmeyen değişken,

$$||z_1 - z_2||^2 = \sum_{k=1}^{p} (z_{1k} - z_{2k})^2 \quad (14)$$

Bilinmeyen parametre fonksiyonları için çekirdek metodu önemli bir rol oynar, eldeki verilerden optimum bir tahmin yapmak zor bir problemdir [20].

2.3. Destek Vektör Makineleri

Destek vektör makineleri yönteminde giriş uzayı yüksek boyutlu bir özellik uzayına dönüştürülmekte ve bu uzayda veriyi sınıflara ayıran çok sayıda hiperdüzlem (lineer sınıflandırıcı) bulunmaktadır. Destek vektör makineleri ile {+1,-1} şeklinde etiketlenebilen iki sınıfa ait veri noktalarının, eğitim verisinden elde edilen bir karar fonksiyonu yardımıyla birbirinden ayrılması amaçlanır. Söz konusu karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırabilecek optimum hiper-düzlem bulunur. Amaç alt düzlemler oluşturabilmek için bir lineer vektör oluşturmaktır; wx+b=0.

$\{x^{(i)}, y^{(i)}\}$, i=1,…,m şeklinde kaydedilmiş m adet eğitim verisi çiftinden oluşan bir D eğitim kümesi için aranan optimum düzleme ait eşitsizlik aşağıdaki gibidir.

$$y^{(i)}(\vec{w}\vec{x} + b) - 1 \geq 0 \quad (15)$$

Burada $x^{(i)} \in R_N$ olup N-boyutlu bir uzayı, $y^{(i)} \in \{-1,+1\}$ sınıf etiketlerini, w ağırlık vektörünü (hiper-düzlemin normali) ve b eğilim değerini göstermektedir. Optimum hiperdüzlemin bulunması için bu düzleme paralel olan ve destek vektör sınırlarını oluşturan iki düzlemin bulunması gerekmektedir. Optimum hiperdüzlemin destek vektörlerine ait lineer doğru denklemleri denklem 16'da verilmiştir.

$$\vec{w}\vec{x} + b = \pm 1 \quad (16)$$

Optimum hiperdüzlemin sınırının maksimuma çıkarılması için pozitif ve negatif etiketli veriler arasındaki uzaklığın maksimum yapılması gerekmektedir. Problem denklem 17'de belirtilen sınırlı optimizasyon problemine dönüştürülür.

$$min\frac{1}{2}\|w\|^2 \quad (17)$$

kısıtlamalar: $y^{(i)}\left(\vec{w}\vec{x} + b\right) \geq 1$, i=1,..,m

SVM için hata minimizasyon fonksiyonu aşağıdaki denklem ile ifade edilmektedir.

$$C(w) = \frac{1}{2}\|w\|^2 + c\sum_{i-1}^{n} \xi_i \quad (18)$$

kısıtlamalar: $y^{(i)}\left(\vec{w}\,\vec{x}+b\right)\geq 1-\xi_i$, ve $\xi_i \geq 0$

c > 0 parametresi sınıflandırıcı hata dengesini oluşturmaktadır, modelin karmaşıklığı vektörler arasındaki mesafe ve arttıran yapay değişkenler (c ve ξi parametreleri) ile kontrol edilmektedir. Bu değişken yanlış sınıflandırma durumunun cezasıdır öyle ki eğitim verisi hatasız sınıflandırılamadığı zaman, yanlış tarafta sınıflandırılan verinin destek vektörüne ne kadar uzakta olduğunu kontrol etmek için kullanılır. Eğer çok uzaktaysa bu veri uzayı lineer olarak ayrıştırılamıyor anlamına gelmektedir, bu durumda lineer olmayan bir ayrıştırma vektörü kullanılır. Lineer olmayan ayrıştırma çekirdek olarak adlandırılan fonksiyonlar ile gerçekleştirilir [24].

