Marmara Fen Bilimleri Dergisi 2015, 4: 125-134
DOI: 10.7240/mufbed.89154

Yumurtalık Kanseri Veri Kümesindeki Gen İfadelerinin Veri Madenciliği ile Analizi

Hatice Zehra DEMİRCİOĞLU, Hasan Şakir BİLGE

Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Maltepe, Ankara

ÖZET

Biyoinformatik, moleküllerle ilgili bilgiyi anlamak ve düzenlemek için bilgisayar bilimleri, matematik, istatistik, biyoloji ve fizik gibi karmaşık disiplinlerden türetilen bir alandır. Biyoinformatiğin temel görevi, genomu verilen bir organizma ile ilgili fonksiyonların anlaşılması ve yaşam kalitesinin yükseltilmesidir. Biyoinformatik veritabanları nükleotid dizileri, protein dizileri, makro moleküler üç boyutlu (3D) yapılar gibi farklı veri türlerinden oluşabilmektedir. Biyoinformatik alanında elde edilen devasa boyuttaki verileri veri madenciliği yöntemleri kullanarak işlemek büyük önem kazanmaktadır. Bu çalışmada bioinformatik alanında veri madenciliğinin günümüz disiplinleri arasında geldiği noktaya değinilmiş ve kanser veri kümeleri ile veri madenciliği üzerine yapılan çalışmalar ve gerçekleştirilen uygulamalar incelenmiştir. Yapılmış uygulamalar ışığında yumurtalık kanseri verilerinin çeşitli öznitelik seçme ve sınıflandırma yöntemleri ile modellenerek algoritmaların doğruluk oranları incelenip karşılaştırılmıştır.

Anahtar kelimeler: Biyoinformatik, veri madenciliği, öznitelik seçimi, yumurtalık kanseri, boyut indirgeme, sınıflandırma.

Analysis of Gene Expressions in Ovarian Cancer Data Set by using Data Mining

ABSTRACT

Bioinformatics is a field that is derived from the complicated disciplines such as computer sciences, mathematics, statistics, biology, and physics, in order understand and organize the knowledge with molecules. The fundamental role of bioinformatics is to understand the organism that is given its genomes and to increase the quality of the standard life. Bioinformatics data bases may consist of different data types such as nucleoid sequences, protein sequences, 3D structures of the macro molecules. The processing of the huge amount of bioinformatics data is one of the exciting area for researchers. In this study, state-of-the-art of the data mining in bioinformatics is shortly explained and the studies are investigated that have performed on the cancer databases by using data mining. Ovarian cancer database is used and different feature selection and classification methods are implemented and the results are compared.

Keywords: Bioinformatics, data mining, feature selection, ovarian cancer, dimension reduction, classification.

I. GİRİŞ

Son yıllarda biyoinformatik, moleküler biyoloji ve DNA genom teknolojilerinin gelişmesiyle birlikte ortaya çıkan yeni bir bilim dalıdır. Çığ gibi büyüyen genetik araştırmalarda büyük verilerle uğraşıldığı için bunları depolayacak geniş veritabanlarına ihtiyaç duyulmuştur. Bu veritabanları ve bilgisayarlı hesaplamalar kullanılarak biyolojik problemlerin çözümlenmeye çalışılması biyoinformatik olarak tarif edilmiştir [1].

Etkin çoklu bir dizi hizalama işleme için önerilen ilk yaklaşım [2], bu yaklaşımın CLUSTAL'da uygulaması [3], protein yapı analizi ve tahmininde ilk yapay zeka (kural tabanlı uzman sistemlere benzer işlemleri kullanan) uygulamalarından biri [4], Karlin'in istatistiksel çalışmasına dayalı bir dizi eşleştirme algoritmasının uygulaması [5], threading kullanılarak protein yapı tahmininin ilk uygulaması [6] gibi çalışmalar biyoinformatik alanında yapılan çalışmalardan bazılarıdır.

Biyoinformatiğin gelişimiyle büyük boyuttaki verilerin depolanması amacıyla gen bankaları kurulmuştur. Bu veri bankalarından önemli ve anlamlı verilerin elde edilebilmesi için veri madenciliği yöntemleri kullanılmaktadır. Veri madenciliğinin daha iyi anlaşılmasında farklı tanımların incelenmesi etkili olacaktır. Veri madenciliği anlamlı örüntü

Sorumlu yazar/Corresponding author: Hasan Şakir BİLGE, Tel: +90 (312) 582 3346, **e-posta:**bilge@gazi.edu.tr,
Gönderilme/ Submitted: 29.05.2015, **Düzenleme/Revised:** 24.11.2015, **Kabul/Accepted:** 30.12.2015

ve kuralları bulmak için büyük miktardaki verilerin analizi ve keşfidir [7]. Veritabanlarında saklı kalmış verilerin istatistik, matematik ve örüntü tanıma teknikleri kullanılarak gözden geçirilmesiyle yeni ilişki ve örüntülerin bulunması işlemidir [8].

II. VERİ MADENCİLİĞİ

Veri madenciliği aşamaları aşağıda kısaca özetlenmiştir.

2.1. Problem Tanımlama

Bu aşamada veri madenciliği ile sağlanacak bilgi ihtiyaçları tanımlanmaktadır. Örüntülerle ilgili sorular ve veritabanında oluşabilecek ilişkilerdir. Veri madenciliği, birçok nitelik arasındaki var olabilecek ilişkilerin incelenmesi durumunda, kendi sorusunu sunar. Böylece, sonuca gelindiğinde tahmin edilemeyen ilişkilerin bulunmasını sağlar [9].

2.2. Veri Anlama

Veri toplama ile başlayan bu aşamadaki veriler, veri madenciliğinde ham verilerdir. Bu nedenle veri kalitesinin tanımlanması, veri içeriğinin anlaşılması, gizli bilgiden yeni hipotezler oluşturularak farklı değerlendirmelerin yapılması bu aşamadaki adımlardır.

Farklı kaynaklardan gelen verilerin anlaşılmasından önce birbiriyle bütünleşmiş olması gerekir. Daha sonra, tablolardaki birincil anahtar bilgileri düzgün bir şekilde girilmiş olması gerekir, girilmediği takdirde veri tutarsızlıkları olabilir [10, 11].

