T.C.
SÜLEYMAN DEMİREL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİYOİNFORMATİK MİKRO DİZİ VERİLERİ ÜZERİNDE GEN SEÇİMİ VE SINIFLANDIRMA İÇİN YENİ BİR SEZGİSEL YAKLAŞIM GELİŞTİRİLMESİ

Mehmet BİLEN

Danışman
Prof. Dr. Tuncay YİĞİT

II. Danışman
Doç. Dr. Ali HAKAN IŞIK

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
ISPARTA - 2020

TAAHHÜTNAME

Bu tezin akademik ve etik kurallara uygun olarak yazıldığını ve kullanılan tüm literatür bilgilerinin referans gösterilerek tezde yer aldığını beyan ederim.

Mehmet BİLEN

İÇİNDEKİLER

Sayfa

ÖZET

Doktora Tezi

BİYOİNFORMATİK MİKRO DİZİ VERİLERİ ÜZERİNDE GEN SEÇİMİ ve SINIFLANDIRMA İÇİN YENİ BİR SEZGİSEL YAKLAŞIM GELİŞTİRİLMESİ

Mehmet BİLEN

Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Tuncay YİĞİT

II. Danışman: Doç. Dr. Ali Hakan IŞIK

Lösemi, diğer kanser türlerinde olduğu gibi dünya çapında birçok insanın sağlığını tehdit eden ölümcül bir hastalıktır. Lösemi hastalığına ait gen-kanser ve gen-gen ilişkilerinin ortaya çıkarılabilmesi için mikro dizi verileri aynı anda binlerce genin ekspresyon değerini ölçebilmesinden dolayı yaygın olarak kullanılmaktadır. Bununla beraber mikro dizi verilerinin yüksek veri boyutu ve yetersiz örnek sayısı içermesi analiz çalışmalarını zorlaştırmaktadır. Bu çalışmada mikro dizi verilerinde filtreleme, gen seçimi, sınıflandırma ve diğer analizlerin yapılabilmesi için web tabanlı bir arayüz geliştirilmiştir. Ayrıca lösemi verilerinin içerisindeki anlamlı genlerin filtrelenmesi, veri boyutunun indirgenmesi, gen seçimi işlemlerinin gerçekleştirilmesi ve bu veri kümesine ait örneklerin başarılı bir şekilde sınıflandırılabilmesi için yeni bir hibrit gen seçim algoritması tasarlanmıştır. Geliştirilen algoritma filtreleme ve gen seçimi olmak üzere iki adımdan oluşmaktadır. İlk adımda Fisher korelasyon skoru, Willcoxon rütbeler toplamı ve Bilgi kazanımı algoritmalarından oluşan birlik bir gen seçim algoritması tasarlanarak gen filtreleme işlemi gerçekleştirilmektedir. İkinci adımda ise güçlendirilmiş bir genetik algoritma kullanılarak filtrelenmiş genlerin içerisinden en başarılı genler seçilmektedir. Lösemi veri kümesi üzerinde geliştirilen algoritmanın seçmiş olduğu genler ile gerçekleştirilen sınıflandırma işlemleri sonucunda sadece iki gen ile %100 test başarısı elde edilmiştir. Elde edilen performans değerleri literatürdeki çalışmalar ile karşılaştırıldığında geliştirilen algoritmanın %100 LOOCV ve %100 K10 çapraz doğrulama değerine en az sayıda gen seçimi ile ulaşarak daha başarılı olduğu görülmektedir. Geliştirilen algoritmanın seçmiş olduğu genlerin ortaya çıkarmış olduğu biyolojik bulgular incelendiğinde ise Lösemi hastalığının teşhis, tedavi ve ilaç geliştirime aşamalarında kullanılan onkogenleri genleri başarılı bir şekilde belirlediği görülmektedir.

Tez kapsamında geliştirilen web tabanlı arayüz ile dünyanın her yerinden araştırmacıların sadece internet bağlantısına sahip bir cihaz ile kullanabileceği, birçok farklı algoritma ve yaklaşımı farklı birleşimler ile kendi verilerine uygulayabileceği Yapay Zeka çalışma ortamı oluşturulmuştur. Ayrıca literatüre

başarılı bir Birlik (Ensemble) - Hibrit gen seçim algoritması kazandırılmıştır. Geliştirilen algoritmanın yüksek veri boyutuna ve yetersiz örneğe sahip diğer veri kümeleri üzerinde başarılı sonuçlar vereceği düşünülmektedir.

Anahtar Kelimeler: Lösemi, web tabanlı biyoinformatik aracı, birlik algoritma, gen filtreleme, gen seçimi

2020, 123 sayfa

ABSTRACT

Ph.D. Thesis

DEVELOPING A NEW HEURISTIC APPROACH FOR GENE SELECTION AND CLASSIFICATION ON BIOINFORMATIC MICRO ARRAYS

Mehmet BİLEN

Süleyman Demirel University
Graduate School of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Prof. Dr. Tuncay YİĞİT

Co-Supervisor: Assoc. Prof. Dr. Ali Hakan IŞIK

Leukaemia, as in other cancer types, is a deadly disease that threatens the health of many people worldwide. Micro array data are extensively used due to the fact that it can calculate the expression value of thousands of genes simultaneously in order to reveal the gene-cancer and gene-gene relationships related to Leukaemia. Additionally, the fact that micro array data includes a huge data size and insufficient number of samples makes the analysis studies more difficult. A web-based interface was developed in this study to carry out filtering, gene selection, classification and other analyses on micro array data. Moreover, a new hybrid gene selection algorithm was developed in order to filter significant genes within the leukaemia data, to reduce the data size, to perform gene selection processes, and to successfully classify the samples from this data set. The developed algorithm is made up of two steps; filtering and gene selection. In the first step, gene filtering process is carried out by designing an ensemble gene selection algorithm that is made up of Fisher Correlation Score, Wilcoxon Rank Sum, and Information Gain algorithms. In the second step, most successful genes were chosen among the filtered genes by using a reinforced genetic algorithm. 100% success rate was obtained only from two genes as a result of the classification made through the genes chosen by the developed algorithm from the Leukaemia data set. Upon the comparison of the obtained performance values with the ones from the studies in the literature, it is seen that the developed algorithm is more successful by reaching 100% LOOCV and 100% K10 cross validation value with the least number gene selection. When the biological findings that the genes selected by the developed algorithm are studied, it is seen that it also can successfully identify the oncogenes used in the diagnosis, treatment, and medicine development stages of Leukaemia.

Thanks to the web-based interface that was developed within the scope of this thesis, an Artificial Intelligence environment was created where researchers from all over the world can access by a device with internet connection and which can help them practice various algorithms and approaches in different

combinations for their data. Furthermore, a new and successful Ensemble-Hybrid gene selection algorithm was brought into the literature. It is thought that the developed algorithm can yield more successful results for other datasets with huge data size and insufficient samples.

Keywords: Leukemia, web-based bioinformatics tool, ensemble algorithm, gene filtering, gene selection

2020, 123 pages

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Prof. Dr. Tuncay YİĞİT'e teşekkürlerimi sunarım.

Tez çalışmamın başından sonuna her adımında bana rehberlik eden değerli hocam ve ikinci danışmanım Doç. Dr. Ali Hakan IŞIK'a ayrıca teşekkür ederim.

Tez izleme komitesinde bulunan ve tezin gelişimine katkı sağlayan hocalarım Dr. Öğr. Üyesi Asım Sinan YÜKSEL ve Dr. Öğr. Üyesi Ufuk ÖZKAYA'ya teşekkürlerimi sunarım.

Tezimin hiç bir aşamasında beni yalnız bırakmayan aileme ve arkadaşlarıma sonsuz sevgi ve saygılarımı sunarım.

Mehmet BİLEN
ISPARTA, 2020

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

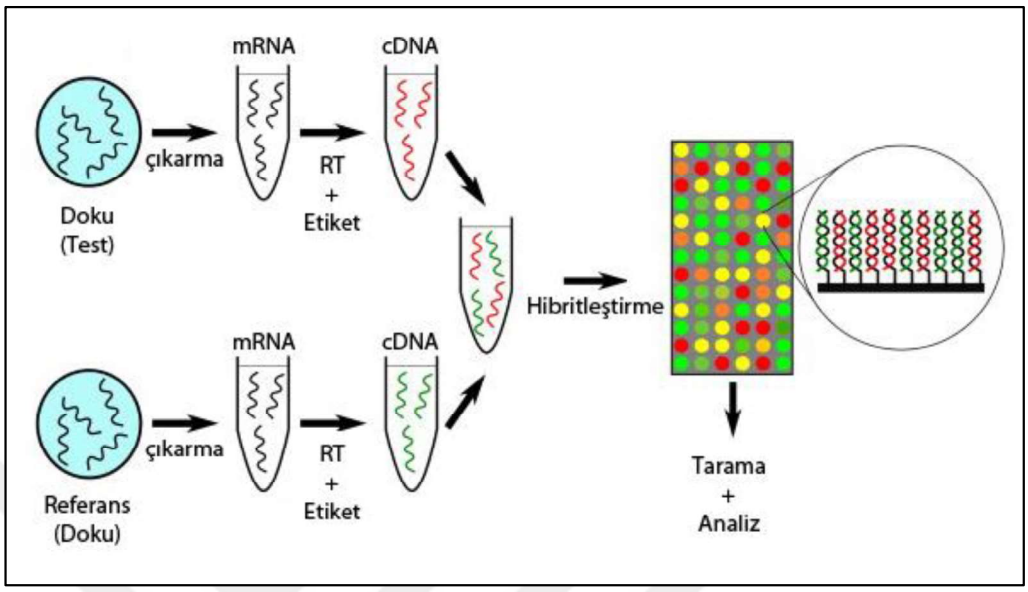
SİMGELER VE KISALTMALAR DİZİNİ

ALL	Akut Lenfoblastik Lösemi
AML	Akut Miyeloid Lösemi
AUC	Eğri Altında Kalan Alan (Area Under Curve)
BK	Bilgi Kazanımı
DNA	Deoksirübo Nükleik Asit
DVM	Destek Vektör Makinesi
FKS	Fisher Korelasyon Skoru
GA	Genetik Algoritma
K 10	K parametresi 10 seçilerek gerçekleştirilen çapraz doğrulama işlemi
K 5	K parametresi 5 seçilerek gerçekleştirilen çapraz doğrulama işlemi
KKO	Karınca Kolonisi Optimizasyonu
KNN	K En Yakın Komşular
LOOCV	Bir tanesi dışarıda bırakarak gerçekleştirilen çapraz doğrulama işlemi (Leave One Out Cross Validation)
MSST	Merkezi Sinir Sistemi Tümörü
NB	Naif Bayes
OKH	Ortalama Kare Hata
PSO	Parçacık Sürü Optimizasyonu
TBA	Temel Bileşen Analizi
WRT	Wilcoxon Rütbeler Toplamı
YSA	Yapay Sinir Ağları

1. GİRİŞ

Kanser insan hayatını ve yaşam kalitesini ciddi ölçüde tehdit eden ölümcül bir hastalıktır. Anormal hücrelerin kontrolsüz olarak bölünerek çoğalmasına ve diğer dokuları istila etmesine neden olan bu hastalığa, her geçen gün daha fazla insan yakalanmakta ve bundan dolayı acı çekmektedir. Dünya genelinde 2018 yılında 18 milyondan fazla yeni kanser vakası ortaya çıkarken, 9 milyondan fazla kanser hastalığına bağlı ölüm gerçekleşmiştir (GLOBOCAN ve WHO, 2019a). Türkiye İstatistik Kurumu'nun verilerine bakıldığında ise %19,7 oranı ile kanser hastalığının, Türkiye'deki ölüm nedenlerinin içerisinde ikinci sırada yer aldığı görülmektedir (TÜİK, 2019). Bir kanser türü olan Lösemi ise kemik iliğindeki kök hücrelerin hızlı bir şekilde çoğalması sonucu ortaya çıkan aynı derecede ciddi bir hastalık olarak tanımlanmaktadır (Davis vd., 2014). Lösemi hastalığı 2018 yılında dünya genelinde 309.006 ölüme neden olmuştur (GLOBOCAN ve WHO, 2019b). Bu rakam kanser kaynaklı ölümler sıralamasında lösemi hastalığının ilk onda yer almasına neden olmaktadır. Ülkemizde ise bu hastalık özellikle çocukluk çağında görülen kanserler içerisinde %32,3 oran ile birinci sırada yer almaktadır (T.C. Sağlık Bakanlığı, 2019). Bu veriler dikkate alındığında, kanser ile mücadelenin, insan hayatının ve yaşam kalitesinin güvence altına alınması adına önemli bir çalışma alanı olduğu anlaşılmaktadır. Bu nedenle, löseminin genetik nedenlerinin ortaya çıkarılması ve hızlı bir şekilde teşhis edilebilmesi için gerçekleştirilen çalışmalar da her geçen gün daha fazla önem kazanmaktadır. Bu kapsamda, binlerce gen ekspresyon değerinin aynı anda ölçülerek analiz edilmesine olanak veren mikro dizi teknolojisi, gen-kanser ve gen-gen ilişkilerinin ortaya çıkarılmasında büyük bir rol oynamaktadır. Mikro dizi yaşayan canlıların genetik yapısının, gen ve protein fonksiyonlarının anlaşılabilmesi için DNA (Deoksirübo Nükleik Asit) seviyesinde biyolojik sinyaller üretilebilmesini ve bu sinyallerin biyolojik yongalar (biyoçipler) üzerinde incelenebilmesini sağlayan bir teknolojidir (Barbulovic-Nad vd., 2006). Binlerce gen ekspresyon bilgisinin tek seferde incelenebilmesi teşhis, tedavi, ilaç üretimi vb. çalışma alanlarına büyük katkılar sağlamaktadır. Bu nedenle kanser araştırmalarında yaygın olarak

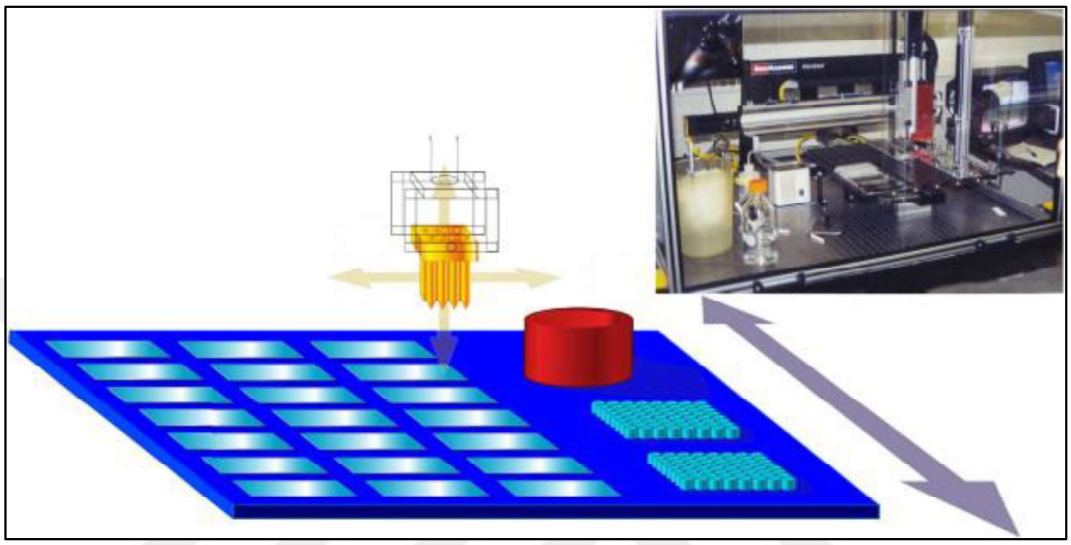
kullanılmaktadır. Mikro dizi verisi elde etmek için izlenen yol en basit hali ile Şekil 1.1'de verilmiştir.



Şekil 1.1. Mikro Dizi elde etme adımları (Ewart vd., 2005)

Genlerin mikro dizi yöntemi ile elde edilmiş gen ekspresyon değerlerine bakıldığında, sağlıklı-sağlıksız veya farklı türdeki kanser tipine sahip hücrelerin genleri birbirinden farklı değerlere sahip olmaktadır. Bu ölçümün yapılabilmesi için ölçülmek istenen dokudan önce mRNA'lar (Mesajcı Ribo nükleik asit) elde edilmektedir. Bu mRNA'lar ters enzim transkripsiyon (Reverse Enzyme Transcription) işleminden geçirilerek cDNA (Tamamlayıcı DNA) elde edilir. Elde edilen her bir örnek bir plaka üzerine yerleştirilir. Daha önceden sağlıklı (test) ve sağlıksız (referans) (ya da bir kanserin farklı türlerinden) olduğu bilinen dokulardan alınmış örnekler ayrı ayrı cDNA'ları elde edildikten sonra farklı renkli floresan boyaları (örnek; sağlıklı-yeşil, sağlıksız-kırmızı) ile etiketlenerek plaka üzerine yerleştirilmiş örneklerle karıştırılır ve etkileşime girmesine izin verilir. Kırmızı işaretlenmiş cDNA'lar sağlıksız örneklere, yeşil işaretlenmiş cDNA'lar da daha çok sağlıklı örneklere bağlanması beklenir ancak birçok örnek tam olarak sağlıklı ya da tam olarak sağlıksız değildir. Bu yüzden spot üzerindeki örnekler farklı oranlarda yeşil ve kırmızı cDNA'larla bağlanarak Kırmızı ile yeşil renk skalası arasında farklı renk yoğunluklarına sahip olurlar. Plaka üzerindeki her bir nokta farklı bir geni temsil etmesi nedeniyle her bir

noktanın renk yoğunluğunun sayısal değeri bir dizi logaritmik fonksiyon ve normalleştirme önişleminden geçirildikten sonra ilgili genin ekspresyon değeri olarak kabul edilir. Şekil 1.2'de bu işlemlerin gerçekleştirildiği düzenek ve bir mikro dizi yazıcı gösterilmiştir.



Şekil 1.2. Bir mikro dizi yazıcı ve gerekli düzenek (Karakach vd., 2010)

Büyük ve yüksek boyutlu biyolojik verilerin mikro dizi gibi teknolojilerle elde edilmesi bu verilerin analizinin yapılabilmesi için yeni bir çalışma alanı gerekliliği doğurmuştur. Hesper ve Hogeweg (1970) paylaştıkları "Biyoinformatik; çalışan bir kavram" adlı bilimsel çalışmada Biyoinformatik kavramını ilk defa ortaya atarak biyolojik verilerin bilgisayar teknolojisinin yardımı ve bilgisayar biliminde kullanılan yaklaşımlar ile analiz edilmesinden bahsetmişlerdir. Gelişen teknoloji ile birlikte her geçen gün elde edilen biyolojik verilerin miktarının ve karmaşıklığının daha da artması bu çalışma alanını daha da önemli hale getirmiştir. Mikro dizi ile elde edilen veriler de biyoinformatik çalışma alanı içerisinde, birçok farklı istatistiksel yöntem kullanılarak analiz edilmektedir. Bunlar ilgileşim (korelasyon) analizi, kümeleme analizi, temel öz bileşen analizi ve regresyon modelleri olarak özetlenebilmektedir (Wang vd., 2018) .

Biyoinformatik alanında mikro dizi verilerinin analiz edilmesinde esnasında bazı güçlüklerle karşılaşılmaktadır. Tek seferde bir dokuya ait binlerce genin ekspresyon değerinin elde edilmesi ve bu değerlerin aynı anda incelenebilmesi kanser araştırmalarına büyük katkı sağlasa da mikro dizi veri elde etme işlem adımlarından da tahmin edilebileceği üzere işlem maliyeti oldukça yüksektir. Bu verilerin elde edilmesinde ki maliyet nedeni ile her ne kadar örnek içerisindeki gen ekspresyon boyutu yüksek olsa da, toplam örnek sayısı çok az olabilmektedir. Bu durum mikro dizi verilerinin genelinde yüksek boyutluluğun laneti (curse of dimensionality) adı verilen bir sorunu ortaya çıkarmaktadır. Örnek sayısının az olduğu durumlarda istatistiksel yaklaşımlar hızlı çalışmalarına rağmen bu probleme sahip verilerin analizinde başarısız olmaktadırlar. Bu durumda daha az örnek sayısı ile başarılı sonuçlar elde edebilen yapay zeka temelli sezgisel algoritmalara başvurulsa da bu kez de gen sayısının fazla olması, verinin algoritmaya uyumsuz olma problemine neden olmaktadır. Bir diğer güçlük ise bir veri için çok güzel sonuçlar üreten istatistiksel veya geleneksel bir yaklaşımın başka bir veri için başarısız sonuçlar üretebilmesidir.

Dünyada ve ülkemizde var olan lösemi gerçeği ve bunun meydana getirdiği sonuçlar dikkate alındığında; Lösemi ve benzer kanser türleri ile mücadelede edilebilmesi için bahsedilen güçlüklerin üstesinden gelebilen bir biyoinformatik aracına duyulan ihtiyaç göze çarpmaktadır. Bu doğrultudan yola çıkarak bu çalışmanın amacı, içerisinde birçok farklı biyoinformatik analizinin gerçekleştirilebileceği web tabanlı bir biyoinformatik aracı geliştirilmesi olarak belirlenmiştir. Ayrıca yeni bir Birlik-Hibrit algoritma geliştirilmesi ile lösemi hastalığına ait verilerin sınıflandırılabilmesi ve gen-kanser, gen-gen ilişkilerinin ortaya çıkartılabilmesi amaçlanmıştır. Bu amaçla, literatürde sıklıkla kullanılan gen filtreleme algoritmaları Fisher korelasyon skoru (FKS, Fisher Correlation Scroe), Bilgi Kazanımı (BK, Information Gain) ve Wilcoxon Rütbeler Toplamı (WRT, Wilcoxon Runk Sum) çalışmanın ilk adımında birlik hale getirilerek ortak bir gen filtreleme işlemi gerçekleştirilmiştir. Filtreleme işlemi sonucunda elde edilen genler sınıflandırma performansına göre bir eşik değerden geçirilerek birleştirilmiştir. İstatistiksel olarak anlamlı olan genlerin belirlenmesi ile

karmaşıklığı daha düşük bir alt küme elde edilerek ikinci adıma aktarılmıştır. İkinci adımda lösemi kanseri ile ilişkili en anlamlı genlerin seçimi, Genetik algoritma (GA) kullanarak sınıflandırma başarısının iteratif bir şekilde artırılmasıyla gerçekleştirilmiştir. İlk adımda seçilen genlerin performansı K en yakın komşular algoritması (KNN, K nearest neighbors) ile ölçülürken ikinci adımda ki genetik algoritmanın uygunluk değeri KNN, Destek vektör makinesi (DVM, Support Vector Machine) ve Naif Bayes (NB) algoritmalarından oluşan birlik bir yaklaşımla belirlenmiştir.

Bu tez çalışması 5 ana başlık altında sunulmuştur. Giriş başlığı altında bu çalışmanın motivasyonunu oluşturan dünyada ve ülkemizdeki kanser vakaları, kanser ile mücadele için kullanılan mikro dizi teknolojisi ve biyoinformatik çalışma alanından bahsedilmiştir. Kaynak özetleri bölümünde literatürde gerçekleştirilen mikro dizi çalışmaları, tez içerisinde geliştirilen algoritmanın literatürdeki yeri ve son olarak lösemi verisi üzerinde yapılan çalışmaların özeti verilmiştir. Materyal ve Yöntem kısmında üzerinde çalışılan veri kümesi, tez içerisinde kullanılan filtreleme, ön işlem, gen seçim ve sınıflandırma algoritmalarının teorik temellerinden bahsedilmiştir. Araştırma Bulguları ve Tartışma bölümünde geliştirilen Web tabanlı biyoinformatik aracının ve Birlik-Hibrit Algoritmanın detayları verilmiştir. Ayrıca bu bölümde geliştirilen algoritma ile gerçekleştirilen deneysel çalışmaların performans değerleri, literatürdeki benzer yaklaşımlarla karşılaştırılması, elde edilen bulguların biyolojik önemi ve tıp literatürüne katkısı incelenmiştir. Sonuçlar ve öneriler kısmında ise bu çalışmadan elde edilen bulgulara, literatüre yapılan katkılara, geliştirilen arayüz ve algoritmanın dezavantajları ile birlikte gelecekte yapılması planlanan çalışmalara yer verilmiştir.

2. KAYNAK ÖZETLERİ

Mikro dizi verileri veri boyutu yüksek ancak örnek sayısı az olan veri kümeleridir. Bu özellikleri ortaya bazı güçlükler çıkarmaktadır. Gen sayısının çok olması bu verilerin sınıflandırılması için kullanacak algoritmaların başarısını büyük ölçüde azaltmaktadır. Gen sayısının azaltılması için literatürde bir çok yöntem önerilse de gen sayısı azaltılırken veri kümesini en iyi temsil eden genlerin seçilmesi başlı başına farklı bir çalışma konusudur. Bu sürecin başarı ile gerçekleştirilmesi, işlem maliyetinin azaltılması, sınıflandırma başarısının artırılması ve kanser ile ilişkisi olan genlerin tespit edilmesi açısından büyük bir önem taşımaktadır. Öznitelik seçimi öznitelikler arasından amaca en uygun alt kümenin belirlenmesi olarak tanımlansada (Dash ve Liu, 1997), bir gen ekpresyon değerinin bir öznitelik olması nedeni ile kavram karmaşısının da önlenmesi amacıyla "öznitelik seçme" işleminden tez çalışmasının geri kalan kısmında "gen seçimi" olarak bahsedilecektir. Başlıca gen seçim yöntemleri filtreleme, sarmal ve hibrit modeller olmak üzere 3 farklı başlık altında toplanabilirken (Dashtban vd., 2018), Gömülü (Embeded) (Sharbaf vd., 2016), çevrim içi (Masoudi-Sobhanzadeh vd., 2019), Birlik (ensemble) (Liu vd., 2010) gibi farklı yaklaşımlarda literatürde karşımıza çıkmaktadır.

Gen seçim yöntemlerinden biri olan filtreleme yaklaşımı, gen alt kümesi belirlenirken genlerin arasındaki korelasyon ve uzaklık gibi istatistiksel verilere bağımlı kalmaktadır (Ruiz vd., 2012). Bu yöntem hızlı bir şekilde istatistiksel olarak anlamsız genleri filtrelese de, sınıflandırma algoritmalarından bağımsız çalışması nedeniyle sınıflandırma başarısı diğer yaklaşımlara göre daha azdır. Hızlı çalışmaları nedeni ile bir çok farklı yaklaşımın ön işlem adımında kullanılmaktadırlar. Bu sayede sezgisel algoritmaların hesaplama maliyetinin düşürülmesi ayrıca verinin algoritmaya uygun olmama probleminin ortadan kaldırılması sağlanabilmektedir.

Sarmal yaklaşımların istatistiksel olarak anlamsız genleri filtrelemek yerine amacı sınıflandırma başarısı yüksek olan alt kümeleri bulmaya çalışmaktadır.

Doğrudan sınıflandırma başarısına odaklandığı için daha yüksek performans sağlayabilmektedir. Ancak veri boyutu yüksek veri kümelerinde potansiyel alt küme sayısının fazla olması ve her bir çözüm kümesinin sınıflandırma başarısının değerlendirilmesi, yüksek bir hesaplama maliyeti gerektirmektedir. Bu yüzden filtreleme yaklaşımı ile kıyaslandığında sınıflandırma performansı yüksek olsa da sarmal yaklaşımlar daha yavaş çalışmaktadır. Genetik algoritma sarmal yaklaşımda gen seçimi için en yaygın kullanılan algoritmalardan biridir (Wang vd., 2011).

Hibrit yaklaşımlar, filtreleme ve sarmal yaklaşımların avantajlarını birleştirerek hem hesaplama maliyetini düşürmekte hem de sınıflandırma başarısının arttırılması için tekrarlı bir yöntem sunmaktadır. Genellikle bu yaklaşımın ilk adımında mikro dizi veri kümesi içerisinden istatistiksel olarak anlamsız genler bir filtreleme yöntemi ile elenmektedir. İkinci adımda ise tekrarlı çalışan bir algoritma ile sınıflandırma başarısı yüksek genler seçilmektedir. FKS ile Yarasa Algoritması (Dashtban vd., 2018), Hücresel otomata (Cellular Automata) ve Karınca kolonisi optimizasyonu (KKO, Ant Colony Optimization) (Sharbaf vd., 2016), BK ve DVM (Gao vd., 2017), Parçacık sürü optimizasyonu (PSO) ve KNN (Kar vd., 2015) gibi literatürde ki bir çok hibrit algoritma buna örnek olarak verilebilir. Buna ek olarak Lee ve Leu (2011) tarafından yapılan çalışmada olduğu gibi ilk önce sarmal yaklaşımın daha sonra filtreleme yaklaşımının kullanıldığı çalışmalar da literatürde karşımıza çıkabilmektedir.

Birlik (Ensemble) algoritmalar ise aynı işlemi yapan birkaç farklı algoritmanın birleştirilerek beraber çalıştırılması ile ortaya çıkmaktadır. Bu yöntemle algoritmaların bireysel olarak elde ettiği performans değerlerinin toplu halde daha iyi hale getirilmesi amaçlanmaktadır (Kim ve Cho, 2008). Bu birleştirme işlemi birkaç farklı şekilde yapılabilmektedir. Birden fazla filtreleme algoritmasının gen filtreleme işlemi yaptıktan sonra sonuçlarının birleştirilerek tek bir alt küme oluşturması şeklinde olabileceği gibi, sarmal yaklaşımların da birlik halinde çalışabileceği birlik yaklaşımlar bulunmaktadır (Bolon-Canedo ve Alonso-Betanzos, 2019). Bu sayede birlik yaklaşımı ile birlikte bir algoritmanın dezavantajlı yönlerinin başka bir algoritma ile kapatılması, birlik halinde daha

başarılı değerler elde edilmesi amaçlanmaktadır. Birlik haline getirilmiş filtreleme ve sarmal yaklaşımlar sırası ile çalıştırılarak, hem birlik hem de hibrit bir çalışma gerçekleştirilebilmektedir (Liu vd., 2009; Azzawi vd., 2016; Mollaee ve Moattar, 2016; Liu vd., 2010; Ghosh vd., 2019).