SVM'de en yaygın kullanılan iki çekirdek fonksiyonu lineer ve Radial Based (RBF) çekirdek fonksiyonlarıdır. Çalışmamızda denklem 19'da belirtilen RBF çekirdek fonksiyonu kullanılmıştır.

$$K(x,l)=\exp\left(-\frac{\|x-1\|^2}{2\sigma^2}\right)=\exp\left(\frac{\sum_{j=1}^{n}\left(x_j-l_j\right)^2}{2\sigma^2}\right) \quad (19)$$

3. Deneysel Sonuçlar

3.1. Gen Dizilim Veri Kümeleri

Bu makalede kolon kanseri [32] ve lenfoma (Diffuse large B-cell lymphoma-DLBCL) kanseri [33] gen dizilimi veri kümesi kullanılarak deneysel çalışma gerçekleştirilmiştir. Kolon kanseri veri kümesi normal ve kanserli hücrelerden alınmış örnekleri içermektedir. Toplam 62 örnek içerisinden 40 tanesi kanserli hastalardan alınan biyopsilerden elde edilen ve "negatif" olarak etiketlenen; 22 tanesi normal hücrelerden alınan ve "pozitif" olarak etiketlenen örneklerden oluşmaktadır. Veri kümesi 6500 gen ifadesi içerisinden hesaplanan en iyi güven değerlerine görese çilmiş olan 2000 özellik içermektedir [32]. Genlerin hangi skor değerlerine göre seçildiği detaylı olarak Princeton Üniversitesine ait gen anlatım projesi sayfasından incelenebilir [34].

Lenfoma kanseri veri kümesi, 4000 özellik ile kaydedilmiş 47 örnek içermektedir. Bu örnekler içerisinden 24 tanesi "germinal centre B-like", 23 tanesi "activated B-like" kategorilerine dahil edilmişlerdir.

3.2. Gen Seçimi Ve Sınıflandırma Sonuçları

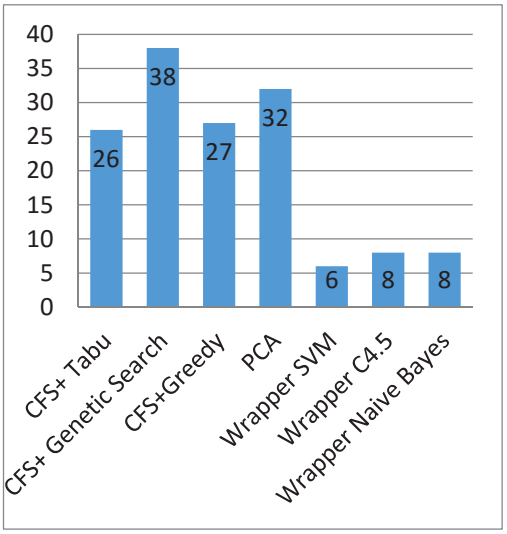
Çalışma kapsamında uygulanan algoritmalar WEKA [35] ortamında gerçekleştirilmiştir. Sınıflandırma algoritmalarının doğruluğu 10'lu çapraz geçerlilik (cross-validation) yöntemi ile test edilmiştir. Kernel LR ve SVM sınıflandırıcıda çekirdek fonksiyonu olarak Gauss çekirdek fonksiyonu seçilmiştir. SVM sınıflandırıcıda destek vektörleri arasındaki uzaklığı maksimum yapacak optimum cost (*c*) ve σ parametrelerinin belirlenebilmesi için *c* [1- 100] aralığında, σ [0,01-10] aralığında değerler test edilmiştir. Uygulamalarda kullanılan her iki gen dizini veri kümesinde, SVM sınıflandırıcı için gerçekleştirilen optimizasyon işlemleri sonucunda cost (c) ve σ parametreleri sırasıyla 35 ve 0.03 olarak seçilmiştir.

Şekil 1 ve Şekil 2'de sırasıyla kolon kanseri ve lenfoma kanseri gen dizini özelliklerinin azaltılması için uygulanan özellik azaltma algoritmalarının sonucunda elde edilen özellik sayıları gösterilmektedir.