2.3. Veri Önişleme

Problem ve hedefler tanımlandıktan sonra sıra veri hazırlama aşamasındadır. Veri önişleme aşaması, verilerin veri madenciliği için hazırlanmasını kapsamaktadır. Veri hazırlama görevi uzun sürede yapılmaktadır. Verinin dönüşümü, temizlenmesi, birleştirilmesi, azaltılması gibi işlemleri içermektedir [12].

2.4. Veri Temizleme

Veri temizleme, eksik değerleri tamamlama, aykırı değerleri belirleme ile gürültüyü azaltma ve verilerdeki tutarsızlıkları giderme gibi birçok teknik içermektedir. Veri madenciliğindeki kirli veriler karmaşaya ve sonuçların güvenilir olmamasına sebebiyet verir. Bu yüzden veri temizleme işlemleri ardından temizlenmiş verilerin kullanılması gerekmektedir.

2.5. Modelleme

Modelleme fazı temel olarak uygun modelleme tekniklerinin belirlenmesi ve uygulanması, eniyileme için model değişkenlerinin düzenlenmesinden oluşur. Gerektiği durumlarda veri hazırlama fazına dönülebilir ve aynı veri madenciliği problemi için birden fazla teknik kullanılabilir [11].

2.6. Değerlendirme

Modelleme yapıldıktan sonra bu modelin başlangıçta belirlenen iş hedeflerinin ne kadarını karşıladığı ölçülmeli ve ortaya çıkan sonucun yaygınlaştırma fazından önce kalite ve etkisi değerlendirilmelidir. Bununla birlikte problemde ele alınacak noktaların yeterli derecede dikkate alınıp alınmadığı kontrol edilmeli ve sonuçların kullanılıp kullanılmayacağı ile ilgili net karar verilmelidir [10, 11, 13].

2.7. Yaygınlaştırma

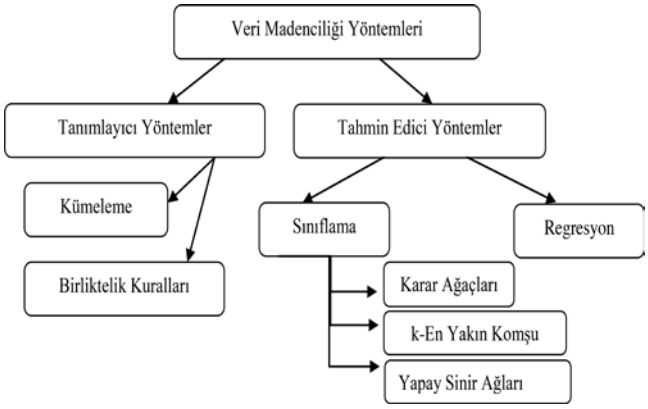
Yaygınlaştırma fazının başarı oranı oluşturulan modelden yararlanılması ile doğru orantılıdır. Ayrıca bu aşamada veri madenciliği çalışmasının sonuçlarının varsa proje sahibine raporlanması gerekmektedir. Veri madenciliği çalışması değerlendirilmesi gereken yeni bir bilgiyi ortaya çıkarır ve bu bilginin proje hedefleri ile birleştirilmesi gerekir [13, 14].

Veri madenciliğinde kullanılan yöntemler, hiyerarşik olarak Şekil 1'deki gibi ele alınabilir.

III. LİTERATÜRDEKİ ÇALIŞMALAR

Chen ve ark. (2014) gen seçimi için karar ağacı sınıflandırıcı ile birlikte parçacık sürüsü optimizasyonunu (PSO) veri kümesinde bulunan binlerce gen içinden daha az sayıda olan bilgi verici olan genleri seçmek için yeni bir model olarak önermişlerdir. Önerdikleri metodun başarısını destek vektör makineleri, kendi kendini düzenleyen harita, geriye yayılım sinir ağı, karar ağacı gibi bilinen yöntemlerin sınıflandırma başarıları ile karşılaştırıldığında daha üstün başarı elde etmişlerdir. Biri Tayvan Ulusal Sağlık Sigortası Araştırma Veritabanından elde edilmiş veriler olmak üzere diğerleri internetten ulaşılabilecek veriler olmak üzere 11 farklı kanser veri kümesi ile çalışmışlardır. Bilinen sınıflandırıcılardan olan destek vektör makineleri ile %72,46, kendi kendini düzenleyen harita ile %52,60, geriye yayılım sinir ağı ile %42,58, karar ağacı ile 93,14 elde edilirken geliştirilen PSO-C4.5 metodu ile %97,26 oranında sınıflandırma başarısı elde etmişlerdir. PSO parametre ayarları ve yerel optimum yakalama sorunu üzerinde daha fazla çalışmalar yapılması gerekmektedir. Genetik algoritma ile bir hibrit metod geliştirilebilir. Bu hibrit metoddaki genetik algoritmanın mutasyon operatörü kullanılarak parçacıkların çeşitlilik göstermesi sağlanarak yerel optimum sorunu çözülebileceği söylenerek gelecekte yapılabilecek çalışmalara dikkat çekmişlerdir [15].

Thanh ve ark. (2015), denetimli öğrenen gizli Markov modeli tasarımı ile elde edilen gen ifade profilleri ile kanser sınıflandırmasına bir yaklaşım sunmaktadırlar. Her bir tümör tipi, gen ifadesi veri olasılığını maksimum yapan gizli Markov model ile modellenmiştir.



Şekil 1. Veri madenciliği yöntemleri.

Bilinen farklı genler analitik hiyerarşi sürecinin (AHP) değişikliğine dayanan yeni bir metod ile seçilmektedir. Geleneksel AHP metodu aksine, değiştirilmiş AHP her bir gen seçimi metodunun sonuçlarının sıralanmasını sağlamaktadır. Gen seçimi metodu olarak t-testi, entropi, alıcı işletim karakteristik eğrisi, Wilcoxon testi ve sinyal gürültü oranı kullanılmaktadır. Değiştirilmiş AHP istikrarlı ve kararlı bir gen alt kümesi oluşturmak için her bir gen seçim metodunun sıralama sonuçlarını birleştirmektedir. Deneysel çalışmalarda gizli Markov model yaklaşımının diğer altı sınıflandırma metoduna göre daha iyi performans gösterdiği görülmüştür. Sonuç olarak AHP ile oluşturulan gen alt kümesi, bilgi kazancı, simetrik belirsizlik, Bhattacharyya uzaklığı ve Relief gibi diğer gen seçim metotlarından daha fazla doğruluk ve kararlılık göstermiştir. Değiştirilen AHP sadece gizli Markov modelleme (HMM) sınıflandırıcısının değil diğer sınıflandırıcıların da sınıflandırma performanslarını artırmıştır. Lösemi, bağırsak kanseri, prostat kanseri, DLBCL (Diffuse Large B-Cell Lymphomas) verileri üzerinde k en yakın komşu (kNN), olasılıklı sinir ağı (PNN), destek vektör makineleri (SVM), çok katmanlı algılayıcı (MLP), bulanık ARTMAP (FARTMAP), grup öğrenen AdaBoost ve yeni önerilen hibrit (AHP-HMM) metodun sınıflandırma başarıları karşılaştırılmıştır. HMM %2,20 ile diğer sınıflandırıcılar arasında en küçük AUC (Area Under Curve) standart sapma değerini vermiştir. DLBCL verisinin HMM ile sınıflandırılması ile diğer sınıflandırıcı sonuçları arasında en yüksek sonuç olan %98,83 doğruluk ve %98,14 AUC değeri elde edilmiştir [16].