Gen seçimi için geliştirilen Birlik-Hibrit yaklaşımların literatürdeki yeri incelendikten sonra bu kısmın devamında tez çalışması boyunca yararlanılan ve tezin gelişimine katkı sağlayan benzer yaklaşımlar paylaşılmıştır.

Literatürde lösemi veri kümesi birçok benzer yaklaşımda gen seçim işlemleri için kullanılmıştır. Furey vd., (2000) yaptıkları çalışmada tüm genleri DVM algoritması ile sınıflandırarak bir LOOCV (Bir örneğin dışarıda bırakıldığı çapraz doğrulama, Leave one out cross validation) değeri hesaplamışlardır. Çalışmanın devamında FKS puanı en düşük olan genleri veri kümesi içerisinden çıkartarak en yüksek LOOCV değeri ile en düşük gen sayısına ulaşmayı hedeflemişlerdir. Sonuç olarak çalışma içerisinde %94,10 LOOCV performans değerine seçilen 500 gen ile ulaşmayı başarmışlardır.

DVM ve GA literatürde gen seçimi ve sınıflandırma işlemleri için yaygın olarak kullanılmaktadır. Peng vd. (2003) gen seçiminin gerçekleştirilebilmesi için bu algoritmalardan oluşan bir yaklaşım önermişlerdir. Bu yaklaşımda ilk önce standart sapma puanı bir eşik değerin altında olan gen ekspresyon değerleri elenmekte daha sonra genetik algoritma ile rastgele kromozonlar (gen alt kümeleri) oluşturularak DVM ile sınıflandırılmaktadır. DVM sınıflandırması LOOCV (Bir örneğin dışarıda bırakıldığı çapraz doğrulama, Leave one out cross validation) ile doğrulanarak elde edilen performans değerleri kromozomların uygunluk değeri olarak kabul edilmektedir. GA her bir tekrarda (iterasyon) bu kromozomlar içerisinden bir geni geliştirmiş oldukları mutasyon operatörü ile silerek ve çaprazlama operatörü ile yeni kombinasyonlar elde ederek tekrar uygunluk değeri hesaplamaktadır. Daha başarılı bireyler oluştuğu sürece bu özyinelemeli gen eleme (RSE, recursive feature elimination) işlemine devam etmektedirler. Geliştirdikleri yaklaşım sayesinde Lösemi veri kümesi içerisinden 6 gen seçerek %100 LOOCV değerine ulaşmayı başarmışlardır.

Dettling (2004) yapmış olduğu çalışmada gen seçiminin gerçekleştirilebilmesi için tekrarlı bir algoritma olan BagBoosting'i kullanmıştır. Bu algoritma genlere sınıflandırma sonucuna etki eden birer ağırlık değeri atamaktadır. Her sınıflandırma işlemi sonunda başarılı sınıflandırma işlemine sahip olan genlerin ağırlıklarını artırmakta, başarısız olanların ağırlıklarını ise azaltmaktadır. Yazar önermiş olduğu yaklaşımın tüm veri kümesi yerine WRS ile filtrelenmiş genler ile çalışması durumunda daha iyi sonuçlar elde edildiğini belirtmiştir. Paylaşmış olduğu sonuçlarda 200 gen ile %4.08 test hatasına sahip bir sınıflandırma işlemi gerçekleştirdiği görülmektedir.

Genlerin önem derecesinin puanlanması için Fu ve Fu-Liu (2005) yapmış oldukları çalışmada yeni bir yaklaşım önermişlerdir. Bu yaklaşım veri kümesi içerisinden bir genin çıkarılmasının sınıflandırma sonucuna etkisi göz önünde bulundurularak ilgili genin seçilmesi ya da seçilmemesidir. Sınıflandırma işlemi DVM, gen eleme işlemi için RFE yöntemini kullanarak yapılan işlemler sonucunda tüm genlerin tekrar etme olasılığı hesaplanmakta ve belirli bir eşik değerin üstünde tekrar sayısına sahip olan genler seçilmektedir. Yazarlar bu yöntem ile seçilen 4 geni kullanarak yapmış oldukları sınıflandırma işlemlerinde 97,06 test başarısı elde etmişlerdir.

Genlerin puanlanması için literatürdeki birçok çalışma istatistiksel yaklaşımlardan farklı oranlarda yararlanmaktadır. Yang vd. (2006) bu yaklaşımları çalışmalarının merkezine alarak literatüre GS1 ve GS2 adında iki benzer gen seçim yöntemi önermişlerdir. Önerilen yaklaşımın temelinde bir genin farklı sınıflardaki ekspresyon seviyelerinin arasındaki farkın istatistiksel olarak ölçülmesi bulunmaktadır. Bu ölçüm çalışma içerisinde kullanılan istatistiksel yaklaşımların her biri için yapılmakta ve her gen puan matrisinde bir puana sahip olmaktadır. Yazarlar geliştirmiş oldukları iki farklı istatistiksel hesaplama yöntemi kullanan GS1 ve GS2 yaklaşımları ile en yüksek puana sahip genleri KNN ile sınıflandırmışlardır. Ve önerilen bu yöntem ile 10 ile 100 arasında değişen bir gen sayısı seçerek ortalama 98,60 LOOCV değerini elde etmişlerdir.

Gen sayısının aşırı fazla, örnek sayısının az olduğu durumlar gen seçim işleminin sadece istatistiksel yöntemlerle başarılı sonuçlar elde etmesini güçleştirmektedir. Bu güçlüğün üstesinden gelebilmek için sezgisel algoritmalar literatürde yaygın olarak kullanılmaktadır. Shen vd. (2008) geliştirmiş oldukları yaklaşımda PSO ve Tabu Arama (TS, Tabu Search) sezgisel algoritmalarını birleştirerek gen seçimi işlemleri için kullanmışlardır. Çalışmada PSO algoritması ile en başarılı gen kombinasyonlarını tekrarlamalı bir şekilde seçmeye çalışmışlardır. Ayrıca algoritma içerisindeki parçacıkların arama yöntemi için yeni bir yaklaşım da önermişlerdir. Belli bir tekrarlama sonrasında PSO'nun yerel çözümlere takılmasını (Local Maxima) önlemek ve daha başarılı gen kombinasyonları elde edebilmek için TS algoritması kullanarak PSO algoritmasını optimize etmişlerdir. Sezgisel ve hibrit algoritmaların gen seçimi işlemlerinde başarılı sonuçlar elde edebilmesi adına bu çalışma literatüre iyi bir katkı sağlamıştır. Geliştirdikleri bu yöntem sayesinde yazarlar 7 gen ile %95,81 test başarısı elde etmişlerdir.

Filtreleme ve sarmal gen seçim yaklaşımlarının bir arada kullanılmasına örnek teşkil etmesi bakımından Li vd. (2008)'nin yapmış oldukları çalışma literatürde önemli bir yere sahiptir. Yazarlar çalışmalarının ilk adımında WRS ile bir gen filtreleme işlemi yaparak istatistiksel olarak anlamsız genlerin veri kümesi içerisinden çıkarılmasını sağlamışlardır. Çalışmanın devamında bir GA kullanarak daha derinlemesine bir şekilde gen araştırması gerçekleştirmişlerdir. GA ile seçilen genlerin başarısı ise bir DVM sınıflandırıcı ile hesaplanmıştır. Bilindiği üzere GA sezgisel olarak çalışan ve her çalıştığında farklı sonuçlar üretebilen bir algoritmadır. Bu nedenden dolayı yazarlar yapmış oldukları çalışmada doğrudan gen seçimi yapmak yerine, defalarca çalıştırılan GA'nın en çok seçtiği genleri birleştirerek nihai gen seçim sonucunu elde etmişlerdir. Bu sayede seçmiş oldukları 7 gen ile %100 test başarısı sağlamışlardır.

Yang vd. (2008) yapmış oldukları çalışma literatürdeki hibrit gen seçimine bir başka örnek olarak karşımıza çıkmaktadır. Bu çalışmada, filtreleme adımında BK ve ilgileşim temelli gen filtreleme işlemi, sarmal kısımda ise PSO kullanan iki

adımlı hibrit bir yaklaşım önerilmiştir. Ayrıca PSO algoritmasının 3 tekrarı boyunca, seçilen genlerin daha başarılı sonuçlar vermemesi durumunda, çözüm uzayının genişletilmesi için bir güçlendirme (robust) yöntemi önerilmiştir. Sonuç olarak yazarlar seçilen 203 genin KNN algoritması ile sınıflandırılması ile %100 test başarısı elde etmişlerdir.

Kanser araştırmaları için yapılan gen seçimi işlemlerinin bir amacı da hastalığa etkisi yüksek olan genlerin (Biomarker ya da okogen) tespit edilmesidir. Mikro dizi veri kümeleri birçok gene sahip olsa da bu genlerin büyük bir bölümü ilgili hastalık için anlamsız olabilmektedir. Anlamsız ve anlamlı genlerin ayırt edilebilmesi için Shim vd., (2009) yapmış oldukları çalışmada öncelikle Gaussian temelli bir kümeleme işlemi gerçekleştirmişlerdir. Daha sonra DVM ile sınıflandırma işlemleri gerçekleştirerek her bir gene başarıları oranında bir ağırlık atamışlardır. Bu ağırlıkların her tekrarda güncellenmesi sonucunda en başarılı genlerin seçilmesini amaçlamışlardır. Elde ettikleri sonuçlar incelendiğinde birçok farklı deneme yaptıkları ve ortalama 14,2 gen ile 92,20 çapraz doğrulama test başarısı elde ettikleri görülmektedir.

Yu vd. (2009) yapmış oldukları çalışmada KKO kullanarak gen seçim işleri gerçekleştirmişlerdir. Karıncaların izledikleri yolun uygunluk değerini DVM ile hesaplamışlardır. Bununla beraber KKO algoritmasında değişiklik yaparak karıncaların eş zamanlı olarak çalışmasını sağlayarak farklı yollardaki karıncaların birbirlerinin bilgilerden yararlanmasını sağlamışlardır. Böylelikle geleneksel bir KKO algoritmasına göre daha başarılı sonuçlar elde etmişlerdir. KKO ile ortalama 8,6 gen ile %100 LOOCV başarısı elde ederken değişiklik yaptıkları yeni algoritma ile aynı başarı oranını 6,3 gen ortalaması ile elde etmişlerdir.

Mikro dizi verilerinde en uygun genlerin seçilmesi sınıflandırma başarısı ile büyük oranda ilişkilidir. Ancak mikro dizi veri türlerinin özellikleri dikkate alınarak tasarlanmış bir sınıflandırıcı da bu başarının artırılmasına katkı sağlamaktadır. Zainuddin ve Ong, (2011) tarafından yapılan bu çalışma gen seçiminden daha çok seçilen genlerin başarısının artırılması için, mikro dizi

verilerinin yapısına daha uygun bir sınıflandırıcı tasarlanması üzerine yoğunlaşmıştır. Bu amaçla birden fazla aktivasyon fonksiyonunun kullanıldığı bir Dalgacık Sinir Ağı'nı (WNN, Wavelet Neural Network), Bulanık C-Ortalamalar (FCM, Fuzzy C Means) algoritması ile optimize ederek yeni bir sınıflandırma yaklaşımı önermişlerdir. Literatürde önerilen diğer sezgisel algoritmalarda da olduğu gibi bu çalışmada da ön işlem aracı olarak bir filtreleme algoritması kullanılmıştır. T-test kullanılarak yapılan filtreleme işlemleri ile anlamsız olan genler veri kümesi içerisinden çıkartılmış ve daha az karmaşıklığa sahip yeni veri kümesi önerilen algoritmanın girişine verilmiştir. Sonuç olarak bu yazarlar tarafından önerilen bu yaklaşımla 10 gen ile %100 çapraz doğrulama başarısı elde edilmiştir.

Mohamad vd. (2011) yaptıkları çalışmada güçlendirilmiş bir PSO algoritması önermişlerdir. Algoritmanın içerisindeki parçacıkların birbirleri ile olan etkileşimleri sonucunda ortaya çıkan yer değiştirme hızlarına ve yöntemlerine yeni bir alternatif getirerek geleneksel PSO'ya göre daha yüksek sınıflandırma başarısı elde etmişlerdir. Önerdikleri algoritmanın geleneksel PSO'dan daha hızlı çalışması mikro dizi gibi verilerinin analizinde büyük bir önem taşımaktadır. Parçacıkların uygunluk değeri literatürün büyük çoğunluğunda olduğu gibi DVM ile hesaplanırken filtreleme aşamasında BK filtreleme algoritmasından yararlanılmıştır. Önerilen çalışma ile seçilen 2 gen sayesinde yazarlar %100 çapraz doğrulama başarısı elde edilmiştir.

Gen seçimi ve sınıflandırma işlemleri için Bulanık bir Uzman Sistem yaklaşımı öneren Kumar vd. (2012) yaptıkları çalışmanın optimizasyonu için ise yeni bir Genetik Sürü Optimizasyonu Algoritması (GSO, Genetic Swarm Algorithm) geliştirmişlerdir. Bulanık kuralların optimize edilmesi, popülasyon sayısının artırılması ve sürüdeki parçacıklarının hızının belirlenmesi için yeni kurallar önermişlerdir. Karşılıklı bilgi kazanımı (MI, Mutual Information) algoritmasının ilk adımda gen filtrelemesi için kullanıldığı bu yaklaşım ile 10 gen seçerek yapmış oldukları test işlemlerinde %100 çapraz doğrulama başarısı elde etmişlerdir.

Cui vd. (2013) yapmış oldukları çalışmada literatürden farklı bir yol izleyerek gen seçiminden önce özellik çıkarımı gerçekleştirmişlerdir. Özellik çıkarım işlemi için yeni bir seyrek maksimum marj ayırt edici algoritması (Sparse Maximum Margin Discriminant Analysis) önermişlerdir. Bu algoritma özellik çıkarımı yaparak gen ekspresyon seviyeleri yerine bu yeni özellikler ile sınıflandırma işlemi gerçekleştirmektedir. Sınıflandırma başarısı en yüksek olan özelliklerin çıkartılmasından sonra, bu özelliklerin vektörleri geriye doğru çalıştırılarak bu özelliklere en çok etki eden genler seçilmektedir. Yazarlar bu yöntem ile seçilen 8 gen ile %98,66 test başarısı elde etmişlerdir.

Alshamlan vd. (2015) GA ve Yapay Arı Kolonisi algoritmalarını (ABC, Artificial Bee Colony) beraber kullanarak mikro dizi verileri üzerinde gen seçimi işlemleri için yeni bir hibrit algoritma geliştirmişlerdir. Bu hibrit algoritma ile iki sezgisel algoritmanın sahip olduğu avantajlı işlevleri birleştirilerek daha etkili bir gen seçim işlemi yapılmasını amaçlamışlardır. GA keşif (exploration) aşamasında olası çözüm kümelerini yüzeysel olarak araştırmaktadır ancak detaylı araştırma (exploitation) aşamasında yerel minimumlara takılabilmektedir. ABC ise tam tersine iyi bir derinlemesine araştırma yaparken keşif adımında GA'ya göre daha başarısızdır. Bu doğrultuda GA'nın mutasyon işlevi ve buna ek olarak en başarılı arının genlerini tek taraflı olarak diğer arılara aktardığı yeni bir çaprazlama işlevi ABC algoritmasının içerisine dâhil edilmiştir. Çalışmada uygunluk değeri DVM ile yapılan sınıflandırma sonucu ile belirlenirken anlamsız genlerin veri kümesi dışarısında bırakılması için ilk adımda Minimum Artıklık Maksimum Alaka Düzeyi Özellik Seçim (mRMR, Minimum Redundancy Maximum Relevance) algoritmasından yararlanılmıştır. Sonuç olarak yazarlar geliştirmiş oldukları hibrit yaklaşım ile sadece 4 gen seçerek %100 test başarısı elde etmişlerdir.

Kar vd. (2015) yapmış oldukları çalışmada gen seçimi ve sınıflandırma işlemleri için PSO ve KNN algoritmalarından oluşan bir yaklaşım önermişlerdir. Yaklaşım içerisinde PSO algoritmasındaki parçacıkların uygunluk değeri KNN algoritması ile yapılan sınıflandırma işlemleri ile hesaplanmaktadır. Bu uygunluk değerinin iyileştirilmesi için KNN algoritması içerisindeki K parametresinin değeri her bir

tekrarda seçilen genlere göre yeniden belirlenerek uyarlanabilir (adaptive) bir yapı oluşturulmuştur. Bu önerilen yaklaşımla beraber yazarlar 3 gen ile %97,06 test başarısı ve %95,89 çapraz doğrulama başarısı elde etmişlerdir.

Öğrenme Otomatası (LA, Learning Automata) ve GA'nın beraber kullanıldığı bir yaklaşım öneren Motieghader vd. (2017), bu yaklaşımla literatürdeki en iyi sonuçlardan birine ulaşmışlardır. Yazarlar GA'nın mutasyon, çaprazlama ve seçilim operatörlerini kullanırken LA'nın ceza ve ödül operatörlerinden yararlanarak hibrit bir algoritma oluşturmuşlardır. DVM algoritması benzer çalışmalarda olduğu gibi burada da amaç fonksiyonun hesaplanması için sınıflandırıcı olarak kullanılmıştır. Sonuç olarak yazarlar %100 test başarısına sadece 2 gen seçerek ulaşmayı başarmışlardır.

Salem vd. (2017) yapmış oldukları çalışmada literatürde yaygın olarak kullanılan IG ve GA algoritmaları ile gen seçimi gerçekleştirmişlerdir ve amaç fonksiyonunun hesaplanması için GA temelli bir sınıflandırıcı olan Genetik Programlama (GP, Genetic Programming) algoritmasını kullanmışlardır. GA türevi bir sınıflandırıcı kullanarak sınıflandırma ve gen seçimi yapan algoritmalar arasındaki uyumun artırılmasını amaçlamışlardır. GA'nın çıktılarının bir sınıflandırıcı için tekrar uyarlanmasına gerek kalmadan GP tarafından kullanılabilmesi aynı şekilde GP'nin geri bildirimlerinin GA tarafından doğrudan ele alınabilmesi işlem maliyetini azaltmıştır. Bu yöntem sayesinde yazarlar 3 gen ile 97,06 test başarısı elde etmişlerdir.

PSO algoritması literatürde Jain vd. (2018) tarafından yapılan çalışmada bir kez daha karşımıza çıkmaktadır. PSO algoritmasına bu çalışmada, parçacıkların yeni konumlarının hesaplanmasında yerel çözümlere daha az yakalanması için yeni bir hesaplama önerisi getirilmiştir. İlk adımda gen filtrelenmesi için Korelasyon Temelli Gen filtreleme yaklaşımı tercih edilirken, aynı teorik temellere sahip olması nedeni ile Bayes, amaç fonksiyonunun hesaplanması için sınıflandırıcı olarak kullanılmıştır. Önerilen yaklaşım 4 gen seçimi ile %100 test başarısı elde etmiştir.

Güncel çalışmalar incelendiğinde birden fazla sezgisel algoritmanın hibrit ve memetik (Memetic) yapıda kullanımının daha da yaygınlaştığı ve elde edilen sonuçların her geçen gün iyileştiği görülmektedir. Baliarsingh vd. (2019) yaptıkları çalışmalarında yeni bir memetik yaklaşım önermişlerdir. Yazarlar önerdikleri bu yaklaşım ile Sosyal Mühendislik Optimizasyon Algoritması (SEO, Social Engineering Optimization) ve İmparator Penguen Optimizasyon Algoritmasını (EPO, Emperor Penguin Optimization) bir araya getirerek hem gen seçim seçimi hem de sınıflandırıcı olarak kullanılan DVM algoritmasının optimizasyonunu gerçekleştirilmiştir. Yazarların EPO ve SEO'nun birlikte kullanımı ile algoritmanın yerel çözümlere takılı kalmasını engellemeyi amaçlanmıştır. Bu yaklaşım da EPO algoritması olası çözümleri keşfederken, SEO bu çözümleri derinlemesine araştırmaktadır. Filtreleme adımında FKS ve ReliefF algoritması da kullanılan bu yaklaşım ile 7 gen ile gerçekleştirilen sınıflandırma işlemleri sonucunda 98,82 test başarısı elde edilmiştir.

Ghosh vd. (2019) yapmış oldukları çalışmanın, birlik yaklaşımların gen seçim işlemlerinde kullanılması ve bunun sonucunda elde edilen başarı oranı göz önünde bulundurulduğunda literatürde önemli bir yere sahip olduğu görülmektedir. Çalışmanın ilk adımında tek bir filtreleme algoritması yerine Ki-Kare Testi (Chi-Square), ReliefF, Simetrik Belirsizlik (Symmetrical Uncertainty) algoritmalarından oluşan birleştirme (E_u) ve çıkarma (E_1) adı verilen iki farklı birlik yapısı önerilmiştir. Birleştirme tüm seçilen genlerin ortak bir veri kümesine toplanması, çıkarma ise sadece 3 farklı algoritma tarafından seçilen ortak genlerin seçilmesi prensibine dayanmaktadır. Birleştirme yöntemi, gen filtreleme işleminde farklı istatistiksel bilgilerin göz önünde bulundurulması, çıkarma yöntemi ise daha anlamlı genlerin seçilebilmesini amaçlamaktadır. İki yöntemle elde edilen sonuçların ayrı ayrı değerlendirmesi ile E_u yöntemi sadece 2 gen ile %100 test başarısı elde ederken E_1 yöntemi aynı başarıyı 12 gen ile elde edebilmiştir. Çalışma içerisinde seçilen genler KNN, DVM ve YSA algoritmalarının her biri ile sınıflandırılmıştır ve benzer sonuçlar elde edilmiştir.

Literatürün gelişimi incelendiğinde birçok farklı yaklaşım önerilse de güncel çalışmaların büyük bir çoğunluğunun ilk adımda istatistiksel bir yöntem ile filtreleme, ikinci adımda tekrarlı çalışan bir sezgisel algoritma ile sınıflandırıcıya bağımlı bir gen seçim yöntemi izlediği görülmektedir. Her ne kadar sezgisel algoritmalar ikinci adımda daha yüksek performans sergileseler de filtreleme adımı, sezgisel algoritmaların bu performansına olan katkısı nedeni ile araştırmacılar tarafından tercih sebebi olmuştur. Bununla beraber bazı çalışmalar filtreleme bazı çalışmalar ise gen seçimine yoğunlaşırken, sınıflandırıcının iyileştirilmesi de üzerinde durulan konulardan biri olmuştur. Filtreleme adımında FKS, WRT, ReliefF, BK vb. birçok filtreleme yaklaşımı kullanılsa da ikinci adımda sürü tabanlı GA ve PSO algoritmaları en çok tercih edilen algoritmalar olmuştur. Sınıflandırma kısmında ise başlıca tercih edilen algoritmalar DVM ve KNN olmuştur.

Literatürdeki çalışmalar hibrit, birlik, meta-sezgisel gibi ortak çalışan algoritmalar olarak isimlendirilse de bu çalışmalar içerisindeki adımların çoğu zaman birbirinden bağımsız olarak çalıştığı görülmektedir. Özellikle ilk adımda filtrelen genler sezgisel algoritmalara aktarıldıktan sonra filtreleme algoritmaları yaşam döngüsünü tamamlamaktadır. Her bir genin almış olduğu filtreleme puanı, sıralaması ve kaç gen içerisinden seçildiği gibi istatistiksel olarak anlamlı veriler bir sonraki adıma aktarılmamaktadır. Filtreleme adımının daha verimli kullanılabilmesi mümkünken, elde edilen veriler kaybedilmektedir. Filtreleme işlemi sadece ikinci adımdaki algoritmalarla değil literatürün büyük çoğunluğunda sınıflandırıcıdan da ayrı çalışmaktadır. Her ne kadar anlamlı genler bu ilk adımda belirlense de istatistiksel olarak anlamlı olmasa da sınıflandırma başarısı yüksek olan genler elenebilmektedir. Ayrıca bu adımda tek bir filtreleme algoritmasına bağımlı kalınması başka bir istatistiksel yaklaşıma göre anlamlı olup seçilme ihtimali olan genlerin de elenmesine yol açmaktadır.

Literatürde üzerinde durulmayan bir diğer husus ise mikro dizi verilerinin yapısıdır. Her bir mikro dizi farklı bir karakteristik özelliğe sahiptir. Bir veri için iyi sonuçlar elde eden bir algoritma başka bir veri için çoğu durumda başarılı

sonuçlar elde edememektedir. Bu durum algoritmanın bir veriyi ezberleyip iyi sonuçlar elde etmesine rağmen aslında literatürün geneline gerçek anlamda bir katkı sağlamamaktadır.

Yapılan çalışmalarda literatürde görülen eksiklerin giderilmesi için;

- Gen filtreleme adımında birden fazla algoritmadan oluşan bir yaklaşım önerilerek bir algoritma tarafından seçilmeyen anlamlı genlerin diğer bir algoritma ile seçilebilme olasılığı artırılmıştır.
- Filtreleme adımında filtrelenen genlerin sınıflandırma performans değerlerinin hesaplanarak istatistiksel olarak anlamlı olmasına rağmen başarısız olabilen genlerin dikkate alınması, filtreleme adımı ile sınıflandırma adımı arasındaki bağın güçlendirilmesi amaçlanmıştır.
- Filtreleme adımında seçilen genlerin önem derecesinin hesaplanması ve bu bilginin diğer adımlarda işlenebilmesi için yeni bir hesaplama yöntemi önerilmiştir.
- Farklı türdeki sınıflandırma algoritmalarından oluşan bir birlik algoritmanın gen seçimi işlemlerinin performansının değerlendirilmesi için kullanılması sayesinde geliştirilen yaklaşımın veri kümesinin ezberlenmesi problemine karşı güçlendirilmesi amaçlanmıştır.

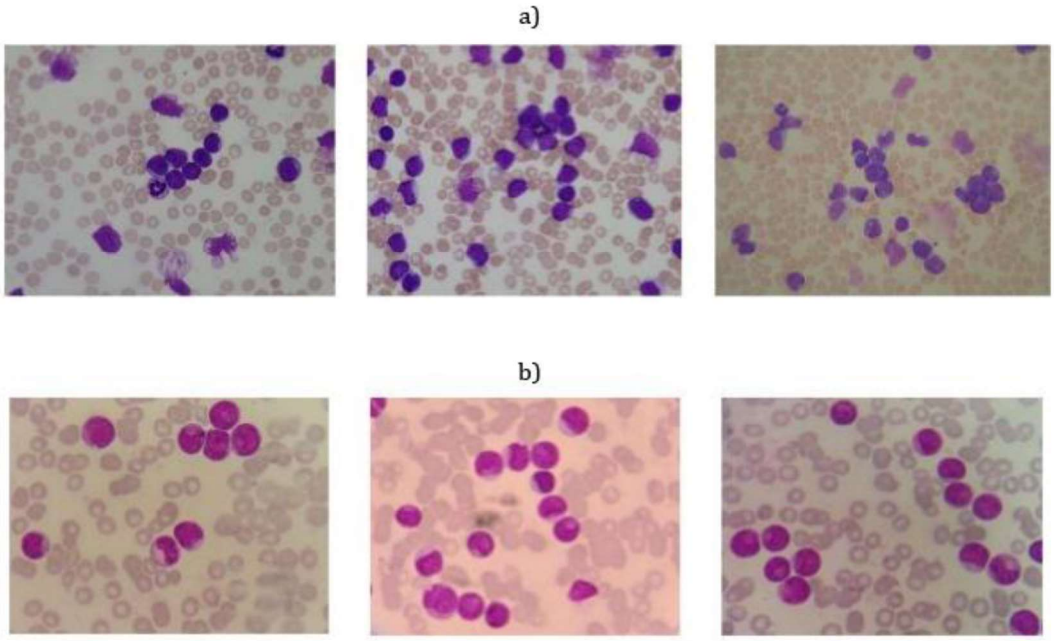
3. MATERYAL ve YÖNTEM

Mikro dizi teknolojisi canlı dokuların DNA seviyesinde incelenmesini ve bu sayede kanser gibi ciddi hastalıkların önceden teşhis edilebilmesi, biyolojik nedenlerinin açığa çıkarılması yada bu hastalık için ilaç geliştirilebilmesi gibi bir çok konuda önemli bir yere sahiptir. Tek seferde binlerce genin incelenebilmesine olanak vermesi nedeni ile bu teknoloji büyük bir avantaja sahiptir. Ancak bu avantajla birlikte veri boyutunun çok büyük olması, örnek sayısının az olması gibi bazı dezavantajları da bulunmaktadır. Bu dezavantajların ortaya çıkardığı güçlüklerin üstesinden gelebilmek için filtreleme, gen seçimi ve sınıflandırma gibi bir çok farklı yöntemin ve algoritmanın bir arada kullanılması gerekmektedir. Bu kısmında Lösemi veri kümesinin detayları verilerek bu verinin analizi için tez çalışması boyunca yararlanılan yaklaşımlardan ve algoritmalardan bahsedilmektedir.