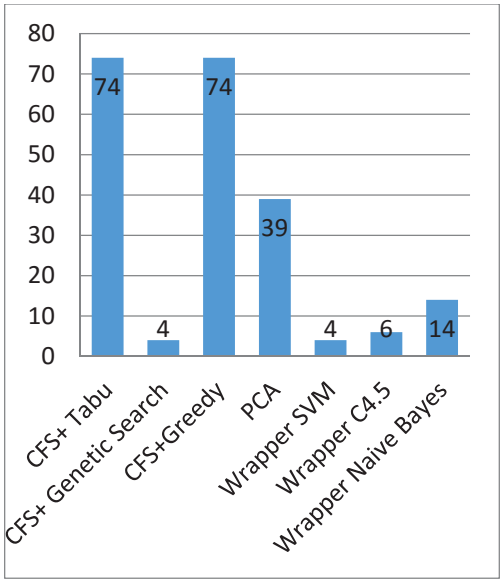
İki sınıflı gen dizini verilerinin sınıflandırılabilmesi için 50 veya biraz daha fazla bilgi verici özelliğin seçilmesini yeterli bulan [36] çalışmaların yanı sıra bir veya iki genin hastalığın tanımlanmasında yeterli olduğunu gösteren çalışmalar [37][38] da mevcuttur. İdeal bir sayı bulunmamakla birlikte seçilecek olan gen sayıları veri kümelerine göre farklılık göstermektedir. Her gen dizilimi veri kümesi kendine özgü gen özellikleri içerdiği için farklı sayılarda seçilen genler ile başarılı sınıflandırma yapılabilmektedir [38].

Çalışmamızda, Multinominal LR, Kernel LR, Simple LR ve SVM sınıflandırma algoritmaları, farklı özellik azaltma yöntemleri sonucu elde edilen özellik kümelerine uygulanmışlardır. Özellik azaltma yöntemlerinden uygulananlar ve bunlar için elde edilen sınıflandırma doğruluk oranları kolon kanseri veri kümesi için Şekil 4' de lenfoma veri kümesi için şekil 5'de gösterilmektedir. Kernel LR sınıflandırıcının γ ve σ parametreleri için optimum değerler 0.02 ve 0.2 olarak belirlenmiştir.

Kolon kanseri veri kümesinden elde edilen sonuçlara göre Multinominal LR, Kernel LR ve Simple LR yöntemleri farklı özellik azaltma teknikleri ile uygulandığında SVM sınıflandırıcıdan daha başarılı sonuçlar vermektedir. LR' nun en başarılı sonuçları sezgisel arama tekniklerinden olan Tabu arama ile kullanıldığında %90,9 doğrulukta başarı elde edildiği görülmektedir.



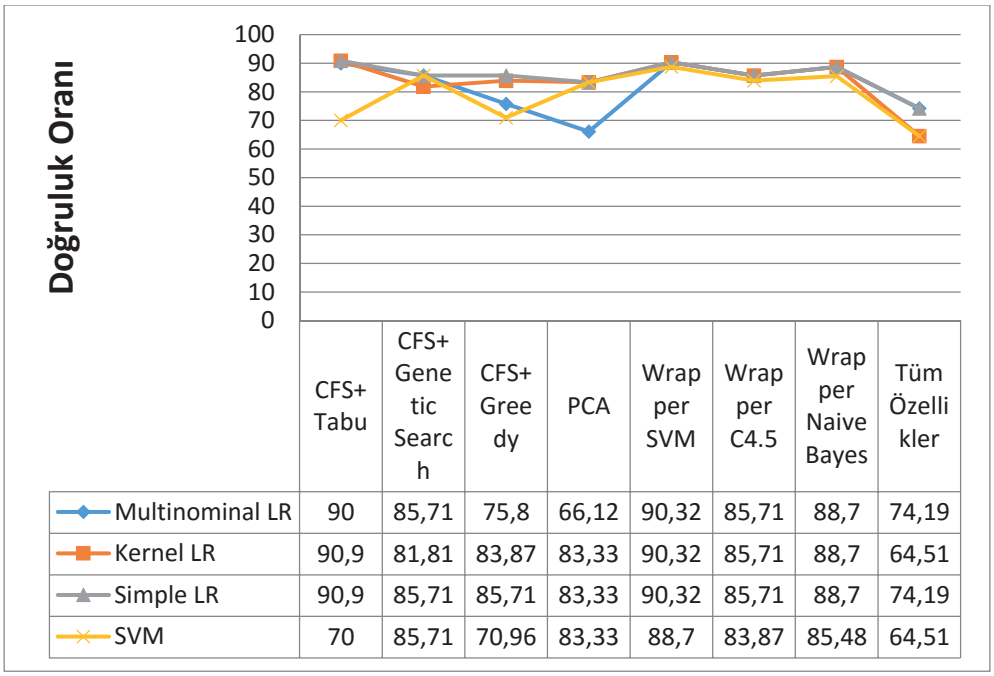
Şekil-1:Kolon kanseri için seçilen özellik sayıları