Jin ve ark. (2015), yaptıkları çalışma ile çoklu destek vektör veri açıklama tabanlı hızlı bir öznitelik seçme metodu önermektedir. Tekrarlı olarak ilgisiz öznitelikler çıkarılarak özyinelemeli bir öznitelik eleme tasarısı önerilmektedir. Önerilen metot çoklu SVDD-RFE (MSVDD-RFE)' dir. Bu metot her bir sınıf için alakalı gen alt kümesini bağımsız bir şekilde seçmektedir. Bu seçilen alakalı gen alt kümeleri birleşerek nihai gen alt kümesini oluşturmaktadır. MSVDD-RFE metodunun etkinliği ve doğruluğu beş genel mikro dizilim veri kümesi üzerinde geçerliliği sağlanmıştır. Bu önerilen metot diğer metotlardan daha hızlı ve daha efektiftir. Leukemia, Colon, Tumor ve Novartis veri kümeleri üzerinde ortalama %90 üzerinde başarı yakalanmıştır. Lung Cancer veri kümesinde istenilen başarı yakalanamamıştır. Önerilen metodun bu sınıf üzerindeki sınıflandırma başarısını artırmak için metoda grup öğrenmesi çalışması eklenmesi düşünülmektedir [17].

H. Banka ve S. Dara (2015), ikili parçacık sürüsü optimizasyonu tabanlı Hamming uzaklığı yöntemi önermişlerdir. Hamming uzaklığı, önemli öznitelikleri seçmek için ikili parçacık sürüsü optimizasyonundaki parçacık hızlarını güncelleme amacıyla yaklaşık değer olarak verilmektedir. Hesaplanan yaklaşık değer Hamming uzaklıkları kullanan HDBPSO yöntemiyle gen ifade verilerindeki önemli öznitelik alt kümelerinin daha iyi performans ile bulunabileceği görülmüştür. Leukemia, Colon, DLBCL veri kümeleri üzerinde önerilen HDBPSO öznitelik seçme yöntemi uygulanarak çeşitli sınıflandırıcılar ile bu yöntemin başarısı ölçülmüş ve diğer öznitelik seçme yöntemlerinin başarıları ile karşılaştırılmıştır. Verilerin %50'si eğitim, %50'si test olarak kullanılmış ve 10 kat çapraz doğrulama yapılmıştır. Colon veri kümesi için önerilen metot ile elde edilen öznitelik alt kümesi LibLinear, SVM, MLP ve J48 sınıflandırıcılar ile sınıflandırıldığında %100 başarı göstermiştir. Lymphoma veri kümesi için önerilen metot LibLinear sınıflandırıcı ile %100 başarı göstermiştir. Leukemia veri kümesi için ise önerilen metot LibLinear, SVM, RF ve MLP sınıflandırıcı ile %100 başarı göstermiştir. Her bir veri kümesi için diğer öznitelik seçme yöntemlerinin sınıflandırma başarıları, önerilen öznitelik seçme yönteminin sınıflandırma başarısından kötü çıkmıştır [18].

E. Lotfi and A.Keshavarz (2014) gen ifade verilerinin sınıflandırılması için temel bileşen analizi (PCA) ve beyin duygusal öğrenme (BEL) ağı tabanlı yeni hibrit bir yöntem önermişlerdir. BEL ağı nöropsikolojik özellikleri yansıtan duygusal beynin sayısal sinir modeli halidir. Bu sınıflandırıcının önemli bir ayırt edici özelliği hesaplama

karmaşıklığı diğer sınıflandırıcılardan daha az olmasıdır. Çalışmada 5 kat çapraz doğrulama kullanılmıştır. Yeni önerilen hibrit PCA-BEL yöntemi ile küçük yuvarlak mavi hücreli tümörler (SRBCTs), yüksek dereceli gliomalar (HGG), akciğer (lung), kolon (colon) ve meme (breast) kanseri veri kümeleri sınıflandırılarak bulunan sınıflandırma başarıları sırasıyla %100, %96, %98,32, %87,40 ve %88' dir [19].

Devi ve ark. (2015) karşılıklı bilgi (MI) tabanlı gen seçimi ve destek vektör makineleri (SVM) kullanarak hibrit bir yöntem önermişlerdir. Genler ve sınıf etiketleri arasındaki karşılıklı bilgi önemli genleri anlamak için kullanılır. Seçilen genler SVM sınıflandırıcıyı eğitmek için kullanılmış ve sınıflandırıcının testi bir çıkarımlı çapraz doğrulama (LOOCV) kullanılarak değerlendirilmiştir. Lenfoma ve kolon kanser veri kümeleri üzerinde çalışılmıştır. Kolon kanseri veri kümesi için karşılıklı bilgi (MI) ile bulunan 3 gen ile eğitilen sınıflandırıcı doğrulukları kNN ile %61,29, ANN ile %61,29, SVM (doğrusal) ile %74,19, SVM (Radyal) ile %64,51, SVM (quad) ile %38,70, SVM (pol) ile %64,51 bulunmuştur. Lenfoma veri kümesi için karşılıklı bilgi (MI) ile bulunan alakalı 4 gen ile eğitilen sınıflandırıcı doğrulukları kNN ile %90,9, ANN ile %100, SVM (doğrusal) ile %100, SVM (Radyal) ile %90,9, SVM (quad) ile %86,36, SVM (pol) ile % 90,9 bulunmuştur [20].