3.1. Lösemi Veri Kümesi

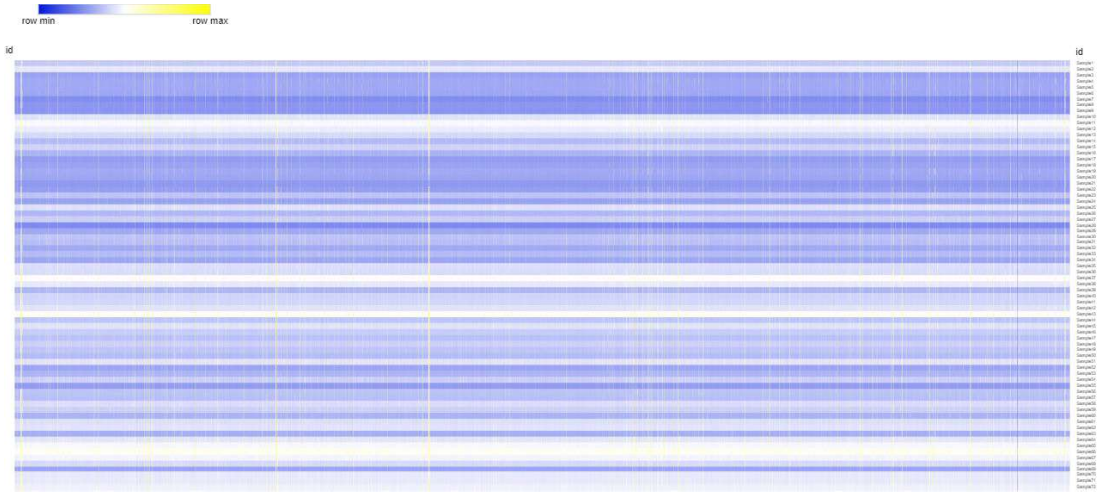
Lösemi kemik iliğindeki kök hücrelerinin hızlı bir şekilde çoğalması sonucunda ortaya çıkan, insan hayatını ve yaşam kalitesini tehdit eden bir hastalıktır. Akut lösemi kavramı ise hastalığın ortaya çıktıktan sonra agresif bir şekilde ilerlemesi ve kısa sürede ölümcül hale gelmesi durumunu olarak tanımlamaktadır (Geary, 2000). Akut löseminin, Akut Lenfoblastik Lösemi (ALL, Acute Lymphoblastic Leukemia) ve Akut Miyeloid Lösemi (AML, Acute Myeloid Leukemia) olmak üzere iki farklı türü bulunmaktadır. AML miyeloid hücrelerinin, ALL ise lenfoblast hücrelerinin aşırı hızlı artışı nedeni ile ortaya çıkmaktadır. Bu iki hastalığın birbirinden ayırt edilebilmesi, bu hastalıklara neden olan genetik nedenlerin ortaya çıkartılabilmesi teşhis, tedavi ve ilaç geliştirme çalışmaları için önemli bir yere sahiptir.

ALL ve AML hastalığına sahip dokulardan alınmış örneklerin mikroskop altında ki görünümü Şekil 3.1'de verilmiştir.



Şekil 3.1. ALL(a) ve AML(b) örneklerinin mikroskobik görünümü (Fatonah vd., 2018)

Bu çalışmada löseminin AML ve ALL türlerine ait dokuların mikro dizi teknolojisi ile elde edilmiş verileri üzerinde çalışılmıştır. Golub vd., (1999) tarafından hazırlanan bu verinin ilk 38 örneği lösemi teşhisi konmuş ama henüz kemoterapi almamış hastaların kemik iliğinden alınmış dokularından çıkarılmıştır. 27 örnek ise 1980 ve 1999 arasında Dana-Farber Kanser Enstitüsü tarafından tedavi edilen ALL hastalığına sahip hastalardan alınmıştır. Veri kümesinin diğer kısmı ise Kanser ve Lösemi Grubu B (Cancer and Leukemia Group B) adı verilen Amerika'da faaliyet gösteren bir araştırma grubuna ait verilerden rastgele seçilmiştir. Toplamda 72 adet örneğe sahip bu veri kümesi içerisinde 47 adet ALL, 25 adet AML verisi bulunmaktadır. Her bir örnek içerisinde 7129 adet gen ekspresyon değeri ve bir adet de sınıf parametresi (AML, ALL) bulunmaktadır. Veri kümesi içerisindeki genlerin ekspresyon seviyeleri Şekil 3.2'de ki sıcaklık haritası (Heatmap) üzerinde gösterilmiştir.



Şekil 3.2. Lösemi Mikro Dizi Gen Ekspresyon Seviyelerinin Sıcaklık Haritası üzerinde gösterimi

Sıcaklık haritaları genlerin birbirleri ve kanser ile olan ilişkilerinin tek seferde incelenebilmesi için görsel bir çıktı veren etkili araçlardır. Ancak mikro dizi verilerin sahip olduğu karmaşıklık verilen şekilde de görüldüğü üzere doğrudan bir inceleme işlemini zorlaştırmaktadır. Bu yüzden anlamsız genlerin veri kümesinden ayrıştırılması, en yüksek öneme sahip genlerin seçilmesi hastalığın yapısının farklı araçlar ile incelenebilmesi için de önemlidir.

3.2. Filtreleme ve Öznitelik Çıkarma

Detaylı bir gen seçimi için kullanılan tekrarlı sezgisel algoritmalar ve son adımda çalışan sınıflandırma algoritmaları, yapıları gereği yavaş çalışmakta ve barındırdıkları karmaşık denklemler nedeni ile yüksek bir hesaplama maliyetine sahiplerdir. Yüksek boyutlu bir veri bu algoritmaların çok daha fazla işlem yaparak uzun süreler boyunca çalışmasına ya da doğrudan uyumsuzluk sorununa yol açabilmektedir. Dolayısı ile veri boyutu bu algoritmalar için büyük önem taşımaktadır. Bu yüzden mikro dizi verilerinin detaylı bir gen seçim işlemi ya da sınıflandırma işlemine tabi tutulmadan önce veri boyutunun azaltılması için bir filtreleme veya öznitelik çıkarma adımına ihtiyaç duyulmaktadır. Ancak veri boyutunun azaltılırken anlamlı genlerin ve genler arasındaki ilişkilerin taşıdığı bilgilerin kaybolmaması için bu adımın dikkatlice planlanması gerekmektedir.

Bu başlıkta veri kümesinin analizinin ilk adımında kullanılan ve tez çalışmasında yararlanılan filtreleme ve öznitelik çıkarma yöntemlerinden bahsedilmiştir. Normalizasyon işleminin, algoritmaların başarısına doğrudan katkı sağlaması ayrıca bazı algoritmaların sadece normalize edilmiş verilerle çalışabilmesi nedeni ile ön işlem için kullanılan normalizasyon yönteminden de bahsedilmiştir.

3.2.1. Fisher Korelasyon Skoru

FKS basit ve hızlı çalışan bir algoritma olması nedeni ile makine öğrenmesinde ve mikro dizi verileri üzerinde yapılan çalışmaların ilk adımında filtreleme amacı ile sıklıkla tercih edilen algoritmalardan biridir (Xiong vd, 2001). FKS'nin amacı farklı sınıfları temsil eden ancak özellikler arasındaki mesafelerin aynı olduğu alt kümeler bulmak olarak özetlenebilir (Gu vd., 2012). Mikro dizi verisi içerisindeki genleri FSK ile puanlamak ve en yüksek puana sahip genleri seçebilmek için Denklem (3.1)'den yararlanılmıştır (Duda vd., 2001).

$$x_i = \frac{\sum_{j=1}^{c} n_j (u_{ij} - u_i)^2}{(\sigma_i)^2} \tag{3.1}$$

x_i i geninin Fisher Korelasyon Skorunu verirken, her bir j sınıfı için denklemde verilen u_{ij} j sınıfındaki i. gene ait ekspresyon değerlerinin ortalamaları, u_i i. gene ait tüm gen ekspresyon değerlerinin ortalamasını göstermektedir. σ_i ise standart sapmayı temsil etmektedir. Son olarak n_j, j sınıfı içerisindeki toplam örnek sayısını ifade etmektedir. Bu hesaplama kullanılarak bir genin farklı sınıflara ait örneklerdeki ekspresyon seviyelerinin anlamlı bir değişim gösterip göstermediği ölçülebilmektedir ve bu değişimin büyüklük seviyesine göre bir puanlama yapılmaktadır. Böylece farklı sınıflarda en fazla değişim gösteren gen en yüksek puana sahip olmaktadır. Yapılan FKS puanı sıralamasından sonra düşük puana sahip genler veri kümesinden elenerek filtreme işlemi tamamlanmaktadır.

3.2.2. Bilgi Kazanımı

Bilgi kazancı (BK), veri kümesi içerisindeki örnekleri seçilen bir gene veya sınıf parametresine göre bölümledikten sonra ortaya çıkan entropideki azalma olarak tanımlanmaktadır (Mitchell, 1997). Bilgi kazanımı bir genin farklı sınıflar içerisindeki dağılımını (varyans, variation) ölçmektedir. Entropi prensibini kullanan bu yöntem ile bir genin farklı sınıflar içerisinde almış olduğu değerlerin değişimi incelenerek ne kadar bilgi taşıdığı ortaya çıkarılmaktadır. Eğer seçilen gen farklı sınıflarda farklı değerler almıyorsa FKS'da da olduğu gibi daha az bilgi taşıdığı veya yüksek entropi taşıdığı söylenebilir. Bunun için önce sınıf parametresinin entropisinin hesaplanması gerekmektedir. Bu hesaplama için Denklem (3.2)'den yararlanılmıştır (Shannon, 1948).

$$H(A) = -\sum_{a \in A} P(a) log\ P(a) \tag{3.2}$$

H(A) sınıf parametresi için olasılık dağılımının entropisini, P(a) a sınıfına ait her örneğinin tüm veri kümesi içerisindeki olasılığını temsil etmektedir. P(a) a sınıfına ait örnek sayısının, tüm örneklerin sayısına bölümü ile Denklem (3.3)'de gösterildiği gibi hesaplanmaktadır.

$$P_{(a)} = \frac{n_a}{n} \tag{3.3}$$

Verilen denklemde n tüm örnek sayısını temsil ederken $n_a$sadece a sınıfına ait örneklerin sayısını temsil etmektedir. Entropi hesaplamasından sonra her bir gen için sınıfa bağımlı entropi Denklem (3.4)'de verilen yöntem ile hesaplanmaktadır (Huang vd., 2007). Ancak mikro dizi veri kümelerinde veriler sürekli olduğu için olasılık hesaplama işleminden önce her bir öznitelik için bir eşik değer belirlemek gerekmektedir. Çalışmada bu eşik değer her bir genin kendi içerisinde ortalamasını alarak belirlenmiştir. Bu eşik değere göre ortalamanın üstünde ve altında olan değerler iki farklı olasılığı işaret edecek

şekilde süreksiz verilere dönüştürülmüştür. Son olarak bilgi kazanımının hesaplanması için Denklem (3.5)'de verilen hesaplamadan yararlanılmaktadır.

$$H(A|B) = -\sum_{b \in B} P(b)(\sum_{a \in A} P(a/b)) \log P(a/b)) \quad (3.4)$$

$$IG(B) = H(A) - H(A|B) \quad (3.5)$$

Verilen denklemlerde B bilgi kazanımı hesaplanmak istenen geni, $H(A|B)$ B geninin sınıf parametresine bağımlı olasılığını, b ise B genine ait her bir farklı ekspresyon değerinin sayısını ifade etmektedir. Son olarak IG(B), B geninin bilgi kazanımını göstermektedir. Her bir gen için bilgi kazanımı değerinin hesaplanmasından sonra bu değerler sıralanarak düşük puana sahip genler veri kümesi içerisinden temizlenmektedir.

3.2.3. Wilcoxon Rütbeler Toplamı

Wilcoxon Rütbeler Toplamı (WRT) yöntemi filtrelemek istenen veri kümesine ait örnek sayısının az olduğu durumlarda daha iyi sonuçlar vermektedir. Bu yüzden mikro dizi verileri üzerinde iyi bir avantaja sahiptir. WRT'nin hesaplanabilesi için bir genin iki farklı sınıf parametresine sahip örnekleri arasındaki farkın Denklem (3.6)'te verilen yöntem ile bulunması gerekmektedir (Liao vd., 2006).

$$s(g) = \sum_{i \in n_0} \sum_{j \in n_1} I((x_j^{(g)} - x_i^{(g)}) \leq 0) \quad (3.6)$$

Denklemde verilen I, içerisindeki ifadenin doğru olması durumunda 1, yanlış olması durumunda 0 sonucunu üreten bir fonksiyondur. $x_i^{(g)}$ ve $x_j^{(g)}$, g geninin i ve j örneklerinde ölçülen ekspresyon değeridir. n_1 ve n_0 simgeleri, g geninin farklı sınıflardaki örnek sayısını temsil eden değişkenlerdir. $s(g)$, g geni için iki sınıf arasındaki farkı ifade etmektedir. Elde edilen sonuç 0'a veya iki farklı sınıfa ait tüm kombinasyonların toplam sayısına yaklaştıkça ilgili genin daha önemli

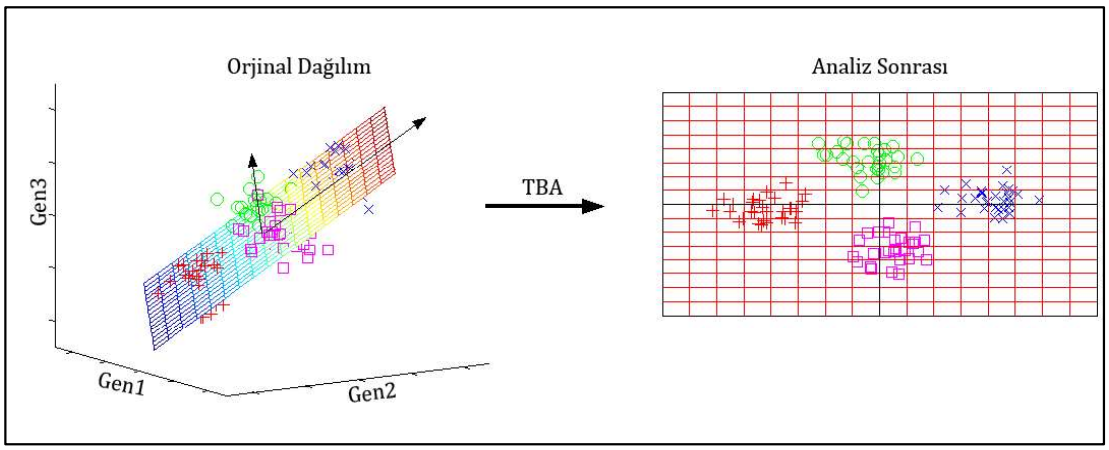
olduğu sonucu ortaya çıkmaktadır. Bu doğrultuda nihai WRT puanının hesaplanması için Denklem (3.7)'den yararlanılmıştır.

$$WRT(g) = \max(s(g), n_1 n_0 - s(g)) \tag{3.7}$$

Denklemde verilen max fonksiyonu içerisindeki parametrelerden en büyük olanını sonuç olarak üretmektedir. Bu hesaplamadan sonra diğer filtreleme yöntemleri ile benzer şekilde en yüksek WRT puanına sahip genler seçilerek diğer genler veri kümesi içerisinden temizlenmektedir.

3.2.4. Temel Bileşen Analizi

Diğer filtreleme yaklaşımlarından farklı olarak Temel Bileşen Analizi (TBA) bir gen ya da öznitelik seçimi için kullanılmamaktadır. Bunun yerine veri kümesi içerisinden önemli bilgileri çıkararak bunların yeni öznitelikler olarak kullanılabilmesini sağlamaktadır. Bu analizin yapılmasındaki başlıca amaç verilerin birbirleri arasındaki varyansının korunarak daha basit bir şekilde gösterilmesidir (Jolliffe, 2002). Verilerin yönleri ve aralarındaki bağıntıların değerleri kaybolmadan verinin daha basit ve anlamlı gösterilebileceği başka bir koordinat sistemine taşınması olarak da tanımlanabilmektedir (Şekil 3.3) .



Şekil 3.3. TBA ile verilerin yeni bir koordinat düzlemine taşınması (Scholz, M., 2006)

TBA analizinin yapılabilmesi için öncelikli olarak her bir gen ekspresyon değerinin bir vektör olarak ele alınması gerekmektedir. Bu vektörlere ait ortalamalar ve vektörlerin birbirlerine bağımlı olarak göstermiş oldukları değişimin yani kovaryansın bulunması gerekmektedir (Adiwijaya vd., 2018). Gen ekspresyon vektörlerine ait ortalama vektörü Denklem (3.8)'de gösterildiği şekilde hesaplanırken kovaryans matrisi Denklem (3.9)'de verilen yöntemle elde edilmektedir.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.8}$$

$$C_X = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T \tag{3.9}$$

Verilen denklemde n toplam gen sayısını ifade ederken, X_i ise i genine ait vektörü göstermektedir. Korelasyon matrisinin bulunmasından sonra öz bileşen vektörü ve öz bileşen değerlerinin hesaplanabilmesi için (3.10)'dan yararlanılmaktadır.

$$C_X V_m = \lambda_m V_m \tag{3.10}$$

Verilen denklem de V_m öz bileşen vektörlerini, λ_m öz bileşen değerlerini ifade etmektedir. Bu noktadan sonra öz bileşen değerlerine göre bir sıralama yapılarak aynı sıradaki öz bileşen vektörleri yeni öznitelikler olarak kullanılabilmektedir. Ancak burada dikkat edilmesi gereken husus seçilen öz bileşen vektörlerine ait öz bileşen değerlerinin toplamının toplam öz bileşen değerlerinin yüzde kaçını kapsadığıdır. Bu oran için net bir hesaplama yöntemi bulunmasa da literatürde %95'i geçen bir değer kabul görmüştür.

TBA karmaşık verilerin daha basit bir şekilde ifade edilerek görsel olarak da incelenebilmesini sağlayan etkili bir istatistiksel yöntemdir. Sınıflandırma başarısına birçok durumda olumlu bir katkı da sağlamaktadır. Ancak

sınıflandırma başarısına ek olarak mikro dizi veri kümelerinin analizinde genlerin hastalıkla olan ilişkisinin de ortaya çıkarılması da istenmektedir. TBA yönteminde elde edilen değerler, gen ekspresyon değerleri yerine kullanıldığı için bu adımdan sonra kanser-gen ilişkilerinin daha fazla incelenmesi mümkün olmamaktadır. Bu yüzden mikro dizi veri kümesinin karakteristik özelliğine ek olarak yapılacak analizin amacı da, kullanılacak algoritmalara karar verilirken üzerinde durulması gereken önemli konulardan biridir.

3.2.5. Normalizasyon

Ön işleme adımı, verilerin anlamlandırılması sürecinde başarıyı artıran etkenlerden biridir. Özellikle sezgisel algoritmalar ham veriler ile daha düşük performans sergilemekte bazı durumlarda ise uyumsuzluk problemi yaşamaktadırlar. Bu nedenle Normalizasyon yaygın olarak kullanılan ön işlem adımlarından biridir. Literatürde min-maks, standart sapma, z-skor, ondalıklı ölçekleme vb. birçok farklı türde normalizasyon yöntemi bulunsa da (Jain vd., 2018), bu çalışmada verilerin 0 ve 1 arasında ölçeklendiği min-maks normalizasyon yönteminden yararlanılmıştır. Min-maks normalizasyonu Denklem (3.11)'de verilen hesaplama ile gerçekleştirilmektedir.

$$N(X_i) = \frac{X_i - min_X}{maks_x - min_x} \tag{3.11}$$

Denklemde verilen X_i, X genine ait i. örnekteki ekspresyon değerini temsil etmektedir. Bununla beraber min_X X genine ait en düşük gen ekspresyon seviyesini, $maks_x$ ise aynı gene ait en yüksek gen ekspresyon seviyesini göstermektedir. Bu sayede her bir genin ekspresyon değerleri 0 ile 1 arasına taşınarak, genler arasındaki orantısız büyüklükten kaynaklanan uyumsuzluk problemleri önlenmektedir.

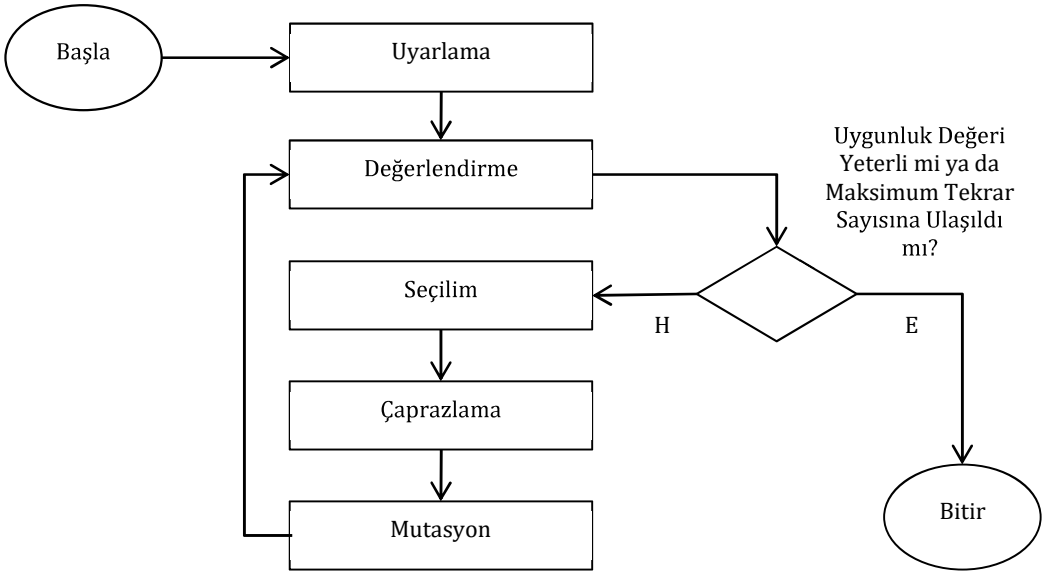
3.3. Gen Seçimi

Gen seçimi filtreleme aşamasından farklı olarak istatistiksel anlama değil sınıflandırma başarısına odaklanmaktadır. Gen seçimindeki temel amaç en az sayıda gen seçerek en yüksek sınıflandırma başarısı elde edebilmektedir. Bu seçim işlemlerinin gerçekleştirilebilmesi için literatürde sürü temelli ve tekrarlı algoritmalar yaygın olarak kullanılmaktadır. Bu çalışmada sınıflandırma başarısı yüksek, kanser-gen ve gen-gen ilişkileri için anlam taşıyan genlerin belirlenebilmesi için sürü temelli ve tekrarlı çalışma prensibine sahip bir algoritma olan GA'dan yararlanılmıştır.

3.3.1. Genetik algoritma

Genetik Algoritma (GA) canlıların hayatta kalma becerilerini taklit eden, sürü prensibi ile çalışarak aynı anda birden fazla olası çözüm yolunu araştıran sezgisel bir algoritmadır (Holland, 1975). Algoritmanın sahip olmuş olduğu doğada ki evrimsel süreçlerden esinlenilerek oluşturulan Seçilim, Çaprazlama gibi işlevler sayesinde en başarılı çözüm kümeleri bir sonraki tekrara aktarılabilmekte, çözüm kümeleri aralarında parametre değişimi yaparak daha başarılı çözüm yolları elde edebilmektedir. En güçlü yönlerinden biri olan mutasyon işlevi sayesinde, GA belli bir örüntüye uymayan ve önceden tahmin edilemeyen parametreleri çözüm kümesi içerisine dâhil edebilmekte ve böylece sadece yerel çözümlere bağlı kalmadan farklı çözüm yollarını da araştırılabilmektedir.

GA'nın literatürdeki gen seçim işlemlerinde elde etmiş olduğu başarılı sonuçlar, sahip olduğu işlevlerinin gen seçiminde büyük bir avantaj sağlaması ve amaca uygun olarak kolaylıkla şekillendirilebilmesi bu çalışmada tercih edilmesinin başlıca nedenlerindendir. GA'nın işlevleri ve öğelerinin daha iyi anlaşılması için Şekil 3.4'de Geleneksel bir GA'nın akış diyagramı verilmiştir.



Şekil 3.4. Geleneksel bir GA'nın akış diyagramı

Uyarlama adımında problemin GA'nın öğelerine uygun hale getirilmesi gerekmektedir. Temel olarak bir GA içerisinde Gen, Kromozom ve Popülasyondan öğelerini barındırmaktadır (Beasley vd., 1993; Man vd., 1996).

Gen; GA'nın bilgi taşıyan en küçük birimidir. Çözüm kümesi içerisindeki parametrelere ait bilgiler genler içerisinde saklanır. Bilgiler genler içerisine İkili kodlama, permütasyon kodlama veya değer kodlama şeklinde yazılabilir. Bu çalışmada ki kullanılan GA'nın amacı gen seçimi olduğu için bu genler üzerinde mikro dizi veri kümesine ait genlerin isimleri tutulmak istenmektedir. Bu isim bir numara ile temsil edilmesi ve bu numaranın aritmetik bir önem taşımaması nedeni ile ikili kodlama yerine değer kodlama tercih edilmiştir. Bu durumda GA'nın gen öğeleri içerisinde bilgi olarak Mikro dizi veri kümesi içerisindeki genlerin numaralarını saklamaktadır. GA ve Mikro dizi veri kümesi içerisindeki Gen kavramının birbiri ile karıştırılmaması için iki kavramın beraber kullanıldığı durumlarda GA'nın gen öğesi için gen(GA), mikro dizi verisi içerisindeki genler için ise gen(MD) kullanılacaktır.

Kromozom; genler(GA)'den oluşan olası bir çözüm kümesidir. Mikro dizi problemine uyarlandığında içerisinde seçilmek istenen gen(MD) sayısı kadar gen(GA) barındırmaktadır.

Popülasyon; GA'nın sahip olmuş olduğu, aynı anda arama yapabilmesine olanak veren tüm kromozomlara verilen isimdir. Problemin karmaşıklığı ve araştırma yapılmak istenen olası çözüm kümelerinin sayısını artırmak için istenilen büyüklükte seçilebilmektedir.

Değerlendirme; Seçilen genlerin başarısının ölçüldüğü önemli adımlardan biridir. Belirlenen bir amaç fonksiyonuna göre her bir kromozomun elde ettiği uygunluk değeri bu adımda hesaplanmaktadır. Amaç fonksiyonları, problemde hesaplanmak istenen, minimize ya da maksimize edilmek istenen bir değer için farklı denklemlere sahip olmakta, her probleme özgü olarak belirlenmektedir. Mikro dizi verilerinde seçilen genlerin sınıflandırma başarısı uygunluk değeri olarak kabul edilirken, bu değerin hesaplanması aşamasında kullanılan sınıflandırma vb. işlemler de amaç fonksiyonu olarak kabul edilmektedir.

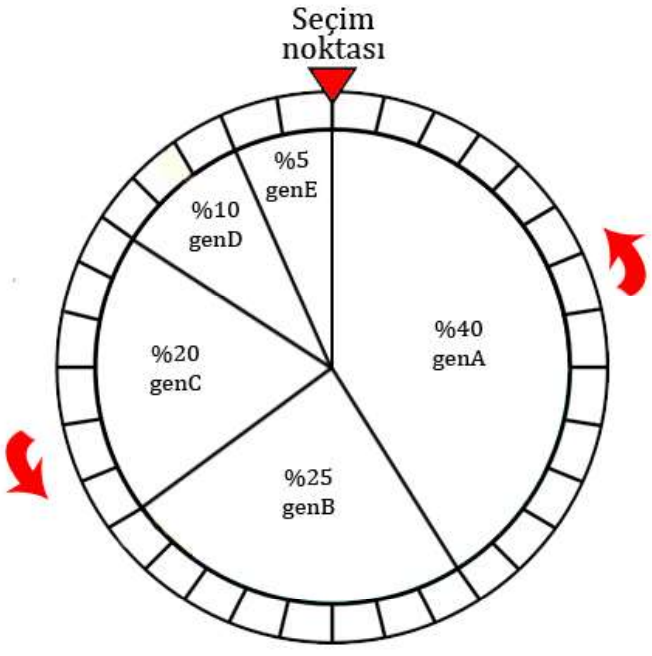
Değerlendirme adımında hesaplanan, kromozomlara ait uygunluk değerlerinden GA'nın birçok işlevi içerisinde yararlanılmaktadır. Seçilim işlevinde en başarılı kromozomların gelecek kuşaklara taşınması, çaprazlanacak kromozomlara karar verilmesi, mutasyona uğrayacak genlerin seçilmesi gibi işlevlerin yanında GA'nın yeterli uygunluk değerine ulaşması durumunda tekrarları durdurması için yapılan karşılaştırma işleminde de bu değer kullanılmaktadır. Bu nedenle problem ile uyumlu, etkili bir amaç fonksiyonunun belirlenmesi GA'nın performansı ve seçeceği genlerin kalitesi için kritik bir öneme sahiptir.

Seçilim adımı, değerlendirme aşamasından sonra eğer istenilen uygunluk değerine erişilmemişse kullanılan bir GA işlevidir. Bu işlevin iki amacı vardır. Bu amaçlardan biri bir sonraki kuşak için en başarılı kromozomların seçilmesi iken diğeri çaprazlama işlevlerinde kullanılacak kromozomların belirlenmesidir.