Şekil-2: Lenfoma kanseri için seçilen özellik sayıları

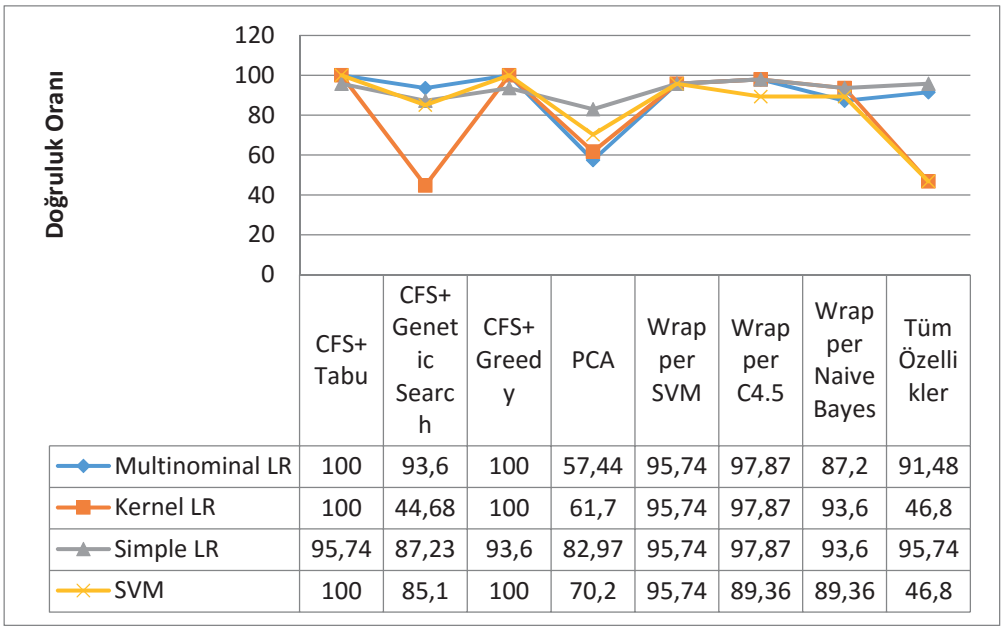
Şekil 4' de görüldüğü gibi Lenfoma kanseri tahmininde LR yöntemleri SVM kadar başarılı sonuçlara ulaşmışlardır, CFS algoritmasının TABU ve aç gözlü arama (Greedy search) ile birlikte kullanılarak özellik seçimi sonucunda LR ve SVM sınıflandırıcılarında benzer doğruluk oranları (%100) elde edilmiştir.