Thanh ve ark. (2015), beş farklı istatistiksel yöntemin gen sıralama hesaplamasını ilişkilendirerek gen seçimi yapan yeni bir yöntem olarak değiştirilmiş analitik hiyerarşi yöntemini (MAHP) önermişlerdir. İki-örnek t testi, entropi testi, alıcı işletimi karakteristik eğrisi (ROC), Wilcoxon testi ve sinyal gürültü oranı olmak üzere beş farklı istatistiksel gene sıralama metodunun hesaplanan sonuçlarını karşılaştırarak gen seçimi yapmışlardır. Bilgi kazancı (IG), simetrik belirsizlik (SU), Relief ve Bhattacharyya uzaklığı (BD) öznitelik seçme yöntemleriyle kıyaslanmıştır. Bir çıkarımlı çapraz doğrulama (LOOCV) ile test ve eğitim kümeleri oluşturulmuştur. DLBCL, lösemi, prostat ve kolon kanseri verileri kullanılmıştır. Doğrusal Ayırma Analizi (LDA), k en yakın komşu (kNN), olasılıklı sinir ağı (PNN), destek vektör makineleri (SVM), çok katmanlı algılama (MLP) sınıflandırıcıları ile önerilen yöntemin ve mevcut bahsedilen diğer gen seçme metotlarının sınıflandırma başarıları her bir veri kümesi için ayrı ayrı bulunarak karşılaştırılmıştır. Lösemi veri kümesi için, en yüksek başarı %97,36 sınıflandırıcılar arasından kNN ve gen seçim metotları arasından bu çalışmada önerilen metot (MAHP) uygulandığında bulunmuştur. Kolon veri kümesi için, en yüksek başarı %87,9 LDA ve MAHP birlikte uygulandığında bulunmuştur. Prostat veri kümesi için, en yüksek başarı %91,18 LDA ve MAHP birlikte uygulandığında bulunmuştur. DLBCL veri kümesi için, en yüksek başarı %98,31 LDA ve MAHP birlikte uygulandığında bulunmuştur [21].

Dajun ve ark. (2014), çeşitli ilişkilerle en önemli genlerin seçimini yapmak için yeni ileri gen seçim algoritmasını (FGSA) önermektedirler. Beş kat çapraz doğrulama kullanılmıştır. Artrit veri kümesinde Elastik Net algoritması ile %88, İleri Gen Seçim Algoritması (FGSA) algoritması ile %91,85 başarı elde edilmiştir. Kolon veri kümesinde üstünde Elastik Net algoritması ile %93,69, FGSA algoritması ile %94,77 başarı elde edilmiştir. Lösemi veri kümesi üstünde FGSA sınıflandırıcı ile %98.41 başarı elde edilmiştir [22].

Jie ve ark. (2015), geliştirilmiş yer çekimi arama algoritmasına dayalı ikili problemlere uygun yeni hibrit bir yöntem önermişlerdir. Bu algoritma genel arama ve yerel aramayı hızlandırmak için sıralı karesel programlama yapmak için parçalı doğrusal kaotik haritalama yapar. Yerçekimi arama algoritmasına parçalı doğrusal kaotik haritalama (PWL) ve sıralı karesel programlama (SQP) algoritmaları katılarak geliştirilmiş yerçekimi algoritması (IGSA) ortaya konulmuştur. Bu algoritma UCI makine öğrenmesi sitesindeki çeşitli öznitelik seçen örneklerle yöntemleriyle karşılaştırılmış, daha az sayıda alakalı gen ve daha iyi başarı elde edilmiştir. Geliştirilmiş yer çekimi arama algoritması (IGSA) 23 lineer olmayan kıyaslama fonksiyonu ile ve 5 sezgisel algoritma ile karşılaştırılarak test edilmiştir. Bu sezgisel algoritmalar sırasıyla Genetik Algoritma (GA), ikili parçacık sürü optimizasyonu (BPSO), Kuantum esinli parçacık sürü optimizasyonu (QBPSO), ikili yerçekimi arama algoritması (BGSA), geliştirilmiş yerçekimi arama algoritması (IGSA)' dır. Wisconsin meme kanseri veri kümesinde önerilen BIGSA optimizasyonu kNN sınıflandırıcıda 26 gen ile diğer optimizasyonlar arasında en yüksek AUC değeri 98.1 elde edilmiştir. kNN sınıflandırıcı için k değeri 1 seçilmiştir. Bir çıkarımlı çapraz doğrulama (LOOCV) ile test ve eğitim kümeleri oluşturulmuştur. PIMA Diabetes veri kümesinde önerilen BIGSA optimizasyonu kNN sınıflandırıcıda 8 gen ile diğer optimizasyonlar arasında en yüksek AUC değeri 75.1 elde edilmiştir [23].

B. Chandra ve K.V. Naresh Babu (2014), dalgacık Radyal Tabanlı Sinir Ağı (WRNN)' nın gen ifade verilerine uygulamayı önermişlerdir. Çivileme fonksiyonu olarak doğrusal olmayan bütünleşmiş ve yangın model ve diğer çivileme aralığı türetilmiş ve Dalgacık Radyal Tabanlı Sinir Ağı (WRNN)' nda kullanılmış ve bu yeni modele Çivileme

Dalgacık Radyal tabanlı Sinir Ağı (SWRNN) adı verilmiştir. Karaciğer tümörü, Genel Kanser Haritası (GCM), Glioma, meme kanseri,11-tümor ve Hepato hücresi veri kümelerinde test edilmiştir. On kat çapraz doğrulama ile test ve eğitim kümeleri oluşturulmuştur. WRNN, standart metot ve SWRNN metodu ile yapılan sınıflandırma başarıları her bir veri kümesi için en yüksek olanı SWRNN metodundan elde edilmiştir. Veri kümelerinde sırasıyla %99,651, %99,79, %98,47, %96,02, %73,79 ve %97,77 sınıflandırma başarıları elde edilmiştir [24].