Eğer GA ilk tekrar da değilse popülasyon sayısından daha fazla kromozom sayısı ortaya çıkacaktır. Bu noktada seçilim işlevi popülasyon parametresini dikkate alarak uygunluk değeri en yüksek kromozomları bir sonraki kuşak için

ayırmakta, daha başarısız olan kromozomlar ise elenmektedir. Bu seçim işlemi sırasında Altın Birey yöntemi adı verilen en başarılı kromozomun çaprazlama veya mutasyon işleminden geçmeden doğrudan yeni kuşağa aktarılması gibi farklı yaklaşımlar da seçilim işlemlerine ek olarak başarının artırılması için kullanılabilmektedir. Bir sonraki kuşak ya da çaprazlama işlevinde kullanılacak kromozomların seçimi için literatürde Rulet Tekerleği (Rulet Wheel), Turnuva (Tournament), Sabit Durum (Steady State), Seçkincilik (Elitism), Rütbe (Rank) vb. birçok yöntem bulunmaktadır. Çalışma içerisinde ise Rulet Tekerliği tercih edilmiştir.

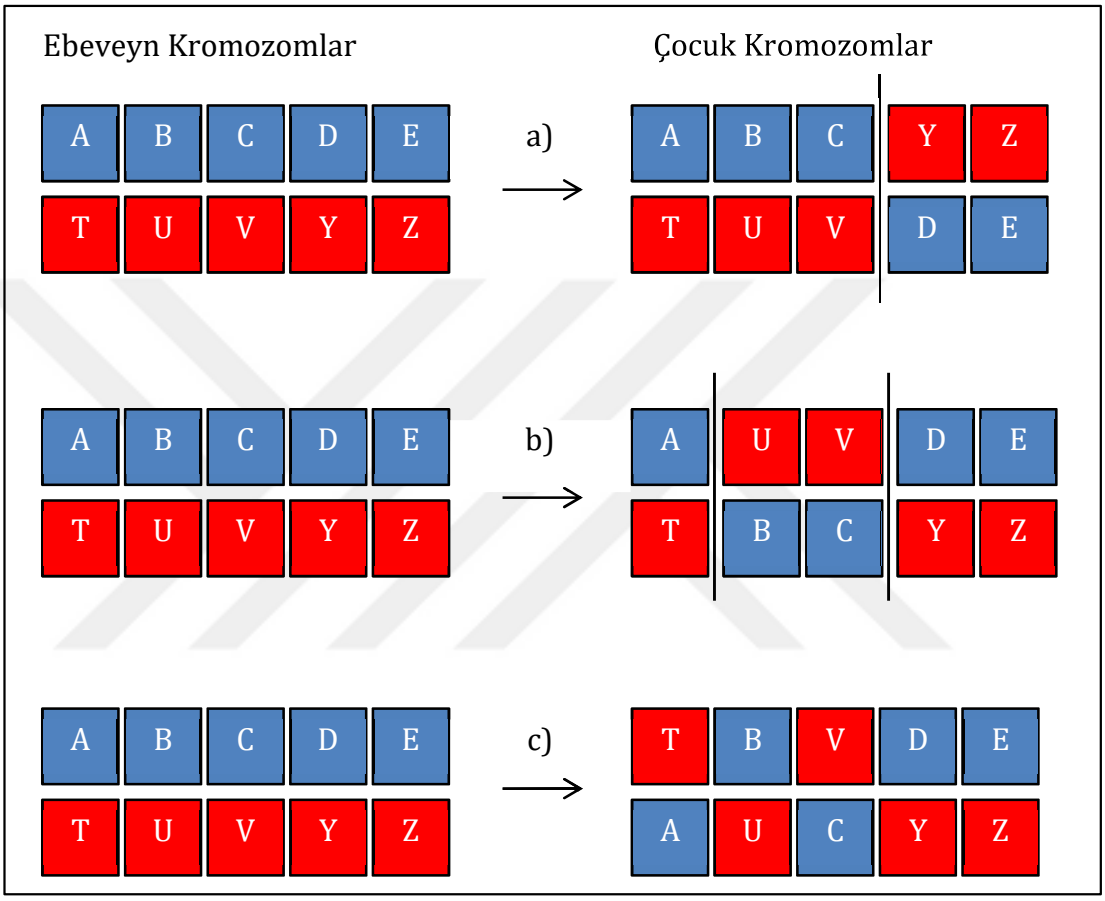
Her bir kromozom için uygunluk değerinin hesaplanmasından sonra bu olasılıklar, uygunluk değerleri oranında yer kaplayacak şekilde rulet tekerleğine yerleştirilir (Şekil 3.5). Rastgele bir sayı üretilerek, bu sayının tekerlek üzerinde isabet ettiği noktadaki kromozom seçilir. Bu yöntem sayesinde uygunluk değeri yüksek olan kromozomdan düşük olana doğru seçilme olasılığı düşmekte ancak bu şans hiçbir zaman sıfır olmamaktadır.



Şekil 3.5. Olasılıkların Rulet Tekerleği Üzerinde Gösterimi

Çaprazlama işlevinin temel amacı, başarılı kromozomların birbirleri arasında gen takası yaparak daha başarılı yeni çocuk kromozomlar elde etmektir. Tek

noktalı, çok noktalı ve tek düze yöntemleri başta olmak üzere literatürde birçok farklı çaprazlama yöntemi bulunmaktadır (Şekil 3.6). Diğer işlevlerde de olduğu gibi probleme özel bir çaprazlama yönteminin seçilmesi ya da geliştirilmesi gerekmektedir. Mikro dizi veri kümelerinde gerçekleştirilen gen seçim işlemlerinde Tek Düze yöntem ve türevleri yaygın olarak kullanılmaktadır.



Şekil 3.6. a) Tek noktalı, b) Çok noktalı ve c) Tek Düze Çaprazlama Yöntemi

Mutasyon olası çözüm kümelerinin hep aynı genlere sahip olması durumunda yerel çözümlere takılı kalmasını önlemek için kullanılan bir işlevdir. Mutasyon oranı dikkate alınarak rastgele seçilen bir gende yapılan rastgele değer değişiklikleri olarak tanımlanmaktadır. Mikro dizi verilerinde mutasyon kromozom içerisindeki bir genin, gen havuzu içerisindeki farklı bir gen ile değiştirmesi ile gerçekleştirilmektedir.

GA en başarılı uygunluk değerine sahip kromozomu elde edene ya da maksimum tekrar sayısına ulaşana kadar bu işlevler akış diyagramında görülen sırada tekrar etmektedir. GA tekrarları sonucunda popülasyon içerisindeki en başarılı kromozomun sahip olduğu genler çözüm kümesi olarak kabul edilmektedir.

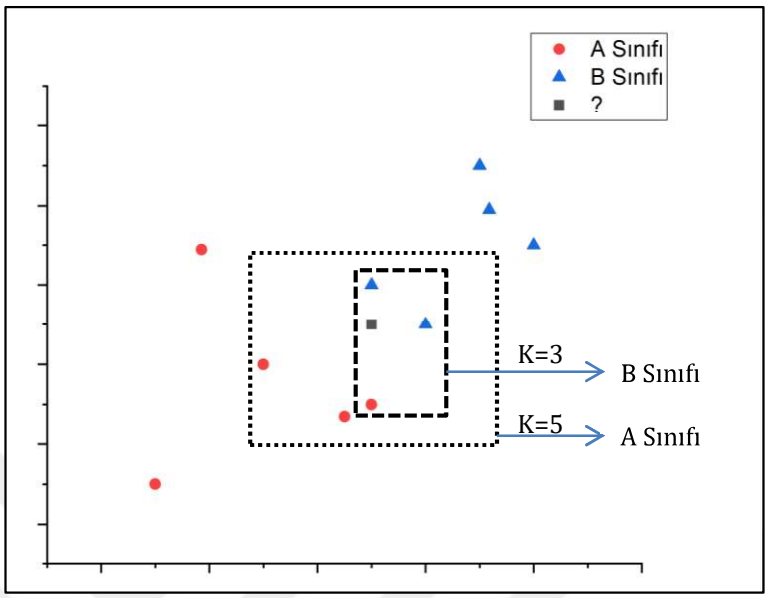
3.4. Sınıflandırma

Mikro dizi verilerinin sahip olmuş olduğu yüksek veri boyutu ve karmaşıklığı sınıflandırma algoritmalarının verimli bir şekilde çalışmasını güçleştirmektedir. Bu yüzden yapılan filtreleme ve gen seçimi işlemleri veri karmaşıklığını büyük ölçüde azaltsa da anlamlı genlerin bu aşamalarda kaybolma riski, ve hali hazırda az veri örneğine sahip olması sınıflandırma başarısını büyük ölçüde düşürmektedir. Bu nedenle mikro dizi verilerinin sınıflandırılması için seçilecek algoritmaların bu zorlukları tolere edebilecek yapıda olması gerekmektedir. KNN, DVM ve Naif Bayes sınıflandırma algoritmaları literatürde mikro dizi sınıflandırma işlemleri için sıklıkla başvurulan güçlü ve etkili sezgisel sınıflandırma algoritmalarıdır.

3.4.1. KNN

KNN, tüm verileri kullanarak öğrenme aşamasına ihtiyaç duymayan ve bu veriler arasındaki uzaklık farkını kullanarak sınıflandırma işlemleri gerçekleştirebilen hızlı bir algoritmadır. KNN sınıflandırmak istenen örneğin gen ekspresyon seviyelerine en yakın olan örneklerin (komşuların) yoğunlaştığı sınıflara bakarak bir tahmin gerçekleştirmektedir (Şekil 3.7). Bir örneğin sınıfına karar verilmesi için o örneğe en yakın tek bir komşuya bakmak yerine en yakın komşulara bakmak sınıflandırma için daha doğru sonuçlar vermektedir (Cover ve Hart, 1967). En yakın kaç adet komşunun sınıf bilgisine bakılacağı K adı verilen parametre ile belirlenirken, örnekler arasındaki uzaklığın bulunması için Öklid, Manhattan ve Minkowski vb. uzaklık hesaplama yaklaşımları kullanılmaktadır. Bu çalışmada literatürde yaygın olarak kullanılan

Öklid uzaklık ölçüm yöntemi tercih edilerek Denklem (3.12)'de verilen hesaplamadan yararlanılmıştır.



Şekil 3.7. KNN Sınıflandırma işleminin grafik üzerinde gösterimi

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{3.12}$$

X ve Y aralarındaki uzaklığın ölçülmesi istenen iki örnek olduğu durumda, X'e ait öznitelikler X_i, Y'ye ait öznitelikler Y_i ile gösterilmektedir. n ise örneklerin sahip olduğu toplam öznitelik sayısını vermektedir. Mikro dizi verilerine ait genler arasındaki mesafe bu yöntemle hesaplanmak istendiğinden X ve Y birer veri örneği ve X_i ve Y_i ise bu örnekler içerisindeki genlerin ekspresyon değerlerini temsil etmektedir. Bu hesaplama ile bir veri örneğe en yakın olan K adet komşu örnek belirlenmektedir ve bu komşu örnekler içerisinde en yoğun olan sınıf, sınıflandırma sonucu olarak kabul edilmektedir.

3.4.2. Naif Bayes

Naif Bayes algoritması, Bayes teoreminden temel alan ve KNN gibi hızlı bir şekilde etkili sınıflandırma sonuçları üretebilen olasılık tabanlı bir sınıflandırma

algoritmasıdır. Bu algoritmada sınıflandırma işlemi, örneklerin sınıflara ait olma olasılığının hesaplanması ile gerçekleştirilmektedir. Her bir örnek için tüm sınıflara ait olma olasılığı tek tek hesaplanarak bu olasılıklar içerisinde en yüksek değere sahip olan sınıf, örneğin sınıflandırma sonucu olarak Denklem (4.6)'da verilen hesaplama ile belirlenmektedir.

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{3.13}$$

Denklemde verilen $\hat{y}$ sınıflandırma sonucunda tahmin edilen sınıfı gösterirken, $\arg\max_{y}$ metodu ifadeyi en yüksek yapan y, yani sınıf değerini seçen bir metottur. $P(y)$, y sınıfına ait örneklerin veri kümesi içerisindeki toplam örnek sayısına bölümünü veren olasılık hesabıdır. n sınıflandırılmak istenen örneğe ait toplam öznitelik sayısıdır. x_i, x örneğine ait i. özniteliğini göstermektedir. $P(x_i \mid y)$ hesaplaması ise süreksiz verilerde her bir özniteliğin y sınıfına ait toplam örnek sayısına bölümü ile yapılabilmektedir. Bu veriler ışığında, bu çalışmada, WRT algoritmasından farklı olarak, mikro dizi gen ekspresyon verilerinin sürekli olması nedeni ile eşik değer yaklaşımı yerine Gaussian metodu kullanılmıştır (Denklem (3.14)).

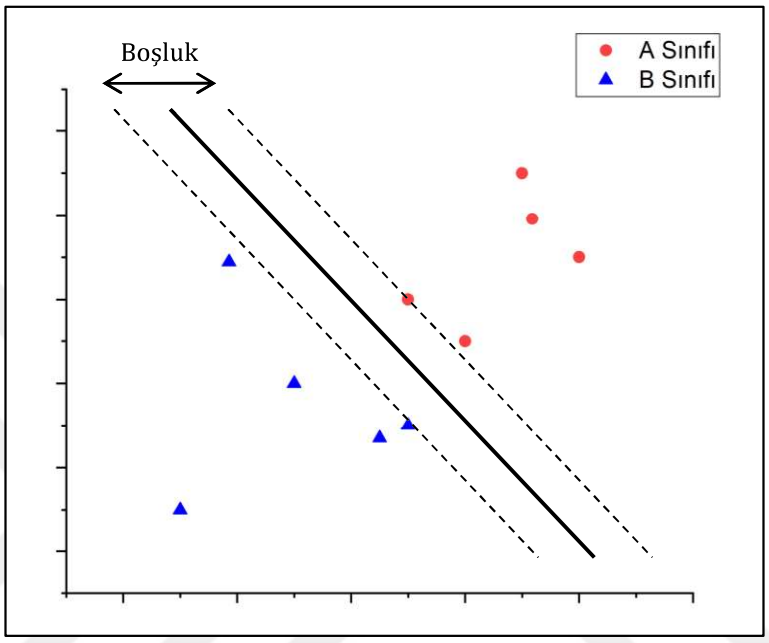
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{3.14}$$

Her bir özniteliğin, y sınıfına ait olma olasılığının hesaplanması için kullanılan bu denklemde σ, y sınıfı içerisindeki örneklerin i genlerine ait ekspresyon değerlerinin standart sapmasını, μ ise ortalamasını temsil etmektedir. exp ise üstel (exponential) fonksiyondur.

3.4.3. Destek vektör makinesi

DVM, sınıflandırma ve regresyon analizi için kullanılan, denetimli öğrenme yöntemine sahip bir makine öğrenme algoritmasıdır (Cortes ve Vapnik, 1995).

Çalışma prensibindeki ana fikir veri kümesine ait farklı sınıflar içerisindeki örneklerin bir düzlem üzerinde birbirinden ayırabilecek vektörlerinin çizilebilmesi, böylece yeni verilerin bu vektöre göre konumuna bakılarak sınıflandırma tahmininin yapılabilmesidir (Şekil 3.8). Bu tahminin yapılması için kullanılan karar verme fonksiyonu Denklem (3.15)'de verilmiştir.



Şekil 3.8. DMV Çalışma prensibi

$$f(X) = sign(\sum_{i=1}^{n} \alpha_i \, y_i \, \emptyset(X) \, \emptyset(X_i)) \tag{3.15}$$

Denklemde verilen $f(X)$ sınıflandırmak istenen X örneğine ait karar verici fonksiyonu temsil etmektedir. sing metodu içerisindeki ifadenin sadece işaretini geri döndürmektedir. Bu işaretin pozitif veya negatif olması X örneğinin hangi y sınıfına ait olacağını belirlemektedir çünkü n veri kümesi toplam örnek sayısı olmak üzere $y_i \in \{-1, +1\}$ her bir örneğin sınıf parametresini temsil etmektedir. α_i çarpanının en iyilenmesi için Denklem (3.16)'de verilen Larange hesaplamasından yararlanılırken, bu hesaplamada ki kısıtlar (3.17)'de verilmiştir. Mikro dizi verilerinin doğrusal olmayan yapısı neden ile $\emptyset(X)\emptyset(X_i)$ skaler çarpımı için $K(X_i, X_j)$ dönüşümü yapılarak Denklem (3.18)'da verilen Radyal Tabanlı Çekirdek fonksiyonu kullanılmaktadır.

$$\max_{a} L_D = \max_{a} (\sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j K(X_i, X_j)), \quad (3.16)$$

$$0 \leq a_i \leq C, \qquad \sum_{i=1}^{n} a_i y_i = 0 \quad (3.17)$$

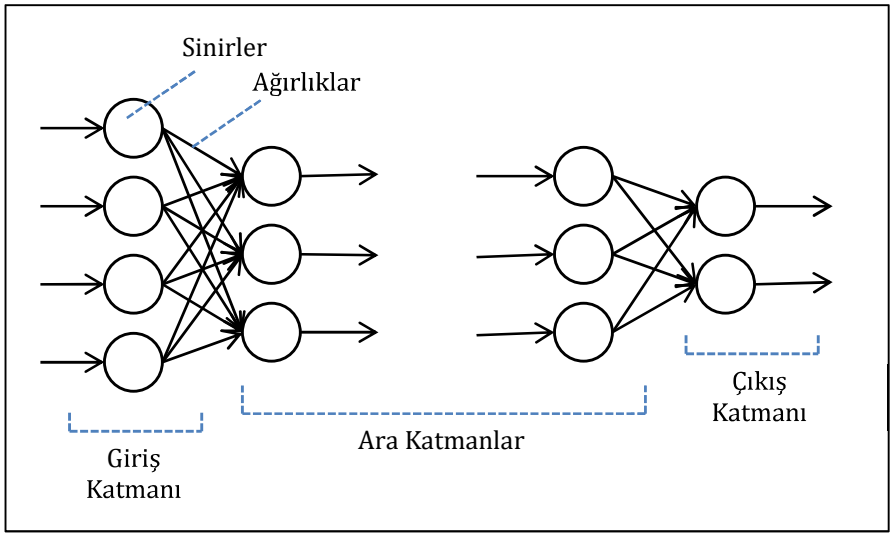
$$K(X_i, X_j) = exp(-\gamma \parallel x_i - x_j \parallel^2), \quad (3.18)$$

Denklemde verilen $C > 0$ olmak üzere karar verme üst sınır katsayısı iken $\gamma > 0$ olmak üzere Radiyal Tabanlı Çekirdek fonksiyonunun gama katsayısıdır.

3.4.4. Yapay sinir ağları

Yapay sinir ağları (YSA) insan beyninin çalışma prensibini ve sahip olduğu sinirlerin matematiksel olarak ifade edilmesi ile insan problem çözme becerisinin taklit edilmesini amaçlayan bir algoritmadır. Bu özelliği sayesinde sınıflandırma başta olmak üzere optimizasyon, tahmin ve örüntü tanıma gibi birçok problemin çözümünde kullanılabilmektedir (Blanton, 1997).

YSA içerisindeki bir yapay sinir hücresinin girdi, ağırlıklar, toplama fonksiyonu, aktivasyon fonksiyonu ve çıktı olmak üzere 5 ana öğesi vardır. Temel bir YSA'da bu sinir hücreleri giriş, ara ve çıkış olmak üzere 3 katman içerisinde sıralı bir şekilde çalışmaktadır (Şekil 3.9). Sınıflandırma amacı ile kullanıldığı durumlarda veri katmanlar üzerindeki sinirler arasındaki ağırlık değerleri, toplama ve aktivasyon fonksiyonu yardımı ile işlenmekte ve bir ya da birden fazla çıkış sinyali üretilebilmektedir.



Şekil 3.9. Temel bir YSA'nın yapısı

Giriş katmanı veri kümesindeki örneklere ait gen ekspresyon değerlerinin YSA tarafından işlenmek üzere içeri alınması için kullanılan katmandır. Her bir YSA çağında ağa verilen örneğe ait gen ekspresyon değerleri giriş katmanındaki sinirlerin çıkış değerleri olarak kabul edilmektedir. Ara veya çıkış katmanındaki sinirlerin çıkışlarının hesaplanması için ise ilk önce toplam net değerleri hesaplanması gerekmektedir. Bu hesaplama için (3.19)'de verilen toplam fonksiyonundan yararlanılmaktadır.

$$Net_i = \sum_{j=1}^{n} w_{ij} Ç_j \tag{3.19}$$

i sinir hücresinin toplam değeri $Toplam_i$ hesaplanmak istendiğinde, n i sinirinden bir önceki katmandaki sinirlerin toplam sayısını verirken w_{ij} bu sinirler ile i siniri arasındaki ağırlıkları, $Ç_j$ ise bu sinirlerin çıkış değerlerini temsil etmektedir. Giriş katmanı haricindeki bir sinir hücresinin çıkış değeri hesaplanırken ise bir aktivasyon fonksiyonundan yararlanılması gerekmektedir. Bu çalışmada bir sinir hücresinin çıkış değerinin hesaplanması için kullanılan sigmoid fonksiyonu Denklem (3.20)'de verilmiştir.

$$Ç_i = \frac{1}{1 + e^{-(Net_i + \beta_i)}} \tag{3.20}$$

β_i i sinirinin düzeltme yani bias değerini ifade etmektedir. Her bir sinirin çıkış bilgisinin elde edilmesinden sonra çıkış katmanındaki sinirlerinin değerleri sınıflandırma sonucu olarak kabul edilmektedir. Ancak öğrenme işlemi gerçekleşmeden önce bu tahminler hatalı olacaktır. Sınıflandırma başarısının arttırılması için ilk başta rastgele belirlenen ağırlıkların güncellenmesi YSA'nın eğitilmesi gerekmektedir. Ağırlıkların güncellenmesi için çalışma içerisinde hatanın geriye yayılması (Backpropagation) yöntemi tercih edilmiştir. Bu yöntemin kullanılabilmesi için öncelikle YSA'nın çıkış sinirlerinin hatalarının hesaplanması gerekmektedir. Çıkış sinirlerinin hataları Denklem (3.21)'de verildiği gibi hesaplanmaktadır.

$$\delta_i = Ç_i(1 - Ç_i)(Y_i - Ç_i) \tag{3.21}$$

Denklemde verilen E_i i sinirinin hatasını ifade ederken, Y_i beklenen çıkış değerini, $Ç_i$ ise YSA'nın gerçekleştirmiş olduğu çıkış değerini ifade etmektedir. Bu hataların hesaplanmasından sonra her bir ağırlıktaki değişimin hesaplanması için Denklem (3.22)'den yararlanılmaktadır.

$$\Delta w_{ij}(t) = \gamma \delta_i Ç_j + \alpha \Delta w_{ij}(t-1) \tag{3.22}$$

$\Delta w_{ij}(t)$ hesaplanmak istenen çıkış katmanındaki i ve bir önceki katmandaki j sinirleri arasındaki ağırlığın değişim miktarını, $\Delta w_{ij}(t-1)$ ise aynı ağırlığın bir önceki çağdaki değişim miktarını temsil etmektedir. γ ağın eğitim katsayısını ifade ederken, α momentum katsayıdır. Tüm ağırlıklar güncellendikten sonra bir sonraki çağa geçilerek en düşük hata oranı ya da tekrar sayısına erişilene kadar aynı işlemler tekrar edilmektedir.

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

Bu kısımda geliştirilen web arayüzünün özellikleri, geliştirilen Birlik-Hibrit algoritmasının detayları, bu algoritma ile elde edilen performans değerleri, literatür karşılaştırmaları ve son olarak elde edilen bulguların biyolojik öneminden bahsedilmiştir.

4.1. Ara Yüz Geliştirilmesi

Gen seçimi mikro dizi verilerinin analizindeki en temel ve en önemli konularından biridir. Kanser ile ilişkili anlamlı genlerin etkili bir şekilde seçilmesi ileri seviyede yapılacak analizlere, teşhis tedavi ve ilaç geliştirme aşamalarına, biyolojik verilerin anlamlandırılmasına büyük katkı sağlamaktadır. Ancak gen seçim sürecinin yönetilmesi, yeni yaklaşımların geliştirilmesi önemli olduğu kadar zorlayıcı da bir konudur. Geliştirilen yaklaşımların mikro dizi verilerinin karakteristik özellikleri nedeni ile genelleştirilememesi karşılaşılan güçlüklerin başında gelmektedir. Bir mikro dizi veri kümesinin analizinde başarılı performans değerleri elde edebilen bir algoritma başka bir veri kümesi için aynı performans değerini yakalaması her zaman mümkün olmamaktadır. Bu yüzden literatürde önerilen birçok yöntem arasından en uygun yöntemin seçilmesi ve bu yöntemde kullanılan algoritmalara ait en uygun parametrelerin belirlenmesi gerekmektedir.

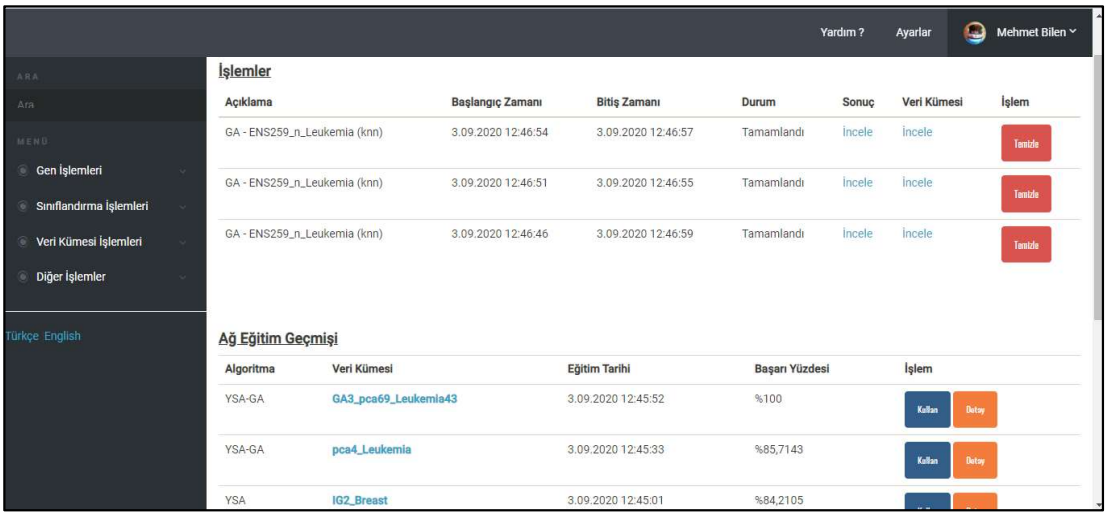
Birden fazla ve farklı türdeki algoritmaların kullanılması gerektiği durumlarda en uygun kombinasyonun belirlenmesi de araştırmacılar için gerçekten zorlayıcı bir konudur. Bu zorlukların üstesinden gelinebilmesi için yapılan çalışmada içerisinde birçok farklı yaklaşıma sahip algoritmanın bir arada kullanılabileceği, farklı kombinasyonlarla yeni yaklaşımlar elde edilebileceği bir arayüz araştırmacıların işini büyük ölçüde kolaylaştıracaktır. Bu amaçla dünyanın her yerinden biyoinformatik alanında çalışan araştırmacıların erişebileceği, içerisinde farklı çalışma prensipleri barındıran filtreleme, öznitelik çıkarma, gen seçimi, sınıflandırma algoritmalara erişim sağlayabileceği, yapılan analizlerin saklanıp paylaşılabileceği bir web arayüzü tasarlanmıştır.

Bu kısımda geliştirilen arayüzün teknik detayları, arayüzün sahip olduğu işlevler, arayüz içerisinde kodlanmış algoritmalar, kullanıcıların arayüzü kullanarak yapabileceği işlemler ve alabileceği raporların detayları verilmiştir.

4.1.1. Geliştirme ortamı ve arayüz özellikleri

Python, R ve Matlab gibi programlama dilleri ve uygulama geliştirme ortamları günümüzde birçok farklı analiz için içerisinde sezgisel algoritmaların kullanıldığı hazır metotlar barındırmaktadır. Az bir kodlama becerisi ile sınıflandırma, gen seçimi ve başka birçok analiz gerçekleştirilebilmektedir. Ancak bu diller ve uygulama geliştirme ortamları ile oluşturulan yazılımların son kullanıcının kullanımına sunulması bir hayli zordur. Geliştirilmek istenen karmaşık algoritmalarda göz önüne alındığında hazır metotlardan daha çok güçlü bir programlama diline, daha işlevsel bir uygulama geliştirme ortamına ihtiyaç duyulmaktadır. Bununla beraber geliştirilen algoritmaların web ortamında çalışabilmesi, veri tabanı ve sunucu desteği gibi ihtiyaçlar da bu tez çalışmasında kullanılacak programlama dilinin ve uygulama geliştirme ortamının seçilmesinde etkili olmuştur.

Bu tez çalışmasında oluşturulan arayüz, güçlü bir uygulama geliştirme ortamına sahip Microsoft Visual Studio kullanılarak C# dilinde kodlanmıştır. Geliştirilen ve kodlanan algoritmaların web ortamında son kullanıcılar tarafından erişilebilir olması için Asp.Net MVC yapısı kullanılmıştır. Bu sayede kullanıcıların hiçbir kurulum gereksinimi duymadan sadece internet erişimine sahip bir cihaz ile algoritmalara erişebilmeleri sağlanmıştır. Ayrıca Asp.Net MVC'in sağlamış olduğu sunucu desteği ile kullanıcıların kendi donanım kaynaklarına ihtiyaç duymadan işlemlerini sunucuda gerçekleştirebilecekleri, MVC ile uyumlu bir şekilde çalışan Microsoft SQL Server sayesinde yapmış oldukları her bir işlemin sonuçlarını saklayabilecekleri bir ortam oluşturulmuştur. Geliştirilen arayüze ait genel görünüm aynı zamanda kullanıcıların oturum açtıklarında karşılarına çıkan arayüz Şekil 4.1'de verilmiştir.