Gen verilerinin sınıflandırılması çalışmaları genelde SVM, NN, CT ve RF gibi makine öğrenmesi yöntemleri ile gerçekleştirilmiştir. Bu çalışmada LR sınıflandırıcının, literatürde güçlü bir sınıflandırıcı olarak kabul edilen SVM sınıflandırıcıdan daha iyi ya da yaklaşık doğrulukta sonuçlar verdiği gözlenmiştir. Bu çalışmanın temel katkısı gen dizilimlerinde özellik azaltma yöntemleri ile uygulanan LR sınıflandırıcısının güçlü bir tahmin aracı olduğu sonucuna varılmış olunmasıdır. Ayrıca, Tabu arama algoritmasının, daha önceki çalışmalarda kullanılmamış olmasına rağmen burada yüksek doğrulukla çalıştığı gözlenmiştir.



	CFS+ Tabu	CFS+ Genetic Search	CFS+ Greedy	PCA	Wrapper SVM	Wrapper C4.5	Wrapper Naive Bayes	Tüm Özellikler
Multinominal LR	90	85,71	75,8	66,12	90,32	85,71	88,7	74,19
Kernel LR	90,9	81,81	83,87	83,33	90,32	85,71	88,7	64,51
Simple LR	90,9	85,71	85,71	83,33	90,32	85,71	88,7	74,19
SVM	70	85,71	70,96	83,33	88,7	83,87	85,48	64,51

Şekil-3: Kolon kanseri



	CFS+ Tabu	CFS+ Genetic Search	CFS+ Greedy	PCA	Wrapper SVM	Wrapper C4.5	Wrapper Naive Bayes	Tüm Özellikler
Multinominal LR	100	93,6	100	57,44	95,74	97,87	87,2	91,48
Kernel LR	100	44,68	100	61,7	95,74	97,87	93,6	46,8
Simple LR	95,74	87,23	93,6	82,97	95,74	97,87	93,6	95,74
SVM	100	85,1	100	70,2	95,74	89,36	89,36	46,8