T.Latkowski ve S. Osowski (2015), en iyi temsil eden gen özniteliklerinin bulunarak sınıflandırıcıya girdi olarak verilmesini sağlayan farklı gen seçim yöntemlerinin bir uygulamasını sunmaktadır. Birkaç gen seçim metodu ile seçilen genler Genetik algoritma ve destek vektör makineleri birlikte uygulanarak başarı elde edilmiştir. 10 kat çapraz doğrulama yapılmış ve tüm verilerin %40'ı test ,%60'ı eğitim olarak kullanılmıştır. %60 seçilen eğitim verisine 8 farklı öznitelik seçme yöntemi uygulanmıştır. Fisher korelasyon analizi (FDA), ReliefF algoritması (RFA), iki örnek t testi (TT), Kolmogorov–Smirnov testi (KST), Kruskal–Wallistest (KWT), aşamalı regresyon metodu (SWR), sınıfla öznitelik ilişkisi (COR), SVM-RFE metodu olmak üzere 8 öznitelik seçme metodu ile seçilmektedir. Her bir yöntemde seçilen öznitelikler genetik algoritma ile tekrar seçilmektedir. Az sayıda bulunan en iyi özniteliklere ait test bilgileri destek vektör makineler sınıflandırıcısı ile sınıflandırılmakta ve başarımı ölçülmektedir. Önerilen yöntemle %86.07 ile diğer sınıflandırıcılar arasında en yüksek başarı elde edilmiştir. Ayrıca aynı sayıda en iyi değil de rastgele seçilen genlerin sınıflandırıcıdaki başarısı %67,16 bulunmuştur [25].

IV. MATERYAL VE YÖNTEM

Bu çalışmada halka açık bir veri kümesi olan yumurtalık kanseri gen veri kümesi kullanılmıştır (OvarianCancer-NCI-PBSII-061902) [26, 27]. Bu veri kümesinde 253 örnek ve 15154 tane gen bulunmaktadır. Bu 253 tane örneğin 91 tanesi sağlıklı, 162 tanesi hasta olarak tanımlamıştır. Yumurtalık kanseri gen ifade veri kümesindeki veriler normalize edilmiştir. Normalize edilen yeni veri kümesinde eksik değerler (NaN) olması ihtimali bulunduğundan 'NaN' değer bulunan genler veri kümesinden çıkarılabilmesi için filtreleme yapılmıştır. Ancak, yumurtalık kanseri gen ifade veri kümesinde eksik değerler olmadığından filtrelendikten sonra da örnek ve öznitelik sayısı aynı kalmıştır. Max-min normalizasyonu kullanılmıştır. Veriler normalize edilerek 0-1 değer aralığına çekilmiştir. Veri kümesinde çok sayıda gen bulunduğundan, öncelikle Fisher korelasyon skorlama (FKS) ve Weltch t testi olmak üzere iki farklı öznitelik seçme yöntemi kullanılarak gen öznitelik sayısı çıkan sonuçlar neticesinde ilgililik sıralaması yapılmıştır. Bu sıralamaya göre öznitelik kümesi ilk 100 örnek alınmış daha sonra ilk 200 örnek alınmış ve 100 artırımlı olarak tüm veri kümesinin sınıflandırma başarıları ölçülmüştür.

Fisher korelasyon skorlama öznitelikler arasındaki ilgiyi hesaba katmaksızın öznitelikler ait bilgiyi elde etmede kullanılan bir istatistiksel yöntemdir. Her bir sınıf için ayrı ayrı özniteliklere karşılık gelen değerlerin ortalaması alınır. Ayrıca, her bir özniteliğin standart sapma değerleri alınarak bir oranlama yapılarak öznitelikler bulunan değerlerle doğru orantılı olarak sıralandırılabilir. Weltch t test ise her bir sınıfın varyans değerlerinin örnek sayısına bölünerek toplanıp karekökünün alınması ve ortalama değerlerin farkına oranlanması ile elde edilmektedir.

V. BULGULAR VE TARTIŞMA

k en yakın komşu (kNN) ve destek vektör makineleri (SVM) sınıflandırıcıları kullanılarak çeşitli sınıflandırma başarıları elde edilmiştir. Sınıflandırma yapılırken %40'ı eğitim, %60'ı test verisi olarak rastgele örnekler seçilmiştir. Sınıflandırıcılardan kNN için 1, 3, 5, 7, 9 ve 11 olarak farklı k değerleri ve SVM için ise doğrusal, radyal, polinomsal ve karesel olarak farklı çekirdek fonksiyonları ile sınıflandırma başarıları ölçülmüştür. Tablo 1 ve Şekil 2' de FKS ile elde edilen değerlere göre ilk 100, ilk 200, ilk 300 ve 100 artırımlı olarak tüm veri kümesindeki genler alınarak kNN sınıflandırıcının farklı değerleri ile sınıflandırma başarıları gösterilmektedir. Tablo 2 ve Şekil 3'te FKS ile elde edilen değerlere göre ilk 100, ilk 200, ilk 300 ve 100 artırımlı olarak tüm veri kümesindeki genler alınarak SVM sınıflandırıcının farklı değerleri ile sınıflandırma başarıları gösterilmektedir. Tablo 3 ve Şekil 4' te WTS ile elde edilen değerlere göre ilk 100, ilk 200, ilk 300 ve 100 artırımlı olarak tüm veri kümesindeki genler alınarak kNN sınıflandırıcının farklı değerleri ile sınıflandırma başarıları gösterilmektedir. Tablo 4 ve Şekil 5'te WTS ile elde edilen değerlere göre ilk 100, ilk 200, ilk 300 ve 100 artırımlı olarak tüm veri kümesindeki genler alınarak SVM sınıflandırıcının farklı değerleri ile sınıflandırma başarıları gösterilmektedir.

Tablo 1'de Fisher korelasyon ile sıralanmış genlerden ilk 100 gen için en yüksek başarı değerleri elde edilmiştir. Gen seçilmeksizin tüm veri kümesine kNN (k=1 için) sınıflandırıcısı uygulandığında %88,74 bulunurken, ilk 100 gen için %99,34 başarı elde edilmiştir. k'nın farklı değerleri

için ilk 100 gende en iyi sınıflandırma başarısı k değerinin 1 alındığı durumda elde edilmiştir. En iyi sınıflandırma başarısı FKS ile ilk sıralanmış 1000 örneğe kadar kNN sınıflandırıcının k'nın farklı değerlerinin hepsinde de tüm veri kümesinden elde edilen başarıdan daha fazla başarı elde edilmiştir. Bu sonuçlardan önerilen ilgililik sıralaması yapılarak sıralandırılmış gen öznitelikleri seçilerek gen veri kümesinin boyutu indirgenerek daha az boyutlu veri kümesinin sınıflandırılması daha iyi başarı elde etmeye yaramıştır.