Şekil 4.1. Kullanıcı arayüzü genel görünüm

Arayüzün sol kısmında gen seçimi, sınıflandırma, veri kümesi ve diğer işlemlere ait arayüzlere ulaşılabildiği bir menü yer almaktadır. Üst kısımda ise oturum işlemleri, ayarlar ve yardım linkleri bulunmaktadır. Tarayıcı uyumluluğuna sahip (Responsive) bir yapıda tasarlanan bu arayüz sayesinde kullanıcılar farklı ekran çözünürlüğüne sahip cihazlarla (bilgisayar, tablet, cep telefonu vs.) arayüzün işlevlerini kolayca kullanabilmektedir. Ayrıca arayüz yerelleştirme (Localization) özelliği sayesinde İngilizce dil desteği de sunulmaktadır.

4.1.2. Gen filtreleme ve seçim arayüzleri

Bu kısımda Şekil 4.2'de verilen gen işlemleri menüsünün altındaki BK, FKS, WRT, Birlik gen filtreleme algoritması, GA, TBA ve normalizasyon arayüzlerinden bahsedilecektir.

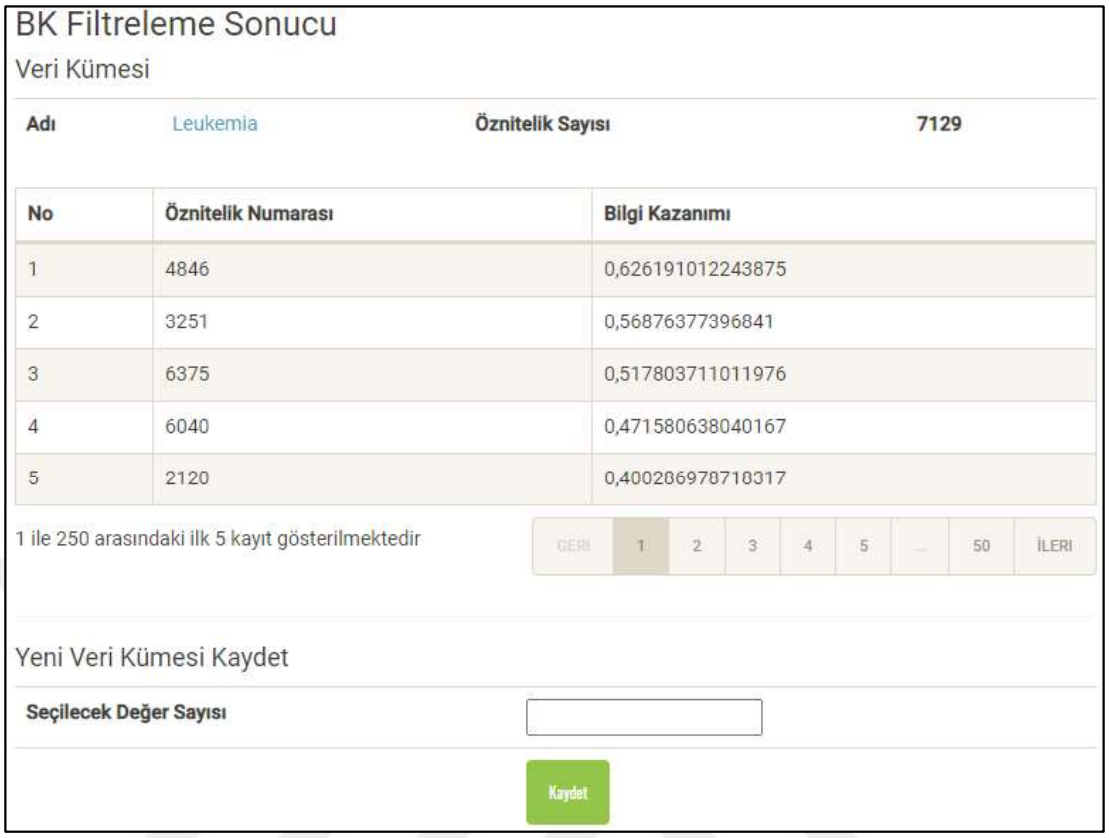
Şekil 4.2. Gen işlemleri menüsü

Bilgi kazanımı filtreleme algoritması arayüzü

Şekil 4.3'de verilen BK arayüzünün ilk adımında kullanıcılar bu algoritma ile gen puanlaması yapmak istedikleri veri kümesini seçebilmekte, seçme işlemi öncesinde bu veri kümesinin detaylarını görebilmektedir.

Şekil 4.3. BK filtreleme arayüzü

Veri kümesi seçim işleminden sonra kullanıcılar BK algoritması ile hesaplanan gen puanlarının listelendiği ve bu puanlara göre filtrelenmek istenen gen sayısının belirlenebildiği dinamik bir arayüze yönlendirilmektedir (Şekil 4.4) . Bu arayüz ile kullanıcılar her bir genin almış olduğu puanı sıralı bir şekilde inceleyerek istedikleri sayıda gen seçebilmekte ve sadece bu genlerin ekspresyon değerlerinden oluşan yeni bir veri kümesi oluşturabilmektedirler.

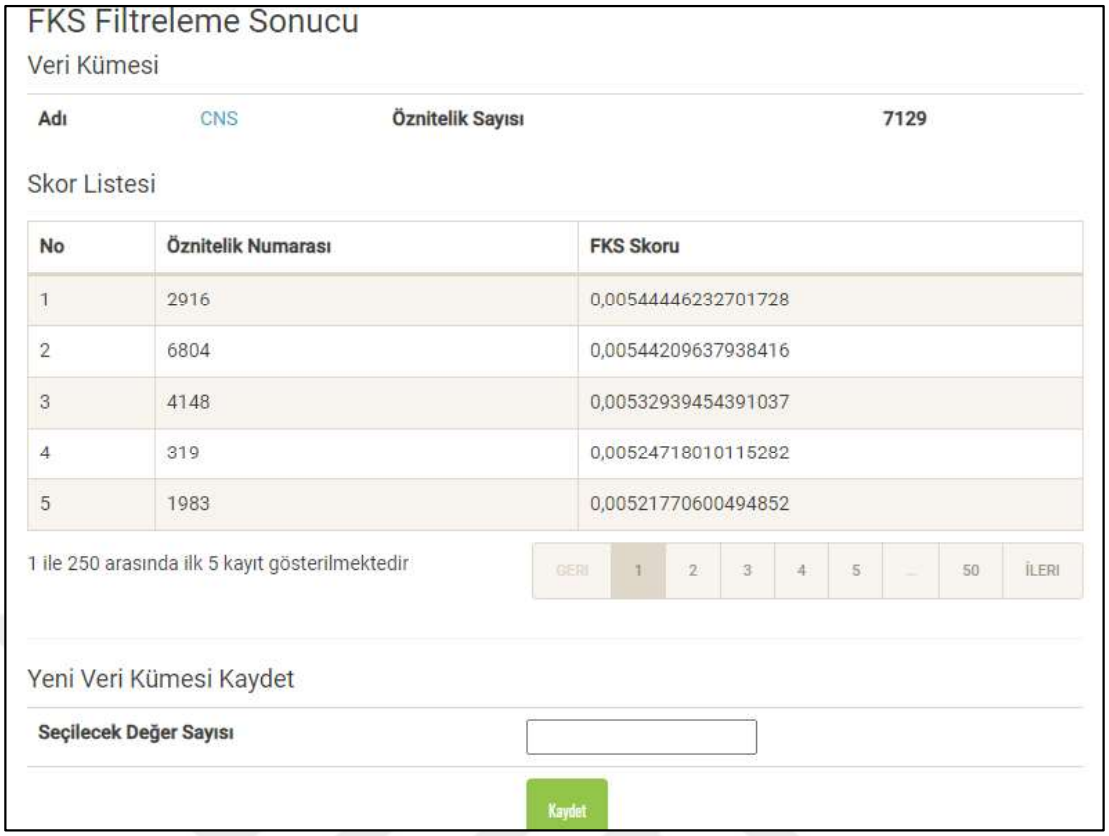


Şekil 4.4. BK sonuç arayüzü

Fisher korelasyon skoru filtreleme arayüzü

Bu arayüzde BKS filtreleme arayüzünde olduğu gibi iki adım bulunmaktadır. İlk adımda filtrelenmek istenen veri kümesinin seçimi (Şekil 4.5), ikinci adımda bu algoritma ile puanlanmış genlerin seçilebilmesine olanak veren kontroller bulunmaktadır (Şekil 4.6).

Şekil 4.5. FKS filtreleme arayüzü

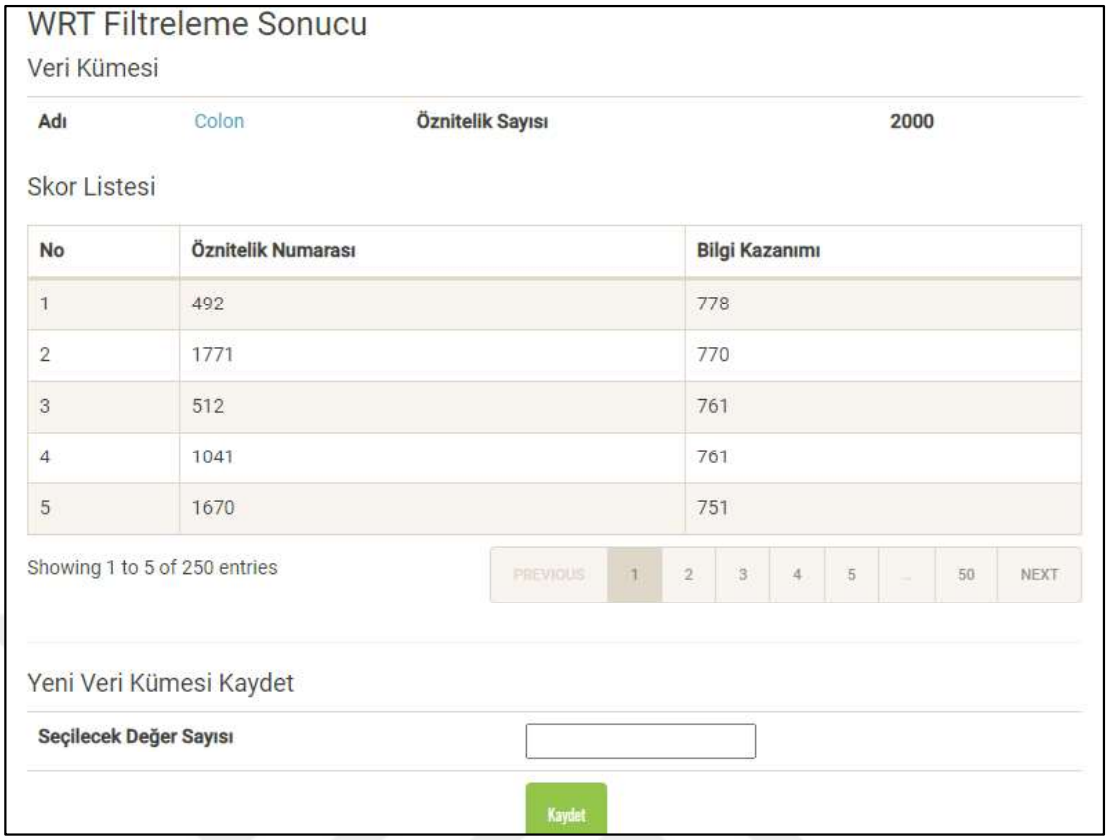


Şekil 4.6. FKS sonuç arayüzü

Wilcoxon rütbeler toplamı filtreleme arayüzü

Filtreleme ve filtreleme sonuçlarının incelenerek seçim yapıldığı arayüz sırası ile Şekil 4.7'de ve Şekil 4.8'de verilmiştir.

Şekil 4.7. WRT filtreleme arayüzü

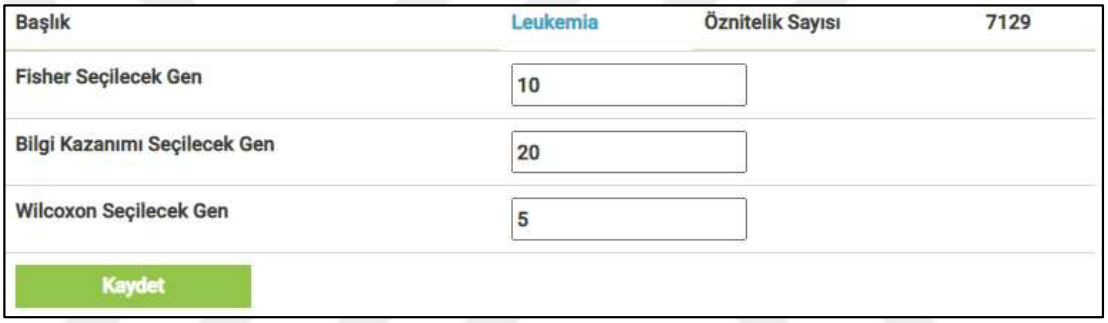


Şekil 4.8. WRT sonuç arayüzü

Birlik gen filtreleme arayüzü

Bu arayüz diğer bahsedilen üç filtreleme arayüzünün görsel ve işlevsel olarak birleştirilmesinden oluşmaktadır. İlk adımda benzer şekilde filtrelenmek istenen veri kümesi seçilmektedir. Buna ek olarak bu adımda veri kümesinin filtrelenme aşamasında hangi algoritmaların kullanılacağı tercih edilebilmektedir (Şekil 4.10). İkinci adımda seçim yapılan her bir algoritma için ayrı ayrı sonuçlar gösterilmekte olup her bir algoritmanın puanlamış olduğu genler arasından farklı sayıda gen seçimi yapılabilmektedir (Şekil 4.10).

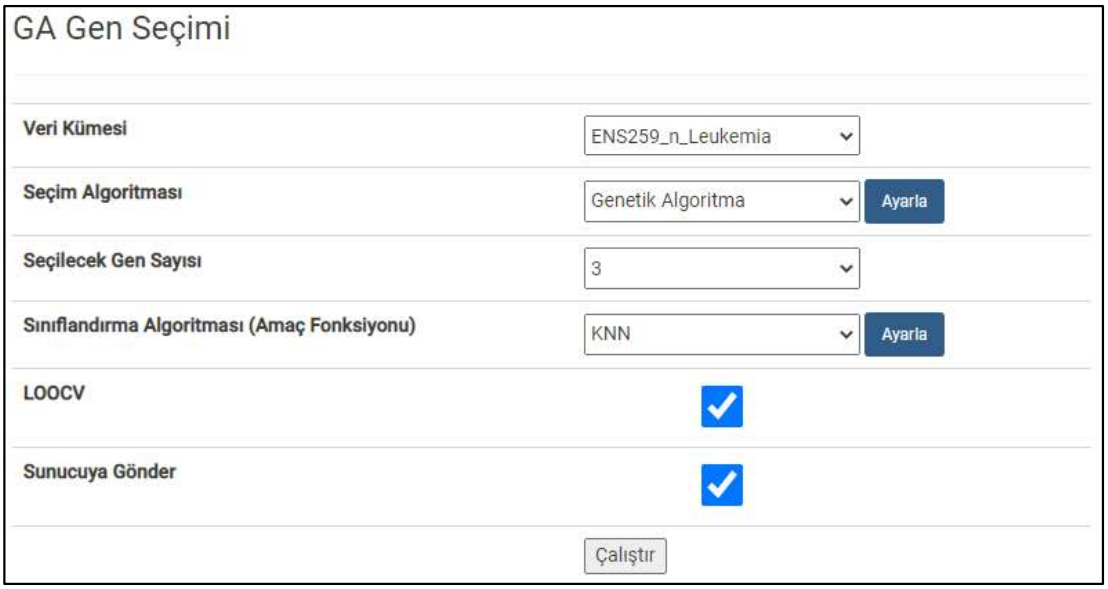
Şekil 4.9. Birlik gen filtreleme arayüzü



Şekil 4.10. Birlik algoritma sonuç ve gen sayısı belirleme arayüzü

Genetik algoritma gen seçim arayüzü

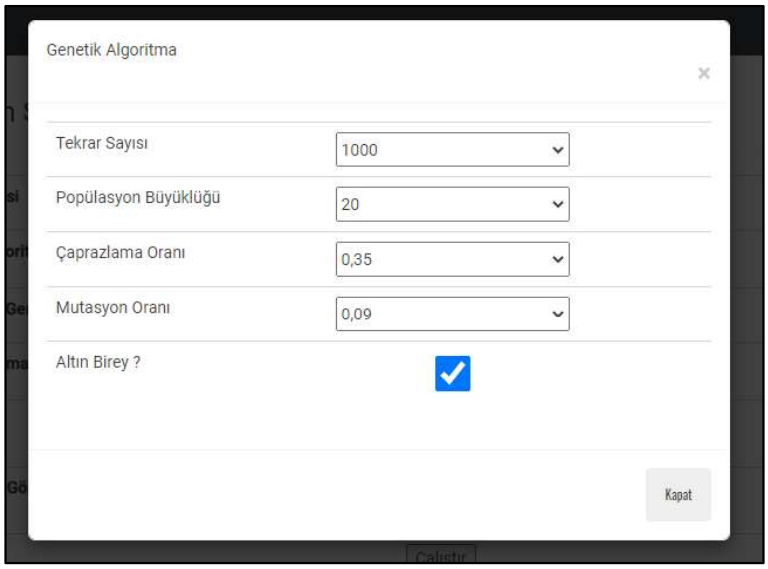
Bu arayüz diğer filtreleme algoritmalarından farklı olarak birçok seçeneğe sahiptir. Kullanıcılar bu arayüz sayesinde GA'nın birçok önemli parametresini seçebilmekte, amaç fonksiyonu için kullanılacak sınıflandırma algoritmasını ve bu sınıflandırma algoritmanın parametrelerini belirleyebilmektedir. Genetik algoritma gen seçim arayüzünün ilk adımı Şekil 4.11'de verilmiştir.



Şekil 4.11. Genetik algoritma gen seçim arayüzü

İlk arayüz üzerinde seçilecek gen sayısı, uygunluk değeri hesaplanırken amaç fonksiyonu olarak kullanılacak sınıflandırma algoritması ile beraber sınıflandırma işleminin LOOCV ile doğrulanıp doğrulanmayacağına karar verilebilmektedir. GA popülasyon fazla ve veri kümesi boyutunun büyük olduğu durumlarda gen seçim işlemi zaman almaktadır. Bu yüzden sunucuya gönder seçeneği ile arka planda gen seçim işlemi gerçekleştirilirken kullanıcılar diğer işlemlerine devam edebilmektedir.

GA'nın ve seçilen sınıflandırma algoritmasının parametreleri ilgili algoritmanın hemen yanındaki ayarla butonu ile gelen açılır pencereler ile belirlenebilmektedir. GA parametreleri için açılan pencere Şekil 4.12'de, sınıflandırma algoritması için açılan pencere Şekil 4.13'de verilmiştir.



Şekil 4.12. Genetik algoritma parametre belirlenmesi için açılan pencere

Şekil 4.13. Sınıflandırma algoritması parametre belirlenmesi için açılan pencere

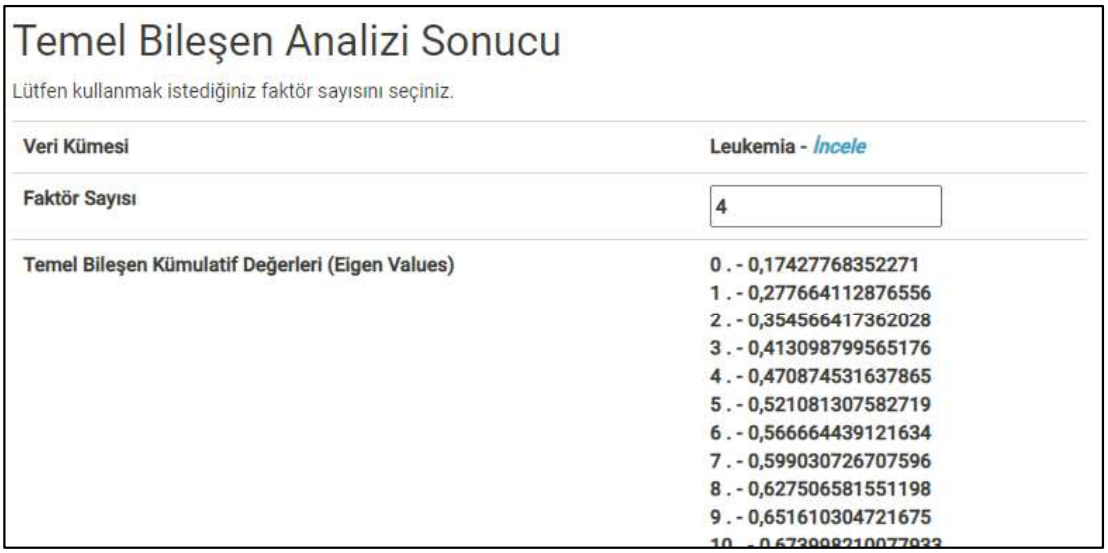
Başlangıç parametrelerinin belirlenmesinden sonra çalıştırılan genetik algoritmanın elde ettiği sonuçlar Şekil 4.14'de verilen rapor üzerinde gösterilmektedir. Bu rapor algoritmanın bulmuş olduğu yerel ve genel çözüm kümelerini hangi tekrarlarda elde ettiğini, hangi genlerin seçildiğini ve hesaplanan LOOCV değerini bir liste halinde sunmaktadır. En iyi tekrara ait bilgiler kırmızı renkte işaretlenmektedir.

Şekil 4.14. GA gen seçim raporu

Temel bileşen analiz arayüzü

Bu arayüz de diğer gen filtreleme arayüzlerde de olduğu gibi ilk adımda analizi yapılmak istenen veri kümesi seçilebilmekte (Şekil 4.15), hesapla komutundan sonra ekrana gelen arayüzde ise gen filtreleme işlemi yerine temel bileşen analizi ile hesaplanmış öz bileşen değerleri listelenmektedir (Şekil 4.16). Kullanıcı bu öz bileşen değerlerinin kümülatif toplamlarına bakarak istediği sayıda faktörü yeni veri kümesi oluşturmak için kullanabilmektedir.

Şekil 4.15. Temel bileşen analiz arayüzü



Şekil 4.16. Temel bileşen analiz sonucu ve öz bileşen listesi

Normalizasyon arayüzü

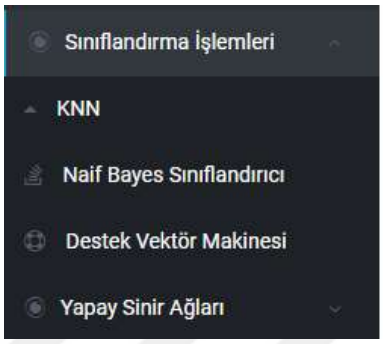
Bu arayüzü kullanarak kullanıcılar, birçok algoritmanın ihtiyaç duyduğu ön işlem adımlarından biri olan, özellikle dengesiz dağılıma sahip olan veri kümelerinin normalizyon işlemini gerçekleştirebilmektedirler. Normalizasyon arayüzü Şekil 4.17'de verilmiştir. Arayüz üzerinden veri kümesi ve değerlerin kaç ondalık basamağa sahip olacağının seçilmesinden sonra hesapla butonu yardımı ile veriler normalize edilerek yeni bir veri kümesi oluşturulmaktadır.

Şekil 4.17. Normalizasyon arayüzü

4.1.3. Sınıflandırma arayüzleri

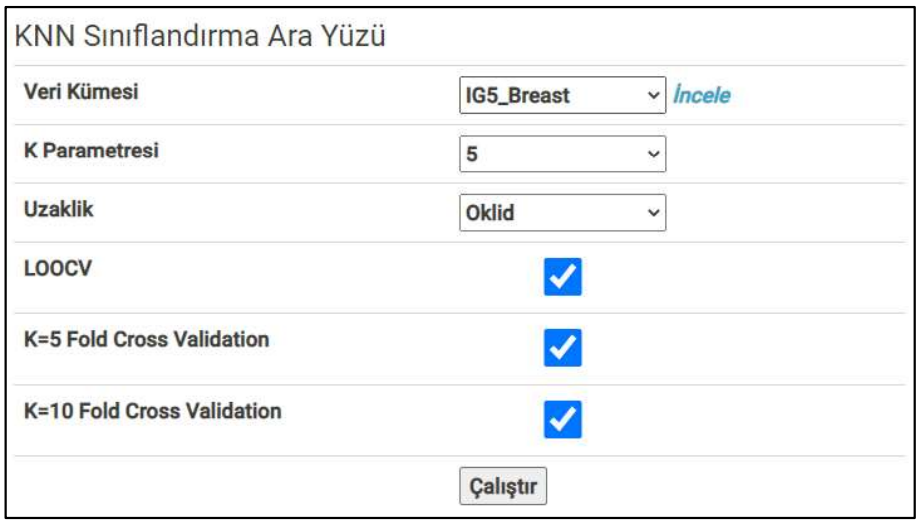
Bu kısımda Şekil 4.18'de ki menü üzerinden erişilen sınıflandırma arayüzlerinden bahsedilecektir. Sınıflandırma arayüzleri de gen seçim

arayüzlerinde olduğu gibi genel olarak iki arayüzden oluşmaktadır. İlk arayüzde sınıflandırma işlemi yapılmak istenen veri kümesi, ilgili sınıflandırma algoritmasının başlangıç parametreler kullanıcılar tarafından seçilebilmektedir. İkinci arayüzde seçilen parametreler sonucunda gerçekleştirilen sınıflandırma sonucuna ait tüm sonuçlar detaylı bir şekilde raporlandırılmaktadır.

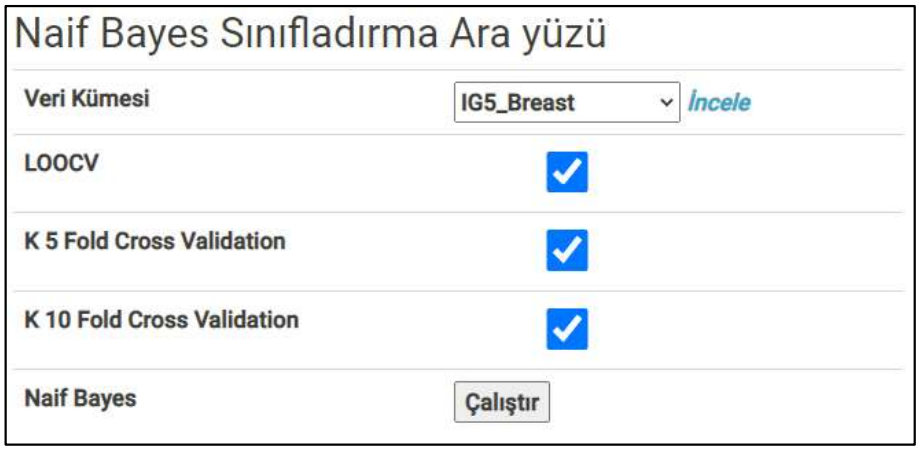


Şekil 4.18. Sınıflandırma işlemleri menüsü

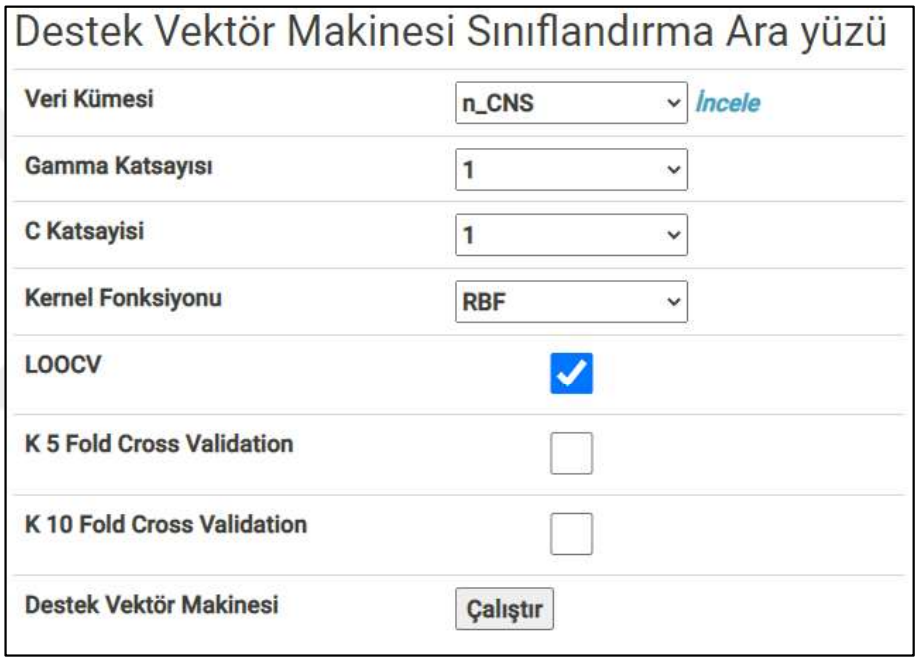
KNN, NB ve DVM arayüzlerinde kullanıcılar sınıflandırılmak istenen veri kümesini, ilgili algoritmanın başlangıç parametrelerini ve sınıflandırma sonuçlarının doğrulanmasında hangi çapraz doğrulama yöntemlerinin kullanılacağını belirleyebilmektedirler. KNN, NB ve DVM sınıflandırma işlemlerine ait parametre seçim arayüzleri sırası ile Şekil 4.19, Şekil 4.20, Şekil 4.21'de verilmiştir.