Şekil-4: Lenfoma kanseri

Çizelge-3. Öğrenme oranları ve farklı sınıfları öğrenme dengeleri

Özellik Azaltma Yöntemi	Veri Kümesi	Val	KernelLR	Simple LR	SVM	Multinominal LR
Özellik seçmeden	**Kolon Kanseri Gen Dizilimi**	F-measure	0,506	0,735	0,506	0,735
		Recall	0,645	0,742	0,645	0,742
		Presicion	0,416	0,735	0,416	0,735
		Acc(%)	**64,51**	**74,19**	**64,51**	**74,19**
	Lenf Kanseri Gen Dizilimi	F-measure	0,42	0,957	0,42	0,915
		Recall	0,468	0,957	0,468	0,915
		Presicion	0,444	0,957	0,444	0,918
		Acc(%)	**46,8**	**95,74**	**46,8**	**91,48**
CFS + Tabu	**Kolon Kanseri Gen Dizilimi**	F-measure	0,9	0,9	0,84	0,9
		Recall	0,91	0,9	0,839	0,9
		Presicion	0,89	0,91	0,843	0,9
		Acc(%)	**90,9**	**90,9**	**83,44**	**90**
	Lenf Kanseri Gen Dizilimi	F-measure	1	0,957	1	1
		Recall	1	0,957	1	1
		Presicion	1	0,957	1	1
		Acc(%)	**100**	**95,7**	**100**	**100**
CFS+ Genetic Search	**Kolon Kanseri Gen Dizilimi**	F-measure	0,819	0,857	0,858	0,858
		Recall	0,816	0,856	0,856	0,856
		Presicion	0,817	0,857	0,856	0,856
		Acc(%)	**81,81**	**85,71**	**85,71**	**85,71**
	Lenf Kanseri Gen Dizilimi	F-measure	0,42	0,872	1	0,936
		Recall	0,468	0,872	1	0,936
		Presicion	0,444	0,884	1	0,944
		Acc(%)	**46,8**	**87,23**	**100**	**93,6**
CFS+Greedy	**Kolon Kanseri Gen Dizilimi**	F-measure	0,904	0,889	0,809	0,759
		Recall	0,903	0,887	0,806	0,758
		Presicion	0,907	0,894	0,819	0,761
		Acc(%)	**90,32**	**88,7**	**80,64**	**75,8**
	Lenf Kanseri Gen Dizilimi	F-measure	1	0,937	1	1
		Recall	1	0,936	1	1
		Presicion	1	0,935	1	1
		Acc(%)	**100**	**93,6**	**100**	**100**
PCA	**Kolon Kanseri Gen Dizilimi**	F-measure	0,835	0,835	0,834	0,662
		Recall	0,833	0,833	0,833	0,661
		Presicion	0,831	0,831	0,832	0,66
		Acc(%)	**83,33**	**83,33**	**83,33**	**66,12**
	Lenf Kanseri Gen Dizilimi	F-measure	0,618	0,831	0,704	0,575
		Recall	0,617	0,829	0,702	0,574
		Presicion	0,616	0,826	0,7	0,574
		Acc(%)	**61,7**	**82,97**	**70,2**	**57,44**
Wrapper SVM	**Kolon Kanseri Gen Dizilimi**	F-measure	0,905	0,901	0,889	0,903
		Recall	0,903	0,903	0,887	0,903
		Presicion	0,901	0,905	0,885	0,903
		Acc(%)	**90,32**	**90,32**	**88,7**	**90,32**
	Lenf Kanseri Gen Dizilimi	F-measure	0,956	0,958	0,957	0,958
		Recall	0,958	0,957	0,957	0,957
		Presicion	0,958	0,956	0,957	0,956
		Acc(%)	**95,74**	**95,74**	**95,74**	**95,74**
Wrapper C4.5	**Kolon Kanseri Gen Dizilimi**	F-measure	0,858	0,858	0,839	0,858
		Recall	0,857	0,857	0,838	0,857
		Presicion	0,856	0,856	0,836	0,856
		Acc(%)	**85,71**	**85,71**	**83,87**	**85,71**
	Lenf Kanseri Gen Dizilimi	F-measure	0,979	0,979	0,893	0,979
		Recall	0,978	0,978	0,893	0,978
		Presicion	0,976	0,976	0,893	0,976
		Acc(%)	**97,87**	**97,87**	**89,36**	**97,87**
Wrapper Naive Bayes	**Kolon Kanseri Gen Dizilimi**	F-measure	0,888	0,888	0,855	0,888
		Recall	0,887	0,887	0,854	0,887
		Presicion	0,886	0,886	0,853	0,886
		Acc(%)	**88,7**	**88,7**	**85,48**	**88,7**
	Lenf Kanseri Gen Dizilimi	F-measure	0,937	0,937	0,893	0,871
		Recall	0,936	0,936	0,893	0,872
		Presicion	0,935	0,935	0,893	0,873
		Acc(%)	**93,6**	**93,6**	**89,36**	**87,2**

4. Tartışma ve Sonuçlar

Gen dizilimleri, DNA dizisi olan genlerin, fonksiyonel protein yapılarına dönüşmesi süreci için kullanılan terimdir. Farklı dokular üzerindeki gen dizilimlerinin farklı noktaları aktif olmaktadır. Dokularda aktif olan genler, üretilen protein yapılarının temelini oluşturduğundan, oluşan değişiklikler hücrelerin fonksiyonlarına etki etmektedir. Hasta olduğu düşünülen dokudan alınan örnek dizilimler ile başka hastalardan alınan örneklerin hasta veya sağlıklı olduğunun ya da hasta ise hangi tür hasta hücre olduğunun tespit edilmesinde büyük önem taşımaktadır [35].

Gen dizilimlerini sınıflandırma problemi, farklı hastalık türlerinin özellikle de kanser hastalıklarının tespitinde çok önemli bir yere sahiptir. Gen dizilimleri üzerinden tahmin yapmak ve kanserli ile kanserli olmayan genleri ayırt edebilmek binlerce gen söz konusu olduğunda oldukça zor bir problemdir. Literatürde bu problemin çözümünde en geçerli sonuçlar makine öğrenmesi yöntemlerinin uygulanması ile elde edilmiştir. Çalışma kapsamında gen dizilimlerinin sınıflandırmasında önceki çalışmalarda yaygın olarak kullanılmamış olan lojistik regresyon sınıflandırıcı önerilmiştir ve yöntem başarısı kanıtlanmış olan SVM sınıflandırıcı ile karşılaştırıldığında benzer başarı sonuçlarının elde edildiği gözlenmiştir.