Tablo 1. FKS sıralı indirgenmiş gen verisinin kNN ile sınıflandırma sonuçları

FKS sıralı indirgenmiş gen verisi	kNN (k=1) %	kNN (k=3) %	kNN (k=5) %	kNN (k=7) %	kNN (k=9) %	kNN (k=11) %
İlk 100 gen	99,34	98,01	98,01	98,01	98,01	98,01
İlk 200 gen	98,68	98,01	98,01	98,01	96,69	96,69
İlk 300 gen	98,68	98,01	97,35	97,35	96,03	95,36
İlk 400 gen	98,68	97,35	96,69	96,69	95,36	96,03
İlk 500 gen	98,01	97,35	96,03	96,03	95,36	95,36
İlk 600 gen	97,35	96,69	95,36	95,36	94,70	94,04
İlk 700 gen	96,69	96,69	96,03	96,03	94,04	93,38
İlk 800 gen	97,35	96,03	96,03	95,36	94,04	94,04
İlk 900 gen	98,01	96,69	96,03	95,36	94,70	94,04
İlk 1000 gen	98,01	96,69	96,03	95,36	94,70	93,38
Tüm veri	88,74	86,75	86,75	87,42	85,43	85,43

Tablo 2'de Fisher korelasyon ile sıralanmış genlerden ilk 100 gen için en yüksek başarı değerleri elde edilmiştir. Gen seçilmeksizin tüm veri kümesine SVM (çf=lineer için) sınıflandırıcısı uygulandığında %98,68 bulunurken, ilk 100 gen için %100 başarı elde edilmiştir. SVM' nin farklı çekirdek fonksiyonları için ilk 100 gende en iyi sınıflandırma başarısı SVM çekirdek fonksiyonunun doğrusal (lineer) seçildiği durumda elde edilmiştir. En iyi sınıflandırma başarısı önerilen metod ile ilk sıralanmış 1000 örneğe kadar SVM sınıflandırıcının tüm çekirdek fonksiyonlarının hepsinde de tüm veri kümesinden elde edilen başarıdan daha fazla başarı elde edilmiştir.

Tablo 2. FKS sıralı indirgenmiş gen verisinin SVM ile sınıflandırma sonuçları

FKS sıralı indirgenmiş gen verisi	SVM (ÇF =Lineer)	SVM (ÇF =Polinom)	SVM (ÇF =Quadratik)	SVM (ÇF =Radyal)
İlk 100 gen	100,00	98,01	97,35	63,58
İlk 200 gen	100,00	98,68	96,03	63,58
İlk 300 gen	100,00	98,68	94,04	63,58
İlk 400 gen	99,34	98,68	94,04	63,58
İlk 500 gen	99,34	98,68	94,70	63,58
İlk 600 gen	99,34	98,68	94,04	63,58
İlk 700 gen	99,34	98,68	93,38	63,58
İlk 800 gen	99,34	98,68	93,38	63,58
İlk 900 gen	99,34	96,03	92,05	63,58
İlk 1000 gen	99,34	94,70	91,39	63,58
Tüm veri	98,68	63,58	43,71	63,58

Tablo 3'te Weltch t testi ile sıralanmış genlerden ilk 100 gen ve k=11 için en yüksek başarı değeri %100 sınıflandırma başarısı elde edilmiştir. Gen seçilmeksizin tüm veri kümesine kNN (k=11 için) sınıflandırıcısı uygulandığında % 85,43 bulunurken, ilk 100 gen için %100 başarı elde edilmiştir. En iyi sınıflandırma başarısı WTS ile ilk sıralanmış 1000 örneğe kadar kNN sınıflandırıcının k'nın farklı değerlerinin hepsinde de tüm veri kümesinden elde edilen başarıdan daha fazla başarı elde edilmiştir. Bu sonuçlardan önerilen ilgililik sıralaması yapılarak sıralandırılmış gen öznitelikleri

seçilerek gen veri kümesinin boyutu indirgenerek daha az boyutlu veri kümesinin sınıflandırılması daha iyi başarı elde etmeye yaramıştır.

Tablo 3. WTS sıralı indirgenmiş gen verisinin kNN ile sınıflandırma sonuçları

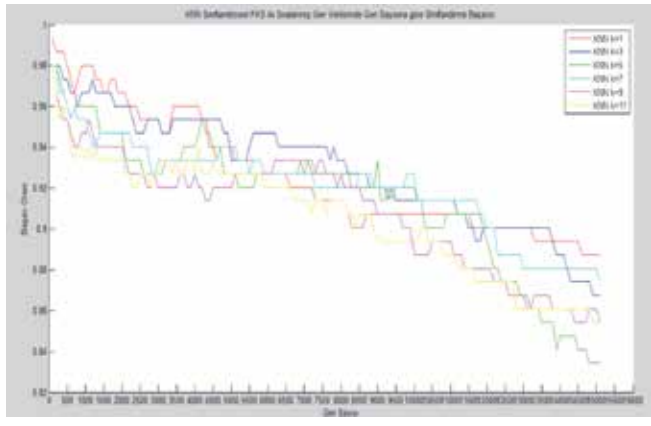
WTS sıralı indirgenmiş gen verisi	kNN (k=1)	kNN (k=3)	kNN (k=5)	kNN (k=7)	kNN (k=9)	kNN (k=11)
İlk 100 gen	98,01	98,68	98,01	98,01	98,01	100
İlk 200 gen	98,68	97,35	98,01	97,35	96,03	97,35
İlk 300 gen	99,34	98,01	98,01	96,69	96,03	96,03
İlk 400 gen	98,01	97,35	97,35	96,69	96,03	94,7
İlk 500 gen	98,01	96,69	96,69	96,03	95,36	94,7
İlk 600 gen	98,01	96,69	96,69	95,36	96,03	95,36
İlk 700 gen	98,01	96,69	96,03	96,03	96,03	95,36
İlk 800 gen	98,01	96,69	96,03	95,36	96,03	94,7
İlk 900 gen	98,68	96,03	96,03	95,36	96,03	95,36
İlk 1000 gen	98,68	96,03	96,69	95,36	95,36	94,7
Tüm veri	88,74	86,75	83,44	87,42	85,43	85,43

Tablo 4'te Weltch t testi ile sıralanmış genlerden ilk 100 gen için %99,34 başarı elde edilmiştir. Gen seçilmeksizin tüm veri kümesine SVM (çf=lineer için) sınıflandırıcısı uygulandığında %98,68 bulunurken, ilk 200 gen için %100 başarı elde edilmiştir. SVM' nin farklı çekirdek fonksiyonları arasında en iyi sınıflandırma yapan doğrusal olurken, en kötü ise radyal çekirdek fonksiyon olmuştur. SVM çekirdek fonksiyonu karesel (quadratic) seçildiğinde en iyi başarı ilk 100 gen için bulunmuştur. En iyi sınıflandırma başarısı önerilen metod ile ilk sıralanmış 1000 örneğin SVM ile sınıflandırılması radyal hariç diğer tüm çekirdek fonksiyonlarının hepsinde de tüm veri kümesinden elde edilen başarıdan daha fazla başarı elde edilmiştir.