Şekil 4.19. KNN sınıflandırma arayüzü



Şekil 4.20. Naif Bayes sınıflandırma arayüzü



Şekil 4.21. DVM sınıflandırma arayüzü

Seçilen veri kümesi ve belirlenen parametreler kullanılarak gerçekleştirilen sınıflandırma işlemi sonucunda elde edilen bulgular, özet, performans değerleri, tüm örneklerin sınıflandırma sonuçları isimli raporlar ve ek grafiklerle sunulmaktadır. Sınıflandırma işlemleri sonucunda elde edilen başarılı sınıflandırılan örnek sayısı, başarı yüzdesi, AUC değeri ve çapraz doğrulama değerlerinin kullanıcılar tarafından incelenebildiği özet rapor Şekil 4.22'de verilmiştir.

Özet

Doğrulama	Örnek Sayısı	Başarılı Sınıflandırma	Başarı Yüzdesi	AUC Değeri	Eğitim için Geçen Süre & Test için Geçen Süre
LOOCV	97	76	**78,3505**	0,78	10 (ms)
Doğrulama	**Toplam Test**	**En Başarılı Test**	**En Başarısız Test**	**Ortalama Başarı**	
K5Fold	5	84,2105	65	74,3158	
K10Fold	10	100	44,4444	76	

Şekil 4.22. Özet rapor

Şekil 4.23'de sınıflandırma sonucunda elde edilen performans değerleri raporu verilmiştir. Bu rapor sayesinde kullanıcılar her bir sınıf parametresi için hesaplanan belirlilik duyarlılık ve doğruluk değerlerine ve bu değerlerin hesaplanmasında kullanılan Doğru Pozitif, Doğru Negatif, Yanlış Pozitif, Yanlış Negatif değerlerini inceleyebilmektedir.

Performans Değerleri

Sınıf	Belirlilik	Duyarlılık	Doğruluk	Detay
relapse	0,8431	0,7174	0,7835	Detay
non-relapse	0,7174	0,8431	0,7835	Detay

Şekil 4.23. Performans değerleri raporu

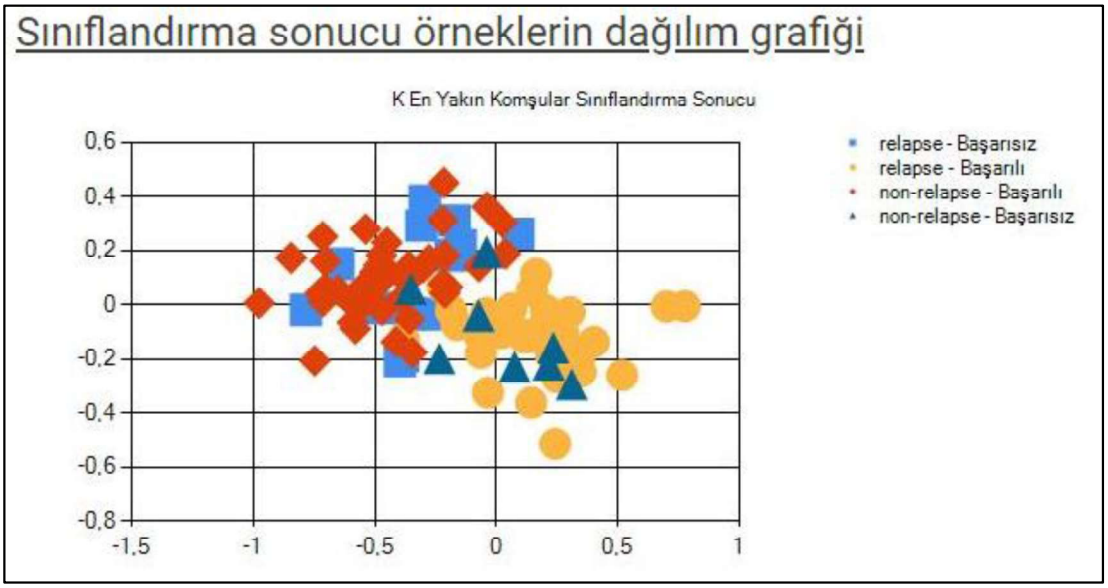
Her bir örneğin ait olduğu gerçek sınıf parametresi, tahmin edilen sınıf, ilgili örneğin gen ekspresyon değerleri ve bu örneklerin KNN için en yakın komşularına ait hesaplanan uzaklık, NB için olasılık değerleri Şekil 4.24'de verilen sınıflandırma detay raporu ile kullanıcılar tarafından incelenebilmektedir.

Her Bir Örneğin Sınıflandırma Detayı

Sıra	Ait Olduğu Sınıf	Sınıflandırma Sonucu	Durum	Öznitelik - 1	Öznitelik - 2	Öznitelik - 3	Öznitelik - 4	Öznitelik - 5	
1	relapse	non-relapse	Başarısız	0,091	0,263	-0,175	-0,167	0,251	Uzaklıkları Goster / Gizle
2	relapse	non-relapse	Başarısız	-0,293	-0,035	-0,167	-0,085	0,096	Uzaklıkları Goster / Gizle
3	relapse	relapse	Başarılı	0,348	-0,247	-0,126	-0,035	0,201	Uzaklıkları Goster / Gizle
4	relapse	relapse	Başarılı	0,266	-0,19	0,194	-0,199	-0,08	Uzaklıkları Goster / Gizle
5	relapse	relapse	Başarılı	0,225	-0,184	0,055	0,011	0,081	Uzaklıkları Goster / Gizle

Şekil 4.24. KNN sınıflandırma detayları (her bir örnek için)

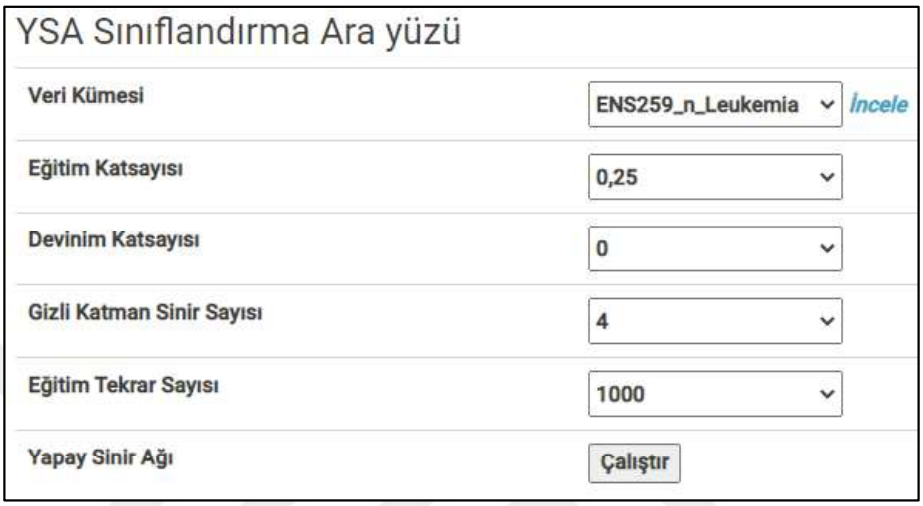
Şekil 4.25'de her bir örneğin sınıflandırma sonucuna göre güncellenen, kullanıcıların başarılı ve başarısız tahminleri inceleyebildiği dağılım grafiği verilmiştir.



Şekil 4.25. Dağılım grafiği

Yapay sinir ağlarının tekrarlı yapısı nedeni ile diğer sınıflandırma algoritmalarındaki raporlara ek olarak farklı bulgular da elde edilmektedir. Bu

yüzden üretilen raporlar farklılık göstermektedir. Ayrıca diğer sınıflandırma algoritmalarında verilen raporlara ek olarak her tekrarda ki başarı oranının, yerel çözümlerin ve en iyi çözümlerin incelenebilmesi için ayrıca bir eğitim grafik raporu oluşturulmaktadır. YSA algoritmasının sınıflandırma arayüzü Şekil 4.26,'da verilmiştir.

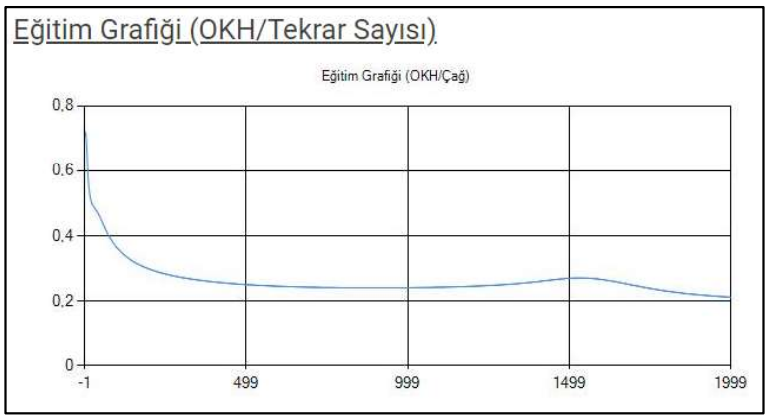


Şekil 4.26. YSA sınıflandırma arayüzü

YSA algoritmaları ile gerçekleştirilen sınıflandırma işlemlerinden sonra oluşturulan özet rapor Şekil 4.27'de verilmiştir. Bu rapor ile kullanıcılar en düşük ortalama kare hata (OKH) ve bu OKH'nın hangi tekrarda gerçekleştiği, eğitim ve test için algoritmanın harcadığı süre gibi bilgileri inceleyebilmektedir.

Şekil 4.27. YSA eğitim ve sınıflandırma özet raporu

Diğer sınıflandırma algoritmalarından farklı olarak sadece YSA için üretilen eğitim grafiği Şekil 4.28'de verilmiştir.



Şekil 4.28. YSA eğitim grafiği

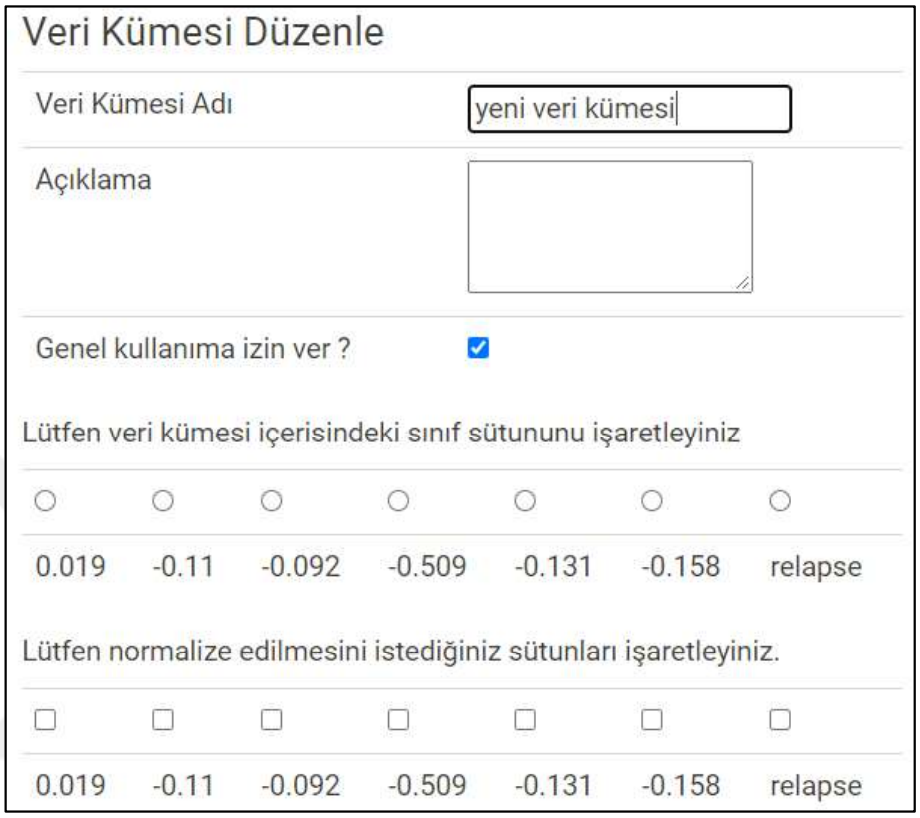
4.1.4. Veri kümesi işlem arayüzleri

Geliştirilen web tabanlı araçta, kullanıcıların kendi veri kümelerini sisteme yükleyerek gen filtreleme, gen seçimi ve sınıflandırma işlemlerini yapabilmeleri için gerekli olan arayüzler tasarlanmıştır. Şekil 4.29'de verilen veri kümesi yükleme arayüzünün ilk adımında kullanıcılar kendi bilgisayarlarındaki dosyaları seçerek yükleme işlemini başlatabilmektedirler.

Şekil 4.29. Veri kümesi yükleme arayüzü

Veri kümesi yükleme işlemi başlatıldıktan sonra sisteme dâhil edilen veri kümesi için gerekli olan bilgilerin girilmesi, sınıf parametresinin hangi öznitelik

ile temsil edileceğinin belirlenmesi, normalize edilmek istenen öznitelikler varsa bunların seçilebilmesi için gereken ayarları içeren ikinci adımda ki arayüz Şekil 4.30'da gösterilmiştir.

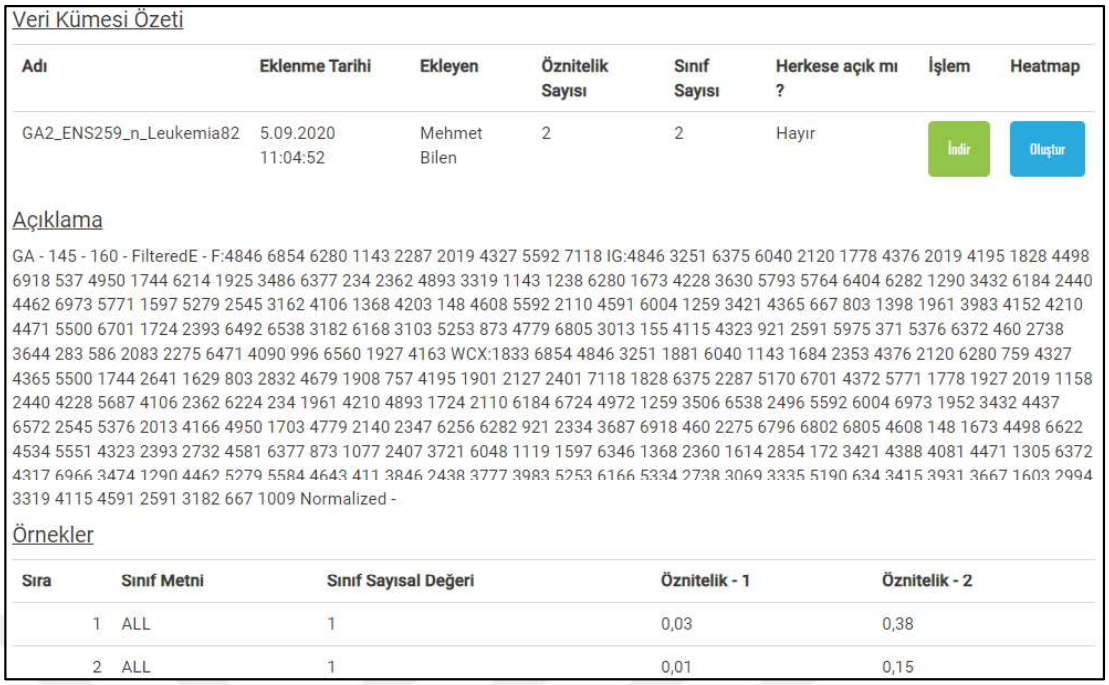


Şekil 4.30. Veri kümesi yükleme ayarları

Kullanıcıların kendi eklemiş oldukları veri kümelerinin ve bunlara ait bilgilerin bir liste halinde inceleyebilmesi için tasarlanmış veri kümesi listeleme arayüzü Şekil 4.31.'de sunulmuştur.

Şekil 4.31. Veri kümesi listeleme arayüzü

Kullanıcılar bir veri kümesine ait detaylı bilgiye ulaşmak istediklerinde Şekil 4.32'de verilen veri kümesi detay arayüzünü kullanabilmektedirler. Bu arayüz ile kullanıcılar veri kümesinin adı, açıklama kısmında daha önce yapılan gen filtreleme ve gen seçim işlemlerinin detayları, gen-öznitelik sayısı, sınıf sayısı, bu veri kümesini sisteme ekleyen kişi, eklenme tarihi ve veri kümesi içerisindeki tüm örnekleri öznitelikleri ile birlikte görebilmektedirler. Bu arayüz içerisinde ayrıca ilgili veri kümesinin indirilebilmesi veya sıcaklık haritasının oluşturulabilmesi için gereken kontroller de bulunmaktadır.



Şekil 4.32. Veri kümesi detay arayüzü

4.1.5. Diğer arayüzler

Web tabanlı araç içerisinde kullanıcıların geçmiş işlemlerini takip edebilmeleri ve bu işlemlerden elde edilen bulguları farklı işlemlerde tekrar kullanabilmeleri için gen seçim geçmişi ve ağ eğitim geçmişi arayüzleri geliştirilmiştir. Buna ek olarak kullanıcıların ihtiyaç duyması halinde gen dizilimlerini hizalayarak puanlama yapabilecekleri Needleman-Wunsch algoritmasına ait bir arayüz de bulunmaktadır.

Gen seçim geçmişi

Bu arayüz kullanıcılara gen seçimini sonucunda elde edilen bulgulara ve oluşturulan yeni veri kümesine erişebilecekleri bir ortam sunmaktadır (Şekil 4.33). Ayrıca yapılan analizin hangi algoritmalar ve parametreler ile gerçekleştiği de bu kısımda görüntülenebilmektedir.

Açıklama	Başlangıç Zamanı	Bitiş Zamanı	Durum	Sonuç	Veri Kümesi	İşlem
GA - ENS259_n_Leukemia (bayes)	5.09.2020 11:02:54	5.09.2020 11:04:52	Tamamlandı	İncele	İncele	Temizle
GA - ENS259_n_Leukemia (bayes)	5.09.2020 11:01:16	5.09.2020 11:02:16	Tamamlandı	İncele	İncele	Temizle
GA - ENS259_n_Leukemia (knn)	3.09.2020 18:04:12	3.09.2020 18:04:16	Tamamlandı	İncele	İncele	Temizle
GA - ENS259_n_Leukemia (knn)	3.09.2020 15:22:35	3.09.2020 15:22:37	Tamamlandı	İncele	İncele	Temizle
GA - ENS259_n_Leukemia (knn)	3.09.2020 12:46:54	3.09.2020 12:46:57	Tamamlandı	İncele	İncele	Temizle
GA - ENS259_n_Leukemia (knn)	3.09.2020 12:46:51	3.09.2020 12:46:55	Tamamlandı	İncele	İncele	Temizle
GA - ENS259_n_Leukemia (knn)	3.09.2020 12:46:46	3.09.2020 12:46:59	Tamamlandı	İncele	İncele	Temizle

Şekil 4.33. Gen seçim geçmişi

Ağ eğitim geçmişi

YSA ile yapılan eğitim işlemleri sonucunda elde edilen değerlerin kullanıcılar tarafından saklanabilmesi daha sonra eğitilmiş olan YSA'ların başka problemin çözümünde kullanılabilmesi için geliştirilen bu arayüz Şekil 4.34'de verilmiştir. Şekil 4.35'de ise bir ağ geçmişinin özeti verilmiştir.

Ağ Eğitim Geçmişi

Algoritma	Veri Kümesi	Eğitim Tarihi	Başarı Yüzdesi	İşlem
YSA-GA	GA3_pca69_Leukemia43	3.09.2020 12:45:52	%100	Kullan Detay
YSA-GA	pca4_Leukemia	3.09.2020 12:45:33	%85,7143	Kullan Detay
YSA	IG2_Breast	3.09.2020 12:45:01	%84,2105	Kullan Detay

Şekil 4.34. Ağ eğitim geçmişi

Şekil 4.35. Ağ eğitim geçmiş detayı

Needleman-Wunsch analiz arayüzü

Sekans analizi ile hizalanmak istenen gen dizilimleri Needleman-Wunsch algoritması ile bu arayüz üzerinden analiz edilebilmektedir. Analizde kullanılan ödül, ceza ve boşluk puanları arayüz üzerinden belirlenebilmektedir. Analiz ayarlarının yapıldığı kısım Şekil 4.36'de verilirken analiz sonucunda elde edilen raporların sunulması için geliştirilen sonuç arayüzü Şekil 4.37'da gösterilmiştir.

Şekil 4.36. Needleman-Wunsch analiz arayüzü

	-	-	A	G	T	C	A	G	T	C
-	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-1	1	0	-1	-2	-3	-4	-5	-6
G	-2	0	2	1	0	-1	-2	-3	-4
G	-3	-1	1	0	-1	-2	0	-1	-2
T	-4	-2	0	2	1	0	-1	1	0
C	-5	-3	-1	1	3	2	1	0	2
A	-6	-4	-2	0	2	4	3	2	1
G	-7	-5	-3	-1	1	3	5	4	3
T	-8	-6	-4	-2	0	2	4	6	5
C	-9	-7	-5	-3	-1	1	3	5	7

Şekil 4.37. Needleman-Wunsch analiz sonuç arayüzü

4.2. Yeni bir Birlik-Hibrit Algoritma Geliştirilmesi

Yapılan literatür araştırmasında mikro dizi verilerinin analizinde araştırmacıların veri boyutunun karmaşıklığı, örnek miktarının azlığı, sınıflandırma başarısının yetersizliği veya seçilen gen sayısının fazlalığı gibi birçok problemle karşılaştığı görülmektedir. Bu zorlukların üstesinden gelinebilmesi için geliştirilen arayüzünde sağlamış olduğu katkı ile birçok farklı algoritma ve yaklaşım denenmiştir. Elde edilen bulgular neticesinde yeni bir algoritma geliştirilerek literatürdeki yaklaşımlara yeni bir öneri getirilmiştir.

Geliştirilen algoritma temel olarak iki adımdan oluşan Birlik-Hibrit bir gen seçim yaklaşımı sunmaktadır. İlk adımda karmaşık veri boyutunun düşürülmesi için FKS, WRS, BK filtreleme algoritmaları geliştirilen birlik yaklaşımı sayesinde ortak bir gen filtrelemesi gerçekleştirirken, ikinci adımda güçlendirilmiş Genetik Algoritma detaylı bir gen seçimi için kullanılmaktadır. Ayrıca GA'nın seçmiş olduğu genlerin her bir tekrar da uygunluk değerinin hesaplanması için birlik bir yapıda çalışan KNN, Bayes ve DVM algoritmalardan oluşan bir amaç fonksiyonu geliştirilmiştir. Geliştirilen algoritmanın sözde kodu (Pseudo Code) Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Geliştirilen hibrit ve birlik algoritmanın sözde kodu

1:	**Filtreleme**
2:	Tüm genleri FKS, BK ve WRT algoritmaları ile ayrı ayrı puanla
3:	Her bir algoritma için en yüksek puanlı genleri kullanarak gen-alt kümeleri oluştur
4:	KNN ile gen-alt kümelerinin LOOCV değerlerini hesapla
5:	Her bir algoritma için en başarılı değerlere sahip gen-alt kümesini seç
6:	Her bir genin önem derecesini hesapla
7:	Seçilen gen-alt kümelerini birleştirerek yeni bir veri kümesi oluştur
8:	**Uyarlama**
9:	Filtrelenen genleri çözüm uzayı (gen havuzu) olarak kullan
10:	Genlerin(MD) sıra bilgilerini genlerin(GA) kodlanmasında kullan
11	N_g (seçilecek gen sayısı), N_c (popülasyon büyüklüğü) olarak belirle
12:	Kromozomları gen havuzu içerisinden N_g sayıda seçilen genler ile oluştur.
13:	N_c adet kromozomdan oluşan bir popülasyon oluştur
14:	**Uygunluk Değerinin Hesaplanması**
15:	**Tekrar et**
16:	**Popülasyondaki her bir C kromozomu için**
17:	Orijinal veri kümesinden tüm örnekleri getir
18:	Kromozomdaki gen numaralarını kullanarak örnekleri yeniden oluştur
19:	**Tekrar et**

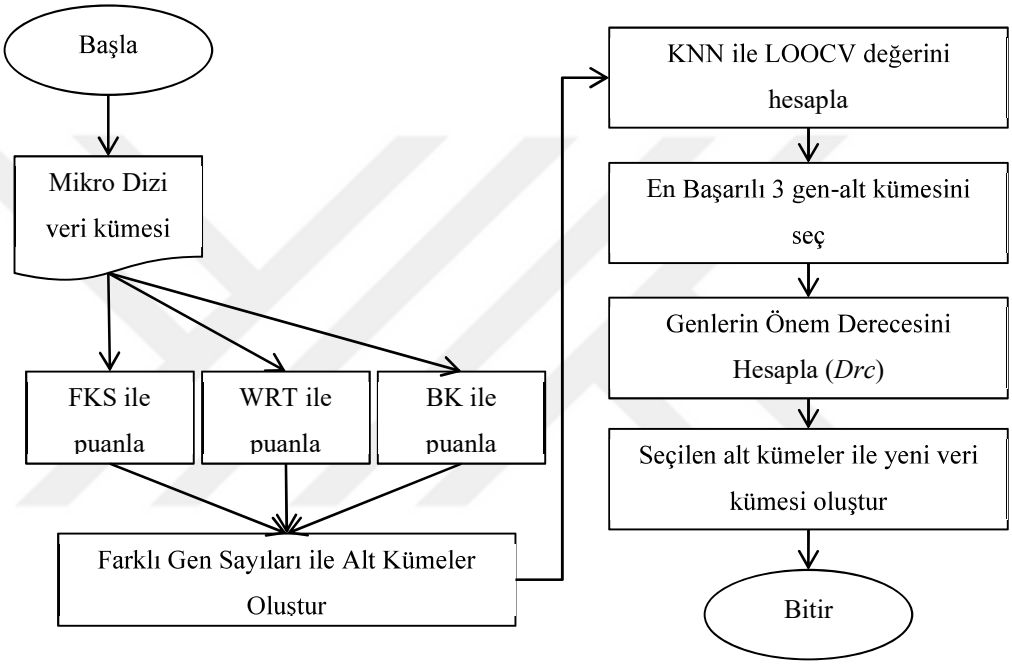
20: **Veri kümesi içerisindeki her bir S örneği için**
21: S örneğini test, geri kalan örnekleri eğitim kümesi olarak ayır
22: S örneğinin sınıfını üç algoritma ile ayrı ayrı tahmin et (KNN, SVM, Bayes)
23: Her bir algoritma için LOOCV değeri hesapla
24: Birlik sınıflandırıcı ile LOOCV değerlerini ağırlık olarak kullanarak yeni sınıf tahmini gerçekleştir
25: C kromozomunun uygunluk değerini Denklem (4.5)'e göre hesapla
26: **Seçilim ve Çaprazlama**
27: Rulet tekerleği yöntemini kullanarak popülasyon içerisinden iki adet kromozom seç, P_1, P_2
28: **Tekrar et**
29: **P_1 kromozomu içerisindeki her bir X geni için**
30: 0 ve 1 arasında rastgele c_1 değeri üret
31: **Eğer** c_1>CR (Çaprazlama Katsayısı) **İse** X genini, P2 içerisindeki aynı sıradaki gen ile değiştir.
32: **Mutasyon**
33: **Tekrar et**
34 **Her bir C kromozomu içerisindeki her bir X geni için**
35: 0 ve 1 arasında rastgele m_1 değerini üret
36: **Eğer** m_1<M (Mutasyon katsayısı)
37: **İse**
38: Gen havuzundan rastgele bir gen seç, X_r
39: 0 ve 1 arasında rastgele m_2 değerlerini üret
40: **Eğer** m_2<I(X_r) **İse** X genini X_r geni ile değiştir **Değilse** git 38
41: **Değilse** git 34.
42: **Değerlendirme**
43: Her bir kromozom için uygunluk değerini 14. satırdaki gibi hesapla
44: En başarılı N_c adet kromozomu bir sonraki tekrar/kuşak için seç
45: **Eğer** uyguluk değeri veya tekrar sayısı istenen değere ulaşmış **İse** Bitir **Değilse** git 26

Geliştirilen Birlik-Hibrit algoritmanın gen filtreleme ve gen seçim aşamalarında kullanılan yaklaşımlar, hesaplamalar ve yöntemler alt başlıklarda sırası ile açıklanmıştır.