Geleneksel sınıflandırma yöntemlerinden olan LR birçok tıbbi sınıflandırma problemlerinde kullanılmasına rağmen, gen dizilimlerinin analizi probleminde yeterince değerlendirilmemiştir. LR modeline ait eğitim ve test zamanları özellik sayısı ve eğitim verilerinin sayılarına göre lineer olarak artmaktadır. LR sınıflandırıcısının başarılı olabilmesi için özellik sayısının azaltılması önem taşımaktadır. Gerekli bilgiyi içeren genleri seçebilmek için sezgisel arama teknikleri ve farklı özellik azaltma yaklaşımları (filter, wrapper, vb.) kullanarak, LR sınıflandırıcısının SVM kadar başarılı sonuç verdiği görülmektedir. Elde edilen sonuçlara göre PCA ile yapılan özellik azaltma işleminden sonra elde edilen başarı oranlarının en düşük olduğu fakat CFS + Sezgisel yöntem ile en yüksek başarı oranlarının elde edildiği, Wrapper tekniği ile özellik azaltıldığı durumlarda ise LR'nun SVM' e göre daha başarılı tahminde bulunduğu görülmektedir.

5. Kaynaklar

[1] **Ben-Dor, A., Shamir, R., Yakhini, Z.,** 1999, Clustering gene expression patterns , J Comput Biol, 6(3):281–97.

[2] **Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A.,** 2000, Signaling and circuitry of multiple Mapk pathways revealed by a matrix of global gene expression profiles, Science, 287: 873–80.

[3] **Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z.,** 2000, Tissue classification with gene expression profiles,In: Proceedings of the Fourth International Conference on Computational Molecular Biology. Tokyo: Universal Academic Press.

[4] **Alizadeh, A., Eisen, M.B., Davis, R.E., Ma C Lossos, I.S., Rosenwald, A.,** 2000, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 403:503–11.

[5] **Wang, X., Gotoh, O.,** 2010, A robust gene selection method for microarray-based cancer classification, Cancer Inf, 9:15–30.

[6] **Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J. S.,** 2005, Incremental wrapper-based gene selection from microarray datafor cancer classification, Pattern Recognition**,** 39: 2383 – 2392.

[7] **Langley, P.,** 1994, Selection of relevant features in machine learning, In: Proceedings of the AAAI Fall Symposium on Relevance.

[8] **Kohavi, R., John,G.,** 1997, Wrappers for feature subset selection, Artif. Intell.1–2: 273–324.

[9] **Alter, O., Brown, P.O., Botstein, D.,** 2000, Singular value decomposition for genomewide expression data processing and modeling, Proc. Natl. Acad. Sci., 97(18).

[10] **Cangelosi, R., Goriely, A.,** 2007, Component retention in principal component analysiswith application to cdna microarray data, Biol. Direct, 2:1–21.

[11] **Liu, K., Li, B., Wu,Q.Q., Zhang, J. , Du, J.X., Liu,G.Y.,** 2009, Microarray data classification based on ensemble independent component selection, Comput. Biol. Med.,39(11):953–960.

[12] **Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A., 2004,**Filter versus wrappergene selection approaches in DNA microarray domains. Artif. Intell. Med., 31(2): 91–103.

[13] **Pohar, M., Blas, M., Turk, S.,** 2004, Comparison of Logistic Regression and Linear. Discriminant Analysis: A Simulation Study", Metodološki zvezki, 1:143-161.

[14] **Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., Mendonça, A.,** 2011, Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests, BMC Research Notes, 4:299.

[15] **Hall, M.A., Smith, L.A.,** 1997, Feature subset selection: A Correlation Based Filter Approach, In International Conference on Neural Information Processing and Intelligent Information Systems. Berlin: Springer, 855-858.