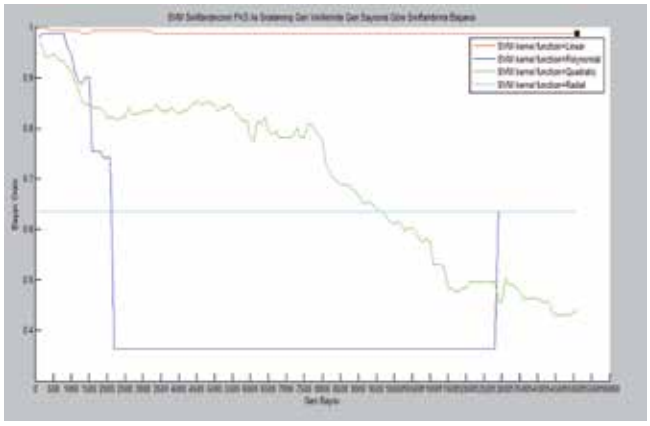
Tablo 4. WTS sıralı indirgenmiş gen verisinin SVM ile sınıflandırma sonuçları

WTS sıralı indirgenmiş gen verisi	SVM (ÇF =Lineer)	SVM (ÇF =Polinom)	SVM (ÇF =Quadratik)	SVM (ÇF =Radyal)
İlk 100 gen	99,34	98,01	97,35	63,58
İlk 200 gen	100	98,68	96,03	63,58
İlk 300 gen	100	99,34	94,04	63,58
İlk 400 gen	100	99,34	9073	63,58
İlk 500 gen	100	99,34	92,05	63,58
İlk 600 gen	100	99,34	93,38	63,58
İlk 700 gen	100	98,01	92,72	63,58
İlk 800 gen	100	98,01	92,05	63,58
İlk 900 gen	99,34	95,36	88,74	63,58
İlk 1000 gen	99,34	94,7	86,09	63,58
Tüm veri	98,68	63,58	43,71	63,58

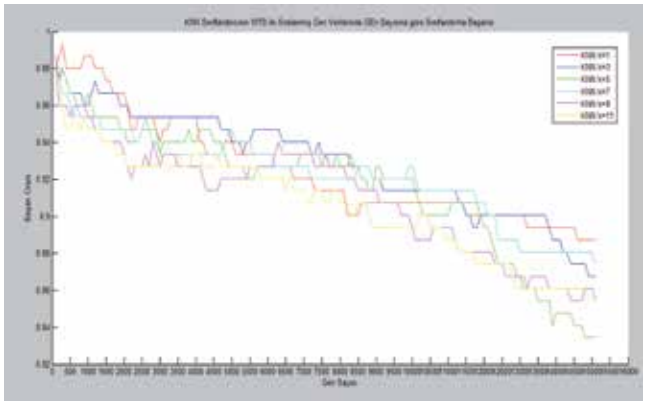
Şekil 2, Şekil 3, Şekil 4, Şekil 5, Tablo 1, Tablo 2, Tablo 3 ve Tablo 4'teki bilgileri içermekte olup ilk 100 örnekten 100 artırımlı olarak tüm veri kümesine kadar seçilen özniteliklerin sınıflandırma başarılarını göstermektedir. Şekil 2 ve Şekil 4 sıralanmış özniteliklerin kNN sınıflandırıcısındaki başarılarını, Şekil 3 ve Şekil 5 ise sıralanmış özniteliklerin SVM sınıflandırıcısındaki başarılarını göstermektedir.



Şekil 2. FKS sıralı indirgenmiş gen verisinin kNN ile sınıflandırma sonuçları.



Şekil 3. FKS sıralı indirgenmiş gen verisinin SVM ile sınıflandırma sonuçları.



Şekil 4. WTS sıralı indirgenmiş gen verisinin kNN ile sınıflandırma sonuçları.

Şekil 5. WTS sıralı indirgenmiş gen verisinin SVM ile sınıflandırma sonuçları.

Bu çalışmada elde edilen performans sonuçları Tablo 5'de literatürdeki benzer çalışmalarla karşılaştırılmıştır. Bu tabloda literatürdeki birçok çalışmadan buradaki çalışmaya benzer olan sınırlı sayıda örnek alınmıştır. Görüldüğü gibi özellikle kullanılan sınıflandırıcılar SVM ve kNN üzerinde yoğunlaşmıştır. Farklılık esas itibarıyla öznitelik seçme algoritmalarındadır. Verilen rakamlar ilgili çalışmalardaki en iyi sonuçları veren durumlardır. Tablo 5'ten de görüldüğü gibi bu çalışmada literatürdeki yüksek başarı oranlarına ulaşılmıştır. Öte yandan bu çalışmada kullanılan öznitelik seçme algoritmaları oldukça kolay uygulanabilen etkili yaklaşımlardır.

Tablo 5. Literatürdeki benzer çalışmalarla karşılaştırma sonuçları

Çalışma	Öznitelik Seçme	Öznitelik Sayısı	Sınıflandırma	Başarı Oranı (%)
Liu (2002) [28]	Korelasyon tabanlı	17	SVM	100
Liu (2002) [28]	Korelasyon tabanlı	17	kNN	100
Liu (2002) [28]	Chi-kare	6136	SVM	100
Kalousis (2007) [29]	SVMRFE [30]	40	SVM	99,56
Saeys (2008) [31]	Simetrik Belirsizlik, Relief, SVMRFE	?	SVM	100
Bu çalışma	WTS	100	kNN	100
Bu çalışma	FKS, WTS	100	SVM	100

VI. SONUÇ VE ÖNERİLER

Bu çalışmada, biyoinformatikte çok boyutlu veri kümelerindeki özellik kanser veri kümelerindeki binlerce öznitelik bilgisinin Fisher korelasyon ve Welcth t testi ile boyut indirgeyerek diğer bir deyişle ilgili öznitelikleri seçerek sınıflandırma başarıları değerlendirilmiştir. Mikrodizilim çiplerinden gelen binlerce gene ait veriler biyologlar tarafından gözle değerlendirilip genlerin hastalık ile alakalılığı hakkında bir kanıya varılmaktadır. Çalışma mikrodizilim çip verilerinden alakalı olan daha az sayıdaki gen üzerinde ilgili çalışmaların yapılabileceğini göstermektedir. Burada özellikle yumurtalık kanseri verileri üzerinde odaklanılmıştır. Literatürde benzer çalışmalar bulunmakla birlikte, bu çalışmada Fisher korelasyon ve

Welcth t testi ile öznitelik seçerken farklı boyutlar analiz edilmiş ve kNN ve SVM sınıflandırma yöntemlerinin farklı parametreleri detaylı bir şekilde incelenmiştir. Bunun sonucunda en iyi boyut ve parametreler tespit edilmiştir.