4.2.1. Birlik gen filtreleme

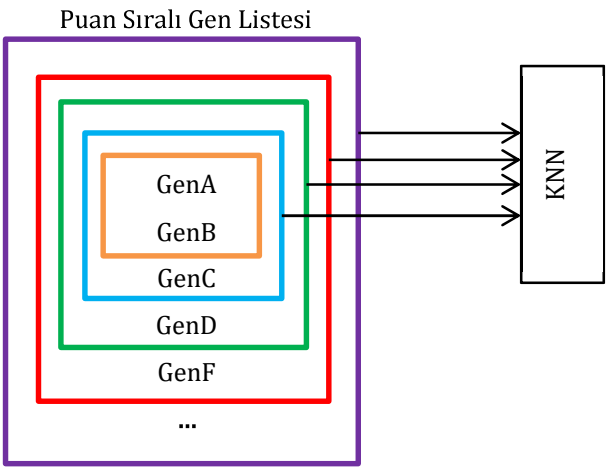
Mikro dizi verilerinin binlerce genin ekspresyon seviyesinin aynı anda incelenmesine olanak vermesi büyük bir avantaj sağlasa da karmaşık ve büyük veri boyutu, veri içerisindeki anlamsız olan genlerin oluşturmuş olduğu gürültü bu verilerin analiz edilmeden önce ön işlemlerden geçirilmesini zorunlu kılmaktadır. İstatistiksel olarak anlamlı genlerin ön işlem adımında belirlenerek diğer genlerin temizlenmesi literatürde yaygın olarak kullanılan ön işlem adımlarından biridir. Ancak veri örnek sayısındaki yetersizlik tek bir istatistiksel hesaplamaya bağımlı kalındığında anlamlı genlerin filtreleme aşamasında elenebilmesine neden olmaktadır. Bir filtreleme yaklaşımında seçilen gen başka bir filtreleme yaklaşımında kanser ile ilişkili ve sınıflandırma

başarısına olumlu katkısı olmasına rağmen elenebilmektedir. Bu durumda ortaya çıkan fakir gen havuzu ikinci adımda çalışan GA'nın performansını düşürmektedir. Bu zorluğun üstesinden gelinebilmesi için FKS, WRT ve BK algoritmalarından oluşan farklı istatistiksel yaklaşımların bir arada kullanıldığı bir birlik algoritma geliştirilmiştir. Bu sayede daha anlamlı genlerin filtrelenebilmesi ve zengin bir gen havuzu elde edilmesi amaçlanmıştır. Birlik gen filtreleme algoritmasının akış diyagramı Şekil 4.38'de verilmiştir.



Şekil 4.38. Birlik algoritması akış diyagramı

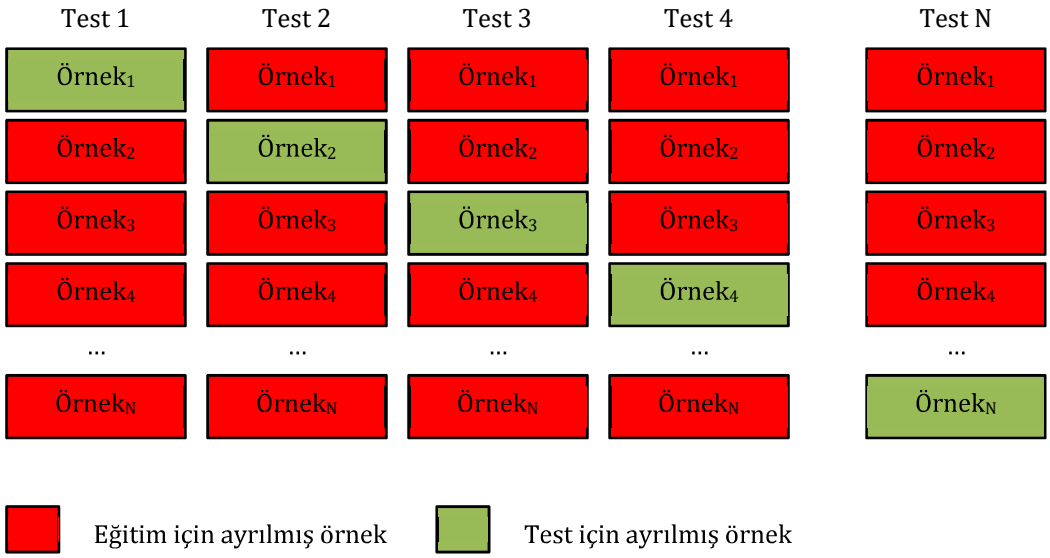
Geliştirilen birlik gen filtreleme algoritması ilk önce tüm genleri FKS, WRT ve BK algoritmaları ile ayrı ayrı puanlamaktadır. Bunun sonucunda üç algoritma için ayrı ayrı gen-puan listesi oluşmaktadır. Bu liste genlerin bireysel olarak sahip oldukları istatistiksel bilgiyi göstermektedir. Bu bilgiye ek olarak bu genlerin sınıflandırma başarıları da hesaplanmak istenmektedir. Bu yüzden bu listelerdeki en başarılı genlerden farklı sayıda gen barındıran alt kümeler oluşturulmaktadır (Şekil 4.39). Oluşturulan bu gen-alt kümelerinin sahip olduğu gen ekspresyon değerleri KNN algoritması kullanılarak LOOCV çapraz doğrulama yöntemi ile başarısı hesaplanmaktadır.



Şekil 4.39. Gen alt kümelerinin oluşturulması

Veri kümelerinin belli bir kısmının eğitim için, belli bir kısmının test için ayrılarak yapılan sınıflandırma işlemlerinden elde edilen sonuçlar her zaman güvenli olmayabilmektedir. Çünkü sınıflandırma modeli eğitim ve test verisini ezberleyerek başarılı sonuçlar elde ediyor gibi gözükse de daha önce görmediği farklı örneklerle karşılaşması durumunda başarısız olabilmektedir. Bu nedenle gen alt kümelerinin sahip olduğu genlere ait ekspresyon değerlerinin sınıflandırma başarısının LOOCV yöntemi ile doğrulanmaktadır.

Bu doğrulama yönteminde veri kümesinin iki parçaya ayrılarak eğitim ve test için tek bir defa kullanılması yerine her bir örnek ayrı ayrı test edilmektedir. Her bir örnek sırayla test işlemi için ayrılırken diğer örnekler eğitim için kullanılmaktadır (Şekil 4.40). Tüm örneklerin ayrı ayrı test edilmesi sonucunda doğrulama işlemi tamamlanmış olmaktadır. Bu testlerden elde edilen sınıflandırma başarıları, Denklem (4.1)'de verilen hesaplama yöntemi ile LOOCV çapraz doğrulama değerinin bulunması için kullanılmaktadır.



Şekil 4.40. LOOCV çapraz doğrulama yöntemi

$$LOOCV_a = \frac{1}{N}\sum_{i=1}^{N} I(a_i = \dot{a}_i) \tag{4.1}$$

Denklem verilen a LOOCV değeri hesaplanmak istenen gen alt kümesinin gerçek sınıf bilgilerini, $\dot{a}$ ise test sonucunda tahmin edilen sınıf bilgilerini göstermektedir. N, örnek sayısını ve bu durumda yapılan toplam test sayısını ifade etmektedir. I ise içerisindeki ifadenin sonucunun doğru olması durumda 1 yanlış olması durumunda 0 sonucunu üreten bir fonksiyondur.

Her bir gen alt kümesinin LOOCV değerinin hesaplanmasından sonra her bir listedeki en başarılı gen alt kümesi seçilerek yeni bir veri kümesinde birleştirilmektedir. Ancak en başarılı LOOCV değerleri her bir filtreleme algoritması için farklı sayıda gen ile elde edilmesi durumunda orantısız bir dağılım ortaya çıkmaktadır. Bir filtreleme algoritması daha az gen sayısı ile yüksek başarı sağlarken diğer bir algoritma çok daha fazla gen sayısı ile aynı başarıya ulaşabilmektedir. İkinci adımda geliştirilen genetik algoritmanın mutasyon operatörünün bu dengesiz dağılımı genlerin önem derecesine göre dikkate alınması istenmekte ve böylelikle genetik algoritmanın olası çözümleri araştırma becerisinin artırılması istenmektedir. Bununla beraber filtrelenen

genlerin filtreleme puanlarına ait bilgininde kaybolmaması ve GA tarafından dikkate alınması istenmektedir. Bu bilgiler doğrultusunda filtrelenen ve sınıflandırma başarısı kullanılarak, genlerin önem derecesinin bulunabilmesi için yeni bir hesaplama yöntemi önerilmiştir (Denklem (4.2)).

$$Drc(x_i) = \left(1 - \frac{n_x}{n_d}\right) . \frac{LOOCV_x}{R_{x_i}} \tag{4.2}$$

Denklemde verilen X seçilen gen-alt kümesini temsil ederken x_i ise bu alt kümedeki i. geni ifade etmektedir. Sırası ile n_d ve n_x, orijinal veri kümesinin toplam gen sayısını ve X gen-alt kümesi için seçilen gen sayısını temsil etmektedir. R_{x_i} ise i geninin filtreleme algoritmasındaki puanlamada bulunmuş olduğu sırayı ifade etmektedir.

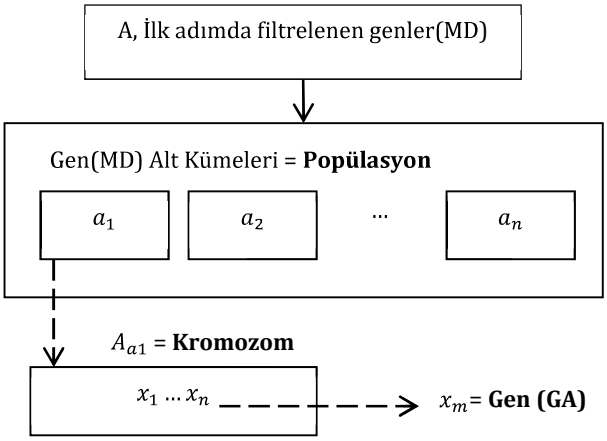
Son olarak veri kümesi içerisinden temizlenen, istatistiksel olarak anlam taşımayan ve sınıflandırma başarısı düşük genler temizlenerek filtrelenmiş yeni veri kümesi daha kapsamlı bir gen seçim işlemi için GA'nın girişine aktarılmaktadır.

4.2.2. Güçlendirilmiş Genetik Algoritma

Bu başlıkta filtreleme aşamasından sonra yeni elde edilen veri kümesi içerisinden daha detaylı bir gen seçimi yapılabilmesi için geliştirilen güçlendirilmiş genetik algoritmanın probleme uyarlanması, amaç fonksiyonu, seçilim, çaprazlama ve geliştirilen mutasyon işlevleri sırası ile açıklanmıştır. Bu kısımda son olarak GA'nın durdurma kriteri için yapılan değerlendirmeye yer verilerek verilmiştir.

4.2.2.1. Uyarlama

Genetik algoritmanın probleme uyarlanması, mikro dizi verilerinin genlerine ait sıra numaralarının kromozomlar içerisindeki genlere aktarılması Şekil 4.41'de verilen akış diyagramında gösterildiği şekilde gerçekleştirilmektedir.



Şekil 4.41. GA Uyarlama Adımı

Şekil içerisinde verilen A, ilk filtreleme adımında filtrelenmiş genlerden oluşan bir veri kümesini ifade etmektedir. a, A veri kümesi içerisinden rastgele oluşturulan gen alt kümelerinden her birini yani kromozomları temsil etmektedir. x değeri ise kromozomların içerisinde taşıdığı ve genlere(MD)'e ait numara bilgilerini tutan gen(GA)'leri göstermektedir.

4.2.2.2. Amaç fonksiyonu ve uygunluk değeri

Amaç fonksiyonunun belirlenmesi gen seçim işlemlerinin başarılı bir şekilde gerçekleşmesi için kritik bir öneme sahiptir. Genlerin sahip olduğu anlamı doğru ölçemeyen bir amaç fonksiyonu başarılı bir gen seçim işlemi gerçekleştiremeyecektir. Sadece bir sınıflandırma algoritması ile elde edilen sınıflandırma başarısı, genetik algoritmanın uygunluk değerinin hesaplanmasında yaygın olarak kullanılmaktadır. Ve sadece test işlemi gerçekleştirilen bu yöntemde sınıflandırma algoritmasının veri kümesini ezberlemesi problemi (Overtraining) ile sıklıkla karşılaşılabilmektedir. Bu sorunun önüne geçilebilmesi için bu çalışmada iki farklı işlem gerçekleştirilmiştir. İlk olarak tek bir sınıflandırma algoritması yerine KNN, DVM ve Naif Bayes algoritmalarının ayrı ayrı LOOCV çapraz doğrulama değerleri hesaplanmaktadır. Daha sonra LOOCV değerleri dikkate alınarak bu üç algoritmanın tahminlerini birleştiren bir birlik yöntem önerilmiştir. Böylece GA'nın veri kümesini ezberleme ihtimaline karşı dayanıklı olması sağlanmıştır.

Sınıflandırma işlemlerinde ortak bir sınıf tahmini yapılması için kullanılan çoğunluk oylaması Denklem (4.3)'de verildiği şekilde hesaplanabilmektedir.

$$B(X_i) = mode(S_1(X_i), S_2(X_i) \dots S_k(X_i)) \tag{4.3}$$

X uygunluk değeri belirlenmek istenen kromozomun seçtiği genler ile oluşturulan yeni veri kümesini, $B(X_i)$ bu veri kümesi içerisinde sınıf parametresi tahmin edilmesi istenen i örneğini, S sınıflandırma algoritmalarının sınıf tahminini ve k toplam sınıflandırıcı sayısını temsil etmektedir. $mode$ parametreler içerisinde en fazla bulunan sınıf değerini sonuç olarak geriye döndüren bir metottur. Bu hesaplama ile en çok oyu alan sınıf parametresi sonuç olarak belirlenmektedir. Ancak sınıflandırma algoritmalarının elde etmiş olduğu LOOCV değerlerinin bu sınıflandırma tahminine etki etmesi istenmektedir. Bu nedenle algoritmaların vermiş olduğu oyların ilgili algoritmanın LOOCV değerleri ile ağırlıklandırılması için Denklem (4.4)'den yararlanılmıştır.

$$B(X_i) = \dot{y} = \arg\max_{y} \sum_{j=1}^{k} LOOCV_j \, I(S_j(X_i) = y)) \tag{4.4}$$

Denklemde verilen $\dot{y}$ hesaplama sonucunda elde tahmin edilen sınıf değerini temsil ederken, $\arg\max_{y}$ metodu denklem içerisindeki ifadeyi maksimum yapan y sınıfını geriye döndürmektedir. I daha önceki hesaplamalarda da olduğu gibi içerisindeki koşulun doğru olması durumunda geriye 1, değilse 0 değerini döndüren bir fonksiyondur. Bu hesaplama sonucunda en başarılı LOOCV değerine sahip sınıflandırma daha yüksek bir oy hakkına sahip olurken düşük LOOCV değerine sahip algoritmanın oyu daha az etkili olmaktadır. En çok oyu olan y sınıfı, algoritma tarafından örneğin sınıf tahmini olarak belirlenmektedir.

Sadece kromozomlar içerisindeki genlere ait ekspresyon değerleri kullanılarak her bir örneğin birlik sınıflandırıcı ile sınıflandırılmasından sonra elde edilen sınıf parametresi tahminlerinin, gerçek sınıf parametreleri ile karşılaştırılması

için oluşturulan ve GA'nın amaç fonksiyonu olarak kullanılan hesaplama yöntemi Denklem (4.5)'de verilmiştir.

$$f(X) = \frac{1}{N}\sum_{i=1}^{N} I(B(X_i) = y_i) \tag{4.5}$$

Denklemde verilen N veri kümesi içerisindeki toplam örnek sayısını, y_i ise her bir örneğin sahip olmuş olduğu gerçek sınıf parametresini temsil etmektedir. I metodu gerçekleştirilen her doğru tahmin için 1, yanlış tahmin için 0 değerini üretmektedir. Sonuç olarak birlik algoritmanın doğru tahmin sayısı ile toplam örnek sayısının bölümünden ortaya çıkan değer ilgili kromozomun uyguluk değeri olarak kullanılmaktadır.

4.2.2.3. Seçilim

GA seçilim işlevi Rulet Tekerleği yöntemi ile gerçekleştirilmiştir. Her bir kromozomun seçilme olasılığı Denklem (4.6)'de verilen yöntem ile hesaplanmaktadır.

$$P_a = \frac{f_a}{\sum_{i=1}^{n} f_i} \tag{4.6}$$

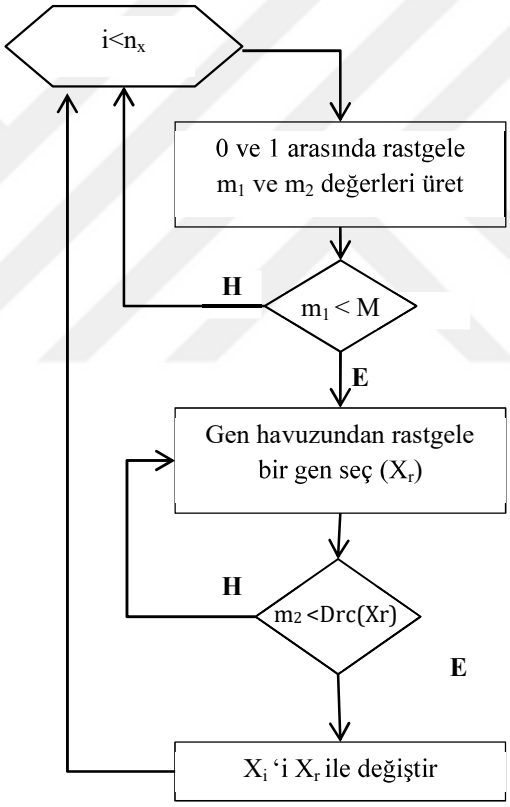
Denklemde verilen P_a, a kromozomuna ait hesaplanmak istenen olasılık değerini, f_a ise aynı kromozoma ait uygunluk değerini ifade etmektedir. n toplam kromozom sayısını temsil ederken f_i popülasyon içerisindeki her bir kromozomun uygunluk değerini göstermektedir.

4.2.2.4. Çaprazlama

Çalışma içerisinde GA'nın çaprazlama işlemleri, genlerin kromozom içerisindeki sırası bir önem taşımadığı için tek düze yöntem kullanılarak gerçekleştirilmiştir. Bu yöntem ile iki kromozom arasında rastgele sayıda ve rastgele noktada gen takası gerçekleştirilmektedir.

4.2.2.5 Mutasyon

Birden fazla filtreleme algoritması ile farklı sayıda seçilen genlerin birleştirilmesi nedeniyle farklı önem derecesine sahip genler aynı gen havuzunda bulunabilmektedir. Mutasyon işlevi uygulanırken önem derecesi yüksek olan genlerin daha az mutasyona uğraması, ve aynı genlerin gen havuzundan mutasyon işlevi ile kromozomlara dâhil edilme ihtimalinin yükseltilmesi için bir mutasyon işlevi geliştirilerek algoritma içerisinde kullanılmıştır. Geliştirilen mutasyon işlevinin akış diyagramı Şekil 4.42'de verilmiştir.



Şekil 4.42. Geliştirilen mutasyon işlevinin akış diyagramı

Şekilde verilen M mutasyon kat sayısını, n_x kromozom içerisindeki toplam gen sayısını, X_i kromozom i. sıradaki geni, Xr ise gen havuzundan rastgele seçilen bir geni ifade etmektedir.

4.2.2.6. Değerlendirme

Mutasyon işlevinden sonra popülasyon içerisindeki tüm kromozomların tekrar uygunluk değeri hesaplanmaktadır. En başarılı kromozomlar bir sonraki kuşağa aktarılarak bir tekrar tamamlanmaktadır. İstenilen uygunluk değerine ya da istenen tekrar sayısına ulaşılması durumunda algoritma sonlandırılarak en başarılı kromozomun içindeki genler, gen seçim sonucu olarak kabul edilmektedir.

4.3. Deneysel Sonuçlar

Geliştirilen Birlik-Hibrit algoritma ile gerçekleştirilen gen filtreleme, gen seçimi ve sınıflandırma işlemleri Intel i7-2670QM 2.20 Ghz işlemcisine ve 8 GB RAM'e sahip bir bilgisayar üzerinde gerçekleştirilmiştir. Elde edilen deneysel sonuçlar şu sıra ile sunulmuştur;

İlk olarak gen filtreleme ve öznitelik çıkarımı işlemleri sonucunda elde edilen bulgular verilerek birlik algoritmanın gen filtreleme süreci ve sonuçları ele alınmıştır. Güçlendirilmiş Genetik Algoritma ile gerçekleştirilen gen seçim işlemleri sonucunda ulaşılan performans değerleri incelenmiş ve literatürdeki diğer yaklaşımlarla karşılaştırılmıştır. Son olarak elde edilen bulguların biyolojik öneminin incelenmesi ile bitirilmiştir.

4.3.1. Filtreleme

Lösemi veri kümesi üzerinde FKS, BK ve WRT algoritmaları ile farklı sayılarda genler filtrelenerek yeni gen-alt kümeleri elde edilmiştir. Elde edilen bu gen-alt kümeleri KNN algoritması ile sınıflandırılmıştır. Gerçekleştirilen sınıflandırma işlemlerinin performansı LOOCV, K5 ve K10 çapraz doğrulama yöntemleri ile ölçülmüştür.

FKS algoritması ile filtrelenen genlere ait LOOCV ve çapraz doğrulama değerleri sırası ile Çizelge 4.2 ve Çizelge 4.3'de verilmiştir. Bu çizelgeler incelendiğinde

FKS algoritmasının en başarılı LOOCV ve çapraz doğrulama değerini 50 gen ile elde ettiği görülmektedir.

Çizelge 4.2. FKS LOOCV değerleri

Gen Sayısı	Başarılı Sınıflandırma	Başarı (%)	AUC	Süre (ms)
2	46	63,8889	0,5830	5
3	51	70,8333	0,6736	2
4	56	77,7778	0,7549	3
5	65	90,2778	0,8787	3
6	62	86,1111	0,8468	3
7	62	86,1111	0,8374	4
8	65	90,2778	0,8787	4
9	65	90,2778	0,8881	5
10	64	88,8889	0,8681	5
25	69	95,8333	0,9494	12
50	**72**	**100**	**1**	**23**
75	70	97,2222	0,9600	34
100	71	98,6111	0,9800	45
150	70	97,2222	0,9600	68
200	70	97,2222	0,9694	89
250	69	95,8333	0,9587	112
500	70	97,2222	0,9600	221

Çizelge 4.3. FKS Çapraz doğrulama değerleri

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	En Yüksek	En Düşük	Ortalama	En Yüksek	En Düşük	Ortalama
2	80	50	67,9048	87,5	37,5	65,3571
3	80	66,6667	73,619	85,7143	50	72,5000
4	92,8571	71,4286	79,2381	100	57,1429	79,1071
5	93,3333	78,5714	88,7619	100	71,4286	90,1786
6	100	73,3333	83,4286	100	62,5	84,8214
7	100	73,3333	87,619	100	62,5	86,2500
8	100	86,6667	91,8095	100	71,4286	88,9286
9	100	86,6667	91,8095	100	75	90,5357
10	100	85,7143	91,7143	100	75	89,1071
25	100	92,8571	95,8095	100	85,7143	97,1429
50	**100**	**92,8571**	**98,5714**	**100**	**100**	**100**
75	100	85,7143	95,8095	100	85,7143	95,8929
100	100	85,7143	97,1429	100	85,7143	97,1429
150	100	85,7143	97,1429	100	85,7143	97,1429
200	100	92,8571	97,2381	100	85,7143	97,1429
250	100	92,8571	97,2381	100	85,7143	97,1429

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	En Yüksek	En Düşük	Ortalama	En Yüksek	En Düşük	Ortalama
500	100	85,7143	95,8095	100	85,7143	95,8929

BK algoritması ile filtrelenerek oluşturulan gen-alt kümelerinin sınıflandırılması sonucunda elde edilen LOOCV değerleri Çizelge 4.4, çapraz doğrulama değerleri ise Çizelge 4.5'de verilmiştir. LOOCV değerleri incelendiğinde 150 gene sahip olan gen-alt kümesinin en başarılı sonuçlara sahip olduğu görülmektedir.

Çizelge 4.4. BK LOOCV değerleri

Gen Sayısı	Başarılı Sınıflandırma	Başarı (%)	AUC	Süre (ms)
2	68	94,4444	0,9481	2
3	67	93,0556	0,9281	2
4	67	93,0556	0,9187	2
5	68	94,4444	0,9387	3
6	67	93,0556	0,9094	3
7	68	94,4444	0,9200	4
8	68	94,4444	0,9200	4
9	68	94,4444	0,9200	5
10	68	94,4444	0,9200	5
25	68	94,4444	0,9200	12
50	68	94,4444	0,9200	22
75	68	94,4444	0,9200	35
100	68	94,4444	0,9200	46
150	**69**	**95,8333**	**0,9400**	**66**
200	68	94,4444	0,9200	90
250	69	95,8333	0,9400	112
500	67	93,0556	0,9000	227

Çizelge 4.5. BK çapraz doğrulama değerleri

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	En Yüksek	En Düşük	Ortalama	En Yüksek	En Düşük	Ortalama
2	100	92,8571	94,381	100	85,7143	94,4643
3	93,3333	92,8571	93,0476	100	85,7143	93,2143
4	100	92,8571	94,4762	100	85,7143	94,6429
5	100	92,8571	94,4762	100	85,7143	94,6429
6	100	85,7143	94,4762	100	75	94,6429
7	100	85,7143	94,4762	100	75	94,6429
8	100	85,7143	95,8095	100	75	94,6429
9	100	85,7143	93,1429	100	75	94,6429

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	**En Yüksek**	**En Düşük**	**Ortalama**	**En Yüksek**	**En Düşük**	**Ortalama**
10	100	85,7143	93,1429	100	75	94,6429
25	100	85,7143	94,4762	100	75	96,0714
50	100	85,7143	93,1429	100	75	96,0714
75	100	85,7143	93,1429	100	75	96,0714
100	100	85,7143	93,1429	100	75	96,0714
150	100	78,5714	93,0476	100	75	94,6429
200	100	85,7143	94,4762	100	71,4286	93,2143
250	100	86,6667	95,9048	100	75	94,6429
500	100	85,7143	93,0476	100	75	93,2143

Çizelge 4.6 ve

Çizelge 4.7'de verilen WRT algoritması ile elde edilmiş LOOCV ve çapraz doğrulama değerleri incelendiğinde 9 gene sahip olan gen-alt kümesinin çizelgede de en başarılı değerlere sahip olduğu görülmektedir.

Çizelge 4.6. WRT LOOCV değerleri

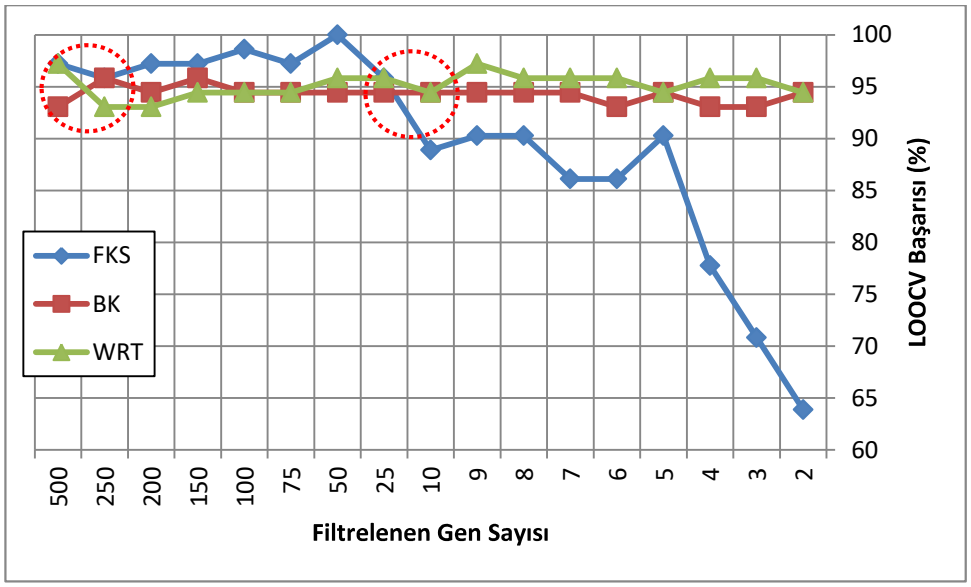
Gen Sayısı	**Başarılı Sınıflandırma**	**Başarı (%)**	**AUC**	**Süre (ms)**
2	68	94,4444	0,9387	5
3	69	95,8333	0,9494	2
4	69	95,8333	0,9494	3
5	68	94,4444	0,9387	3
6	69	95,8333	0,9587	6
7	69	95,8333	0,9587	4
8	69	95,8333	0,9587	4
9	**70**	**97,2222**	**0,9787**	**5**
10	68	94,4444	0,9387	7
25	69	95,8333	0,94	12
50	69	95,8333	0,94	23
75	68	94,4444	0,92	34
100	68	94,4444	0,92	50
150	68	94,4444	0,92	68
200	67	93,0556	0,9	90
250	67	93,0556	0,9	113
500	70	97,2222	0,96	223

Çizelge 4.7. WRT çapraz doğrulama değerleri

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	**En Yüksek**	**En Düşük**	**Ortalama**	**En Yüksek**	**En Düşük**	**Ortalama**
2	100	85,7143	92,9524	100	71,4286	93,0357

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	**En Yüksek**	**En Düşük**	**Ortalama**	**En Yüksek**	**En Düşük**	**Ortalama**
3	100	85,7143	94,2857	100	71,4286	95,7143
4	100	85,7143	94,2857	100	71,4286	95,7143
5	100	92,8571	95,8095	100	85,7143	94,6429
6	100	92,8571	95,8095	100	85,7143	95,8929
7	100	92,8571	95,8095	100	85,7143	94,6429
8	100	92,8571	95,8095	100	85,7143	95,8929
9	**100**	**92,8571**	**97,2381**	**100**	**85,7143**	**97,3214**
10	100	92,8571	94,4762	100	85,7143	94,6429
25	100	86,6667	95,9048	100	75	96,0714
50	100	86,6667	95,9048	100	75	96,0714
75	100	86,6667	94,5714	100	75	94,6429
100	100	86,6667	94,4762	100	75	94,6429
150	100	86,6667	94,4762	100	75	94,6429
200	100	86,6667	94,4762	100	75	94,6429
250	100	86,6667	94,4762	100	75	94,6429
500	100	86,6667	94,4762	100	75	94,6429

Üç ayrı algoritma ile elde edilen LOOCV değerleri grafik üzerinde incelendiğinde (Şekil 4.43) WRT ve BK algoritması ile seçilen genler birbirine yakın değerlere sahipken FKS algoritması ile seçilen genler gen sayısı azaldıkça düşüş eğilimi göstermektedir. Üç algoritma ile seçilen genlerin birleştirildiği veri kümesinin dengeli bir dağılıma sahip olabilmesi için, grafik üzerindeki algoritmaların birbirine en yakın olduğu noktalardan en yüksek başarı oranına sahip olanı seçilmelidir. Ancak grafikten de görüldüğü üzere bu kesişim noktaları her algoritma için en başarılı genleri içermemektedir. Genlerin önem derecelerinin hesaplanması için önerilen (4.2) denklem sayesinde kesişim noktaları göz ardı edilerek her bir algoritmanın en başarılı olduğu noktalar yani farklı gen sayısına sahip gen-alt kümeleri ayrı ayrı seçilebilmektedir. Bu durumda FKS ile 50, BK ile 150 ve WRT ile filtrelenmiş 9 gene sahip en başarılı LOOCV değerlerine sahip gen-alt kümeleri her bir genin önem derecesinin hesaplanmasından sonra tek bir veri kümesi içerisinde birleştirilmektedir. Sonuç olarak 7200 genden filtrelenerek oluşturulan, 209 gen sayısına sahip yeni bir veri kümesi elde edilmiştir.