[16] **Jackson, J.,** 1991, A users guide to principal components, Wiley & Sons, New York.

[17] **Loh, W.,** 2006, Logistic regression tree analysis, In Springer Handbook of Engineering Statistics, 537-551.

[18] **Breiman, L., Friedman, H., Olshen, J., Stone, C.,** 1984, Classification and Regression Trees, Belmont, CA: Wadsworth.

[19] **Le Cessie, S., Van Houwellingen, J.C.,** 1992, Ridge Estimators in Logistic Regression, University of Leiden, the Netherlands. Appl. Statist., 41(1): 191-201.

[20] **Liu, D., Ghosh, D., lin, X.,** 2008, Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models, BMC Bioinformatics.

[21] **Bartenhagen, C.,Klein, H.U.,Ruckert, C.,Jiang, X.,Dugas, M.,** 2010, Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data, BMC Bioinformatics, 11(567).

[22] **Kim , K.J., Cho , S.B.,** 2006, Ensemble classifiers based on correlation analysisfor DNA microarray classification, Neurocomputing, 70:187-199.

[23] **Nguyen, D.V., Rocke, D.M.,**2002, Tumor classification by partial leastsquares using microarray gene expression data, Bioinformatics, 18: 39–50.

[24] **Cortes, C., Vapnik, V.,** 1995, Support-Vector Networks, Machine Learning, 20: 273-297.

[25] **Smith, L.I.,** 2002, A tutorial on Principal Components Analysis.

[26] **Dagliyan, O., Uney-Yuksektepe, F., Kavakli, I.H, Turkay, M.,** 2011, Optimization Based Tumor Classification from Microarray Gene Expression Data.

[27] **Vimaladevi, M., Kalaavathi, B.,** 2014, Cancer Classification using Hybrid Fast ParticleSwarm Optimization with BackpropagationNeural Network, International Journal of Advanced Research in Computer and Communication Engineering, 3(11).

[28] **Paulya, F., Smedbyc, K.E., Jerkemand, M., Hjalgrime, H., Ohlssonf, M., Rosenquist, R., Borrebaecka, C.A.K., Wingrena, C.,** 2014, Identification of B-cell lymphoma subsets by plasma protein profilingusing recombinant antibody microarrays, Leukemia Research, 38: 682–690.

[29] **Yan, Z., Li, J.Xiong, Y., Xu, W., Zheng, G.,** 2012, Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data, Oncology Reports, 28: 1036-1042.

[30] **Thorsteinsson, M., Kirkeby, L.T., Hansen, R., Lund L.R., Sørensen L.T., Gerds, T.A., Jess, P., Olsen, J.,** 2012, Gene expression profiles in stages II and III colon cancers:

application of a 128-gene signature, Int J Colorectal Dis, 27: 1579–1586.

[31] **Bennet, J., Ganaprakasam, C.A., Arputharaj, K.,** 2014, A Discrete Wavelet Based Feature Extraction and Hybrid Classification Technique for Microarray Data Analysis, Hindawi Publishing Corporation The Scientific World Journal.

[32] **Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J.,** 1999, Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA, 96(12):6745-6750.

[33] **Alizadeh,**A.A., , 2000, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 403:503-511.

[34] http://genomics-pubs.princeton.edu/oncology

[35] www.cs.waikato.ac.nz/ml/weka/

[36] **Golub, T.R., Slonim, D.K.,** 1999, Molecular classification of cancer:class discovery and class prediction bygene expression monitoring, Science, 286, 531-537.

[37] **Li, W.,Yang Y.,** 2000, How many genes are needed for a discriminant microarray data analysis?, Critical Assessment of Techniques for Microarray Data Mining Workshop, 137-150.

[38] **Xiong, M., Fang, Z., Zhao, J.,** 2001, Biomarker identification by feature wrappers, Genome Research, 11, 1878-1887.

[39] **Ding, C., Peng, H.,** 2003, Minimum Redundancy Feature Selection from Microarray Gene Expression Data, IEEE, 0-7695-2000-6/03