Tablo 1 ve Tablo 3'teki FKS ve WTS ile seçilmiş genlerin farklı k değerlerinde kNN sınıflandırıcısındaki başarıları karşılaştırıldığında WTS ile sıralanan ilk 100 gen ve k=11 değerinde kNN %100 başarı ile FKS den daha yüksek başarı elde edilmiştir.

Tablo 2 ve Tablo 4'teki FKS ve WTS ile seçilmiş genlerin farklı çekirdek fonksiyonlarında SVM sınıflandırıcısındaki başarıları karşılaştırıldığında iki gen seçimi ve sınıflandırma sonrasında da % 100 başarı elde edilmiştir.

Tablo 1 ve Tablo 2'de FKS ile seçilmiş genler kNN ve SVM sınıflandırıcı ile sınıflandırıldığında SVM (lineer) ile sınıflandırma kNN (k' nın tüm değerlerinde) ile yapılan sınıflandırmadan daha yüksek başarı olan %100 başarı göstermiştir.

Tablo 3 ve Tablo 4'te WTS ile seçilmiş genler kNN ve SVM sınıflandırıcı ile sınıflandırıldığında iki sınıflandırıcıda da %100 başarı yakalanmasına karşın, SVM (lineer) sınıflandırıcının ilk 200, 300, 400 gen de %100 başarısını koruyarak daha kararlı bir yapı gösterdiği görülmektedir.

Çalışmada seçilen ilk 100 gen bilgisi ile yapılan sınıflandırma sonuçları SVM ve kNN sınıflandırıcıda genelde en iyi sonucu vermiştir. Sonuç olarak, yumurtalık kanseri veri kümesi için binlerce gen yerine belirlenen 100 gende inceleme yapılması yeterli olabilecektir.

KAYNAKÇA

[1] Bioinformatics, http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html (Aralık 2015).

[2] Feng, D. F., Johnson, M. S., & Doolittle, R. F. (1985). Aligning amino acid sequences: comparison of commonly used methods. Journal of Molecular Evolution, 21(2), 112-125.

[3] Feng, D. F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. Journal of molecular evolution, 25(4), 351-360.

[4] Lathrop, R. H., Webster, T. A., & Smith, T. F. (1987). Ariadne: pattern-directed inference and hierarchical abstraction in protein structure recognition. Communications of the ACM, 30(11), 909-921.

[5] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

[6] Bowie, J. U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253(5016), 164-170.

[7] Witten, I.H., Frank, E. (2005). Data Mining, Practical Machine Learning Tools and Techniques. 2. baskı, Elsevier Press.

[8] Hand, D., Mannila, H., Smyth, P. (2001). Principles of Data Mining. MIT Press.

[9] Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. Information & Management, 37(5), 271-281.

[10] Hornick, F.M., Marcadé, E., Venkayala, S. (2007). Java Data Mining: Strategy, Standard and Practice a Practical Guide for Architecture, Design and Implementation. Morgan Kaufman.

[11] Larose, D.T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. Wiley Publishing.

[12] Berger, A. M. D. (2005). Identification of factors associated with postoperative pneumonia using a data mining approach. Boston College Dissertations and Theses, AAI3161705.

[13] Edelstein, H.A. (1999). Introduction to Data Mining and Knowledge Discovery. 3. baskı, Two Crows Corporation.

[14] Olson, D.L., Delen, D. (2008). Advanced data mining techniques. Springer.

[15] Chen, K. H., Wang, K. J., Wang, K. M., & Angelia, M. A. (2014). Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. Applied Soft Computing, 24, 773-780.

[16] Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Hidden Markov models for cancer classification using gene expression profiles. Information Sciences, 316, 293-307.

[17] Cao, J., Zhang, L., Wang, B., Li, F., & Yang, J. (2015). A fast gene selection method for multi-cancer classification using multiple support vector data description. Journal of biomedical informatics, 53, 381-389.

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

[19] Lotfi, E., & Keshavarz, A. (2014). Gene expression microarray classification using PCA–BEL. Computers in biology and medicine, 54, 180-187.

[20] Vanitha, C. D. A., Devaraj, D., & Venkatesulu, M. (2015). Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science, 47, 13-21.

[21] Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). A novel aggregate gene selection method for microarray data classification. Pattern Recognition Letters, 60, 16-23.

[22] Du, D., Li, K., Li, X., & Fei, M. (2014). A novel forward gene selection algorithm for microarray data. Neurocomputing, 133, 446-458.

[23] Xiang, J., Han, X., Duan, F., Qiang, Y., Xiong, X., Lan, Y., & Chai, H. (2015). A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method. Applied Soft Computing, 31, 293-307.

[24] Chandra, B., & Babu, K. N. (2014). Classification of gene expression data using spiking wavelet radial basis neural network. Expert systems with applications, 41(4), 1326-1330.

[25] Latkowski, T., & Osowski, S. (2015). Computerized system for recognition of autism on the basis of gene expression microarray data. Computers in biology and medicine, 56, 82-88.

[26] Ovarian Cancer (NCI PBSII Data), http://datam.i2r.a-star.edu.sg/datasets/krbd/OvarianCancer/OvarianCancer-NCI-PBSII.html (Aralık 2015)

[27] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V.A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C. and Liotta, L.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. The Lancet,359, 572–77.

[28] Liu, H., Li, J., & Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome informatics, 13, 51-60.

[29] Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and information systems, 12(1), 95-116

[30] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), 389-422.

[31] Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Machine learning and knowledge discovery in databases (pp. 313-325). Springer Berlin Heidelberg.