Şekil 4.43. Filtreleme sonuçlarının karşılaştırması

500 gen ile yapılan sınıflandırma işlem sürelerinin ortalaması 220.3ms iken WRT algoritması ile seçilen 9 gen ile yapılan sınıflandırma işlemleri sadece 5ms gibi çok kısa bir süre içerisinde gerçekleştirilmiştir. Orijinal veri kümesinin sınıflandırma işlemlerinde geçen süre ise 7200 gen ile 4000ms olduğu dikkate alınırsa gen filtreleme işleminin hesaplama maliyetinin azaltılmasına yapmış olduğu katkının önemi anlaşılmaktadır.

4.3.2. Öznitelik çıkarma

Bu kısımda birlik algoritma ile filtrelenen genlerin TBA analizi sonucunda elde edilen sonuçlar, orijinal veri kümesine ait TBA analizi sonucunda elde edilen değerler ile karşılaştırılmıştır. Böylelikle filtrelenen genlerin taşımış oldukları bilginin, Lösemi kanserinin AML ve ALL türlerini ayırt edici özellik taşıyıp taşımadıkları incelenmek istenmiştir.

Filtrelenen genler ile yapılan faktör analizi sonucunda 70 adet temel bileşen bulunmuştur. Bu öz bileşenler kullanılarak sırası ile 1,2,3..10 faktöre sahip yeni veri kümeleri elde edilmiştir. Yapılan analize göre veri kümesi içerisindeki özniteliklerin taşıdığı anlamın %95'den fazlası 19 temel bileşen ile temsil edilmektedir. Bu yüzden 19 faktörden oluşan veri kümesinin de dâhil edildiği

yeni oluşturulan veri kümelerinin sınıflandırma başarıları LOOCV ve çapraz doğrulama değerleri KNN algoritması ile hesaplanmak istenmiştir. Bu veri kümelerinin elde etmiş oldukları LOOCV değerleri Çizelge 4.8'de, çapraz doğrulama değerleri ise Çizelge 4.9'de verilmiştir.

Çizelge 4.8. TBA ile elde edilen faktörlerin LOOCV değerleri

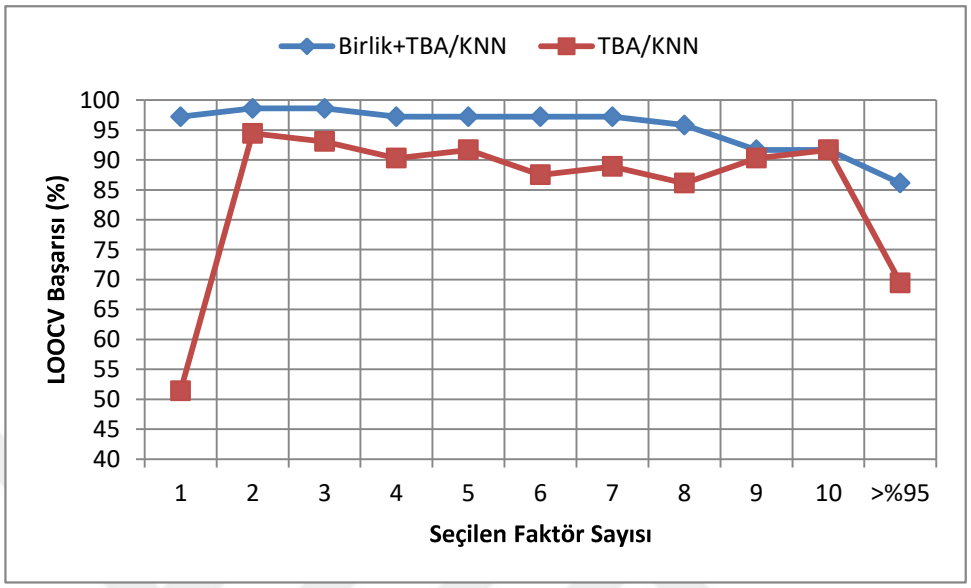
Seçilen Faktör Sayısı	Başarılı Sınıflandırma	Başarı (%)	AUC	Süre (ms)
1	70	97,2222	0,9600	1
2	**71**	**98,6111**	**0,9800**	**5**
3	71	98,6111	0,9800	2
4	70	97,2222	0,9694	3
5	70	97,2222	0,9694	3
6	70	97,2222	0,9694	4
7	70	97,2222	0,9694	4
8	69	95,8333	0,9494	4
9	66	91,6667	0,8894	5
10	66	91,6667	0,8987	5
>%95	62	86,1111	0,8	9

Çizelge 4.9. TBA ile elde edilen faktörlerin çapraz doğrulama değerleri

	K5 Çapraz Doğrulama (%)			K10 Çapraz Doğrulama (%)		
Gen Sayısı	En Yüksek	En Düşük	Ortalama	En Yüksek	En Düşük	Ortalama
1	100	92,8571	97,1429	100	85,714	97,1429
2	**100**	**92,8571**	**98,5714**	**100**	**85,714**	**98,5714**
3	100	92,8571	98,5714	100	85,714	98,5714
4	100	85,7143	97,1429	100	85,714	97,3214
5	100	92,8571	98,5714	100	85,714	97,3214
6	100	92,8571	98,5714	100	85,714	97,3214
7	100	92,8571	97,2381	100	85,714	97,3214
8	100	92,8571	95,8095	100	85,714	95,8929
9	93,3333	86,6667	91,7143	100	85,714	93,0357
10	100	85,7143	93,0476	100	85,714	94,4643
>%95	93,3333	71,4286	83,2381	100	71,429	85,8929

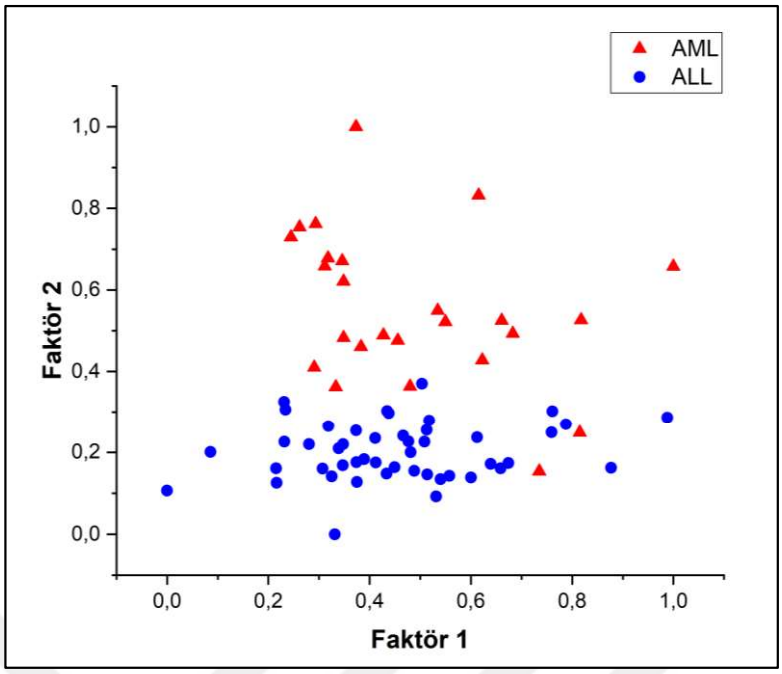
Yapılan sınıflandırma işlemlerinden elde edilen çapraz doğrulama değerleri incelendiğinde 2 faktöre sahip veri kümesinin en yüksek başarıyı elde ettiği görülmektedir. 7200 adet genden birlik filtreleme ve TBA algoritmaları ile boyutu sadece 2 faktöre indirgenen veri kümesi üzerinde gerçekleştirilen

LOOCV testinde sadece bir örneğin sınıflandırma tahmini başarısız olmuştur. Bu sonuçlar ile gen filtrelemesi gerçekleştirilmeden yapılan temel bileşen analizine ait LOOCV değerlerinin karşılaştırılması Şekil 4.44'de verilmiştir.

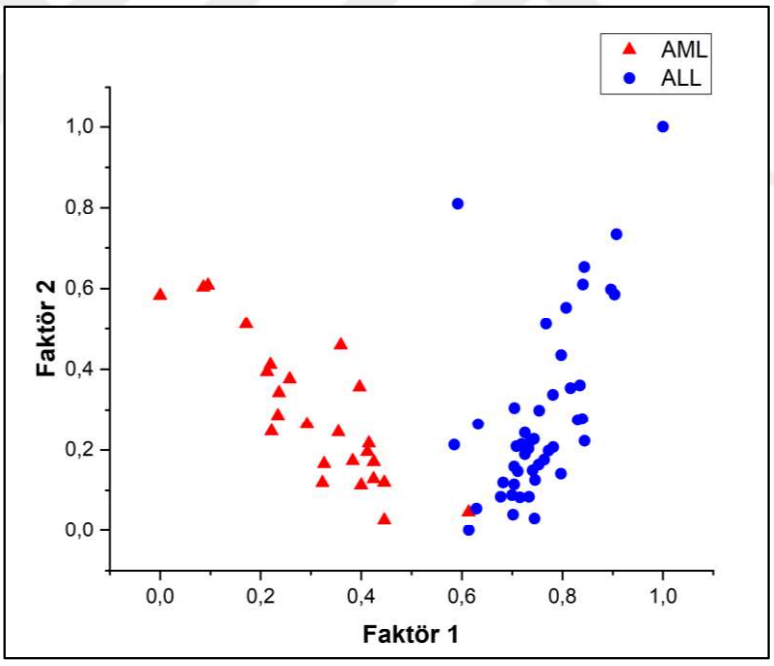


Şekil 4.44. Birlik algoritma ile filtrelenmiş genlerin ve orijinal veri kümesinin TBA sonrası elde edilen LOOCV değerlerinin karşılaştırılması

Verilen grafik incelendiğinde birlik algoritma ile seçilen genlerin daha ayırt edici özelliğe sahip olduğu, tüm veri kümesi ile elde edilen faktörlerden daha başarılı sonuçlar elde ettiği görülmektedir. Aynı sonuç bu veri kümelerinin dağılım grafikleri incelendiğinde de gözle görülür bir şekilde ortaya çıkmaktadır. Veri kümesinin TBA sonrası 2 faktörden oluşan örneklere ait dağılım grafiği Şekil 4.45'de verilirken, birlik algoritma ile filtrelenen genlerden elde edilen 2 faktörlü veri kümesine ait dağılım grafiği Şekil 4.46'de verilmiştir.



Şekil 4.45. Tüm genlerin TBA sonrası faktör dağılım grafiği



Şekil 4.46. Birlik algoritma ile filtrelenen genlerin TBA sonrası dağılım grafiği

4.3.3. Gen seçimi

Daha az gen ile daha yüksek sınıflandırma başarısının elde edilebilmesi için ilk adımda filtrelenen genler ikinci adımda güçlendirilmiş bir genetik algoritma ile daha detaylı bir gen seçim işlemine tabi tutulmaktadır. Bu amaçla üç

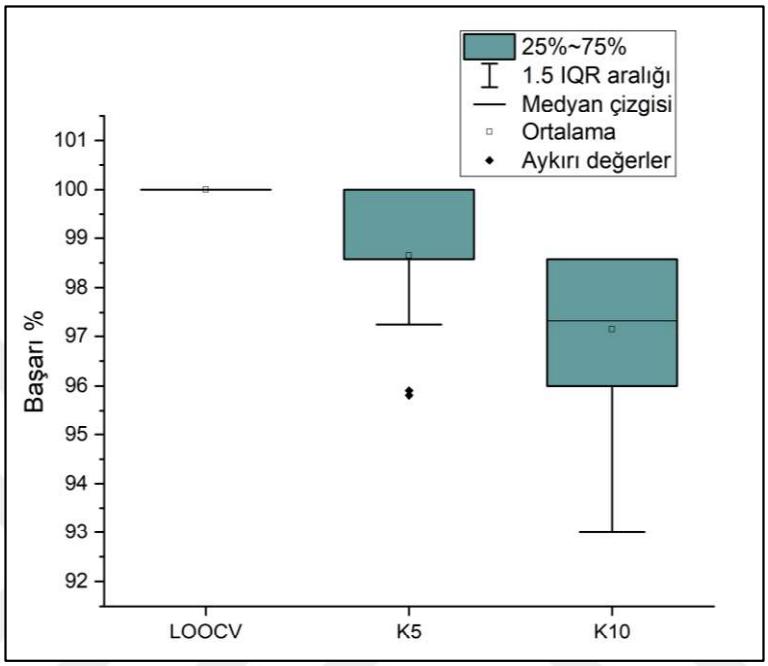
algoritmadan elde edilen ve önem derecesi hesaplanarak birleştirilmiş veri kümesi genetik algoritmanın girişine verilmiştir. Genetik algoritmanın başlangıç parametreleri mutasyon oranı 0.09, çaprazlama oranı olarak 0.35 belirlenmiştir. Genetik algoritmanın popülasyon sayısı ise yeterli sayıda farklı çözüm kümelerini araştırabilmesi için 20 olarak belirlenmiştir. Durdurma kriteri olarak 1000 tekrar sayısına ya da %100 uygunluk değerine erişmesi durumları belirlenmiştir. Genetik algoritma ile gerçekleştirilen 6,5,4,3 ve 2 gen seçimi için ayrı ayrı olmak üzere sınıflandırma ve çapraz doğrulama işlemleri 100 kez tekrar edilerek ortalama değerler alınmıştır. Daha sonra GA tarafından seçilen genler KNN algoritması ile sınıflandırılmıştır. Sınıflandırma başarısının ölçülmesi için veri kümesi %80'i eğitim %20'si test olmak üzere ayrılmıştır. Sınıflandırma performansının daha detaylı incelenebilmesi için ise ilk adımda da olduğu gibi LOOCV, K5 ve K10 çapraz doğrulama işlemleri gerçekleştirilmiştir. 100 tekrar sonucunda elde edilen ortalama başarı değerleri Çizelge 4.10 'da verilmiştir.

Çizelge 4.10. Seçilen genlerin test ve çapraz doğrulama sonuçları

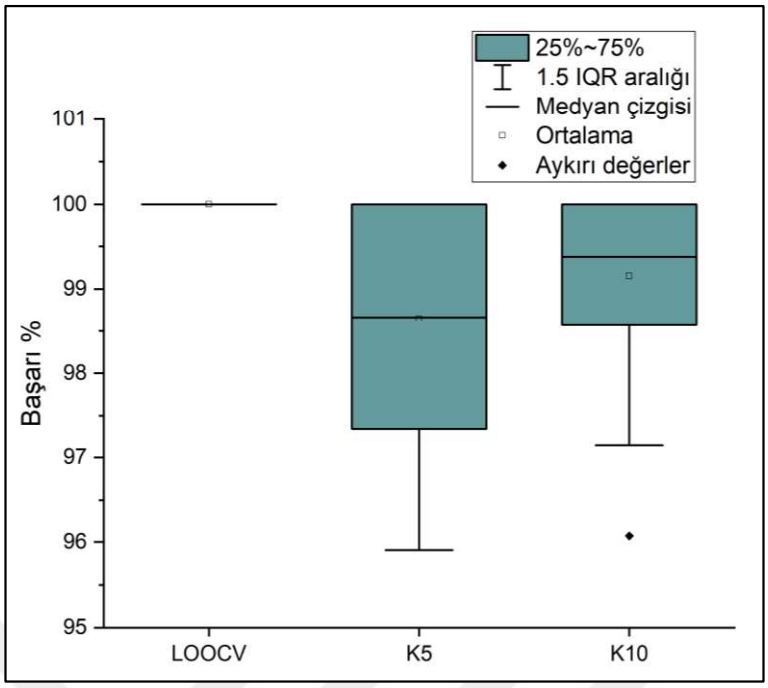
Seçilen Gen Sayısı	Test Başarısı (%)	Test AUC	LOOCV Başarısı (%)	LOOCV AUC	K5 Ort. (%)	K10 Ort. (%)
6	100	1	100	1	100	100
5	100	1	100	1	98.57	97.14
4	100	1	100	1	97.33	97.32
3	100	1	100	1	98.64	99.15
2	100	1	100	1	98.57	97.14

İlk adımda %100 LOOCV değeri en az 50 gen ile elde edilebilirken ikinci adımda kullanılan güçlendirilmiş genetik algoritma ile seçilen genler ile sayesinde aynı başarı oranına sadece iki gen ile ulaşılması sağlanmıştır. Ayrıca daha az gen ile bu başarı oranının sağlanması hesaplama maliyeti ve süresinde (2 ve 3 gen için ortalama 2.5 ms) büyük oranda kazanç sağlamıştır. Sonuç olarak optimum gen sayısı / performans değeri 2 gen için elde edilmiş olsa da çapraz doğrulamalar da dikkate alındığında seçilen üç genin daha iyi sonuçlar elde ettiği anlaşılmaktadır. Seçilen gen sayısı arttırılarak 6 gen seçimi ile elde edilen sonuçlarda da olduğu gibi başarı değerlerinin artırılması mümkün olsa da en az gen ile en başarılı sonuçlar alınmak istenmektedir. Her bir tekrarda elde edilen

LOOCV ve çapraz doğrulama değerinin detaylı bir şekilde incelenebilmesi için kutu grafikleri oluşturulmuştur. İki gen seçimi sonucunda elde edilen değerler ile oluşturulan kutu grafiği Şekil 4.47'de gösterilirken üç gen için oluşturulan grafik Şekil 4.48'de verilmiştir.



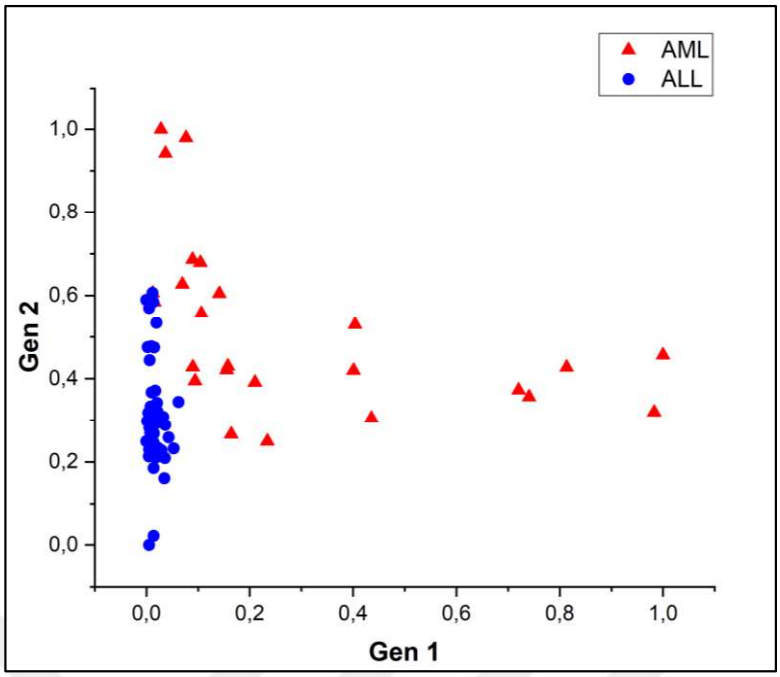
Şekil 4.47. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (2 gen için)



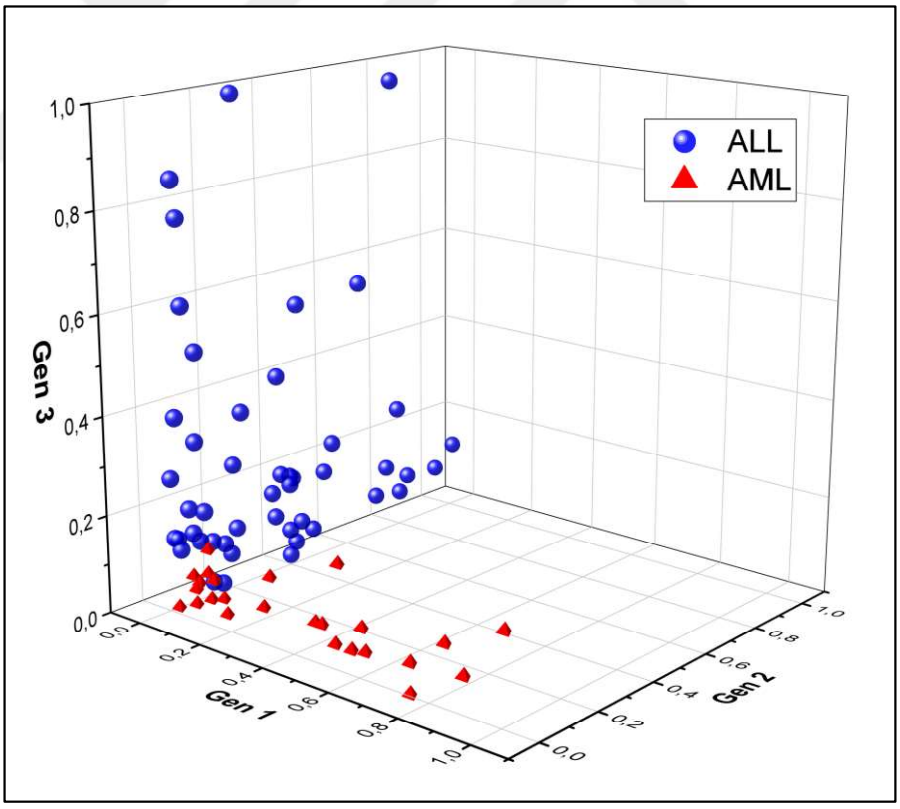
Şekil 4.48. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (3 gen için)

GA'nın ilk durdurma kriteri yani 1000 tekrar sayısına erişmesi koşulu, seçilen genlerin uygunluk değerinin daha önceki tekrarlarda %100'e ulaşmasından dolayı hiçbir zaman gerçekleşmiştir. En çok tekrar sayısı 845, en az tekrar sayısı 1 olurken ortalama tekrar sayısı ise 173,95 olarak gözlemlenmiştir.

İki ve üç gen seçimi ile oluşturulan veri kümelerindeki genlere ait ekspresyon seviyelerinin dağılım grafikleri Şekil 4.49 ve Şekil 4.50 de verilmiştir.



Şekil 4.49. Gen ekspresyon değerlerinin dağılım grafiği (2 gen için)



Şekil 4.50. Gen ekspresyon değerlerinin dağılım grafiği (3 gen için)

Dağılım grafikleri incelendiğinde, her ne kadar 2 gen ile yüksek bir sınıflandırma ve LOOCV başarısı elde edilse de, 3 gen ile yapılan seçimlerde daha ayırt edilebilir bir veri kümesi elde edildiği görülmektedir.

Geliştirilen algoritma farklı örnek ve gen sayısına sahip literatürde sıklıkla kullanılan (Benchmark) farklı mikro dizi veri kümeleri üzerinde de denenmiştir. Bu veri kümeleri üzerinde gerçekleştirilen gen seçim işlemleri sonucunda elde edilen bulgular EK A.'da sunulmuştur.

4.3.4. Sınıflandırma

Geliştirilen Birlik-Hibrit algoritma ile gerçekleştirilen gen seçimi işlemleri sonrasında genlerin sınıflandırılması için KNN, hızlı ve başarılı bir algoritma olması nedeniyle tercih edilmiştir. Geliştirilen modelin sınıflandırma algoritmalarından bağımsız olup olmadığını kontrol edilebilmesi için seçilen genler KNN'ye ek olarak farklı yöntemler ile çalışan sınıflandırma algoritmaları ile de çapraz doğrulama işlemlerine tabi tutulmuştur. Buna ek olarak Kappa ve OKH (Ortalama Kare Hata, MSE, Mean square error) değerleri de hesaplanarak Çizelge 4.11'de verilmiştir.

Çizelge 4.11. Seçilen genlerin farklı algoritmalar ile sınıflandırılması sonucunda elde edilen değerler (K5)

Sınıflandırma Algoritması	Başarı (%)	AUC	Kappa	OKH
DVM	98.57	1	0.967	0.215
NB	97.14	0.996	0.935	0.125
Doğrusal Regresyon	94.38	1	0.872	0.643
YSA	98.57	1	0.967	0.129
Rastgele Orman	95.90	0.991	0.912	0.150

Çizelge incelendiğinde, SVM ve YSA'nın aynı sınıflandırma başarısı ve AUC değerine sahip olduğu, diğer algoritmaların ise bu algoritmalara yakın sonuçlar elde ettiği görülmektedir. Kappa değerleri incelendiğinde tüm algoritmaların Kappa için kritik bir değer olan 0.75'in üzerinde olduğu ve bu nedenle başarılı bir sınıflandırma işlemi gerçekleştirdiği söylenebilmektedir. OKH değerleri karşılaştırıldığında tüm algoritmaların kabul edilebilir bir hata değeri elde ettiği

görülmektedir. Sonuç olarak, seçilen genlerin farklı sınıflandırma algoritmalarıyla birlikte başarılı sonuçlara sahip olması, geliştirilen yaklaşımın sınıflandırıcıdan bağımsız olduğunu ortaya koymaktadır.

Geliştirilen yaklaşımda gen seçim adımında amaç fonksiyonunun hesaplanması için birlik bir sınıflandırma algoritması kullanılmaktadır. Çizelge 4.12'de amaç fonksiyonu için farklı sınıflandırma algoritmaları kullanıldığı durumlarda elde edilen sınıflandırma sonuçları paylaşılmıştır. Birlik sınıflandırma algoritması kendisini oluşturan algoritmaların bireysel olarak elde ettiği başarıdan daha yüksek bir başarı oranı elde ettiği görülmektedir.

Çizelge 4.12. Farklı sınıflandırma algoritmalarının amaç fonksiyonu olarak kullanıldığı durumda elde edilen sınıflandırma sonuçları

Sınıflandırma Algoritması	**LOOCV (%)**	**AUC**	**K5 (%)**	**K10 (%)**
KNN	98.61	0.980	98.57	98.57
DVM	98.61	0.989	97.14	97.14
NB	95.83	0.968	98.57	97.14
Birlik sınıflandırma Algoritması	100	1	98.57	97.14

4.4. Literatür Karşılaştırması

Geliştirilen algoritma ile elde edilen test, LOOCV ve çapraz doğrulama değerleri literatürdeki yöntemler ile karşılaştırıldığında birçok sınıflandırma kriterinde daha başarılı olduğu görülmektedir (Çizelge 4.13).

Çizelge 4.13. Geliştirilen algoritma ve literatürdeki çalışmaların sonuçları

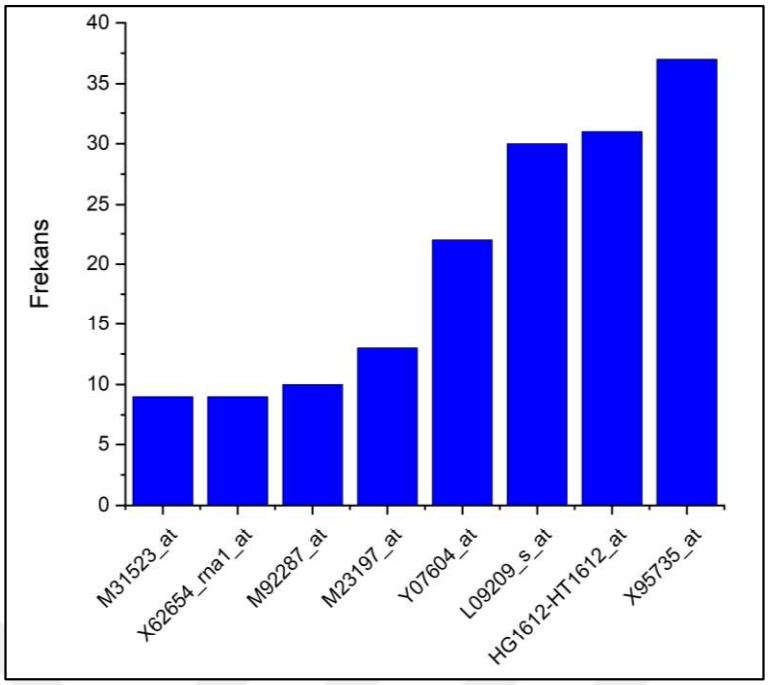
Yöntem	**Gen Sayısı**	**Test (%)**	**LOOCV (%)**	**K5 (%)**	**K10 (%)**
GALA (Motieghader vd., 2017)	2	**100**	-	-	-
E_u (Ghosh vd., 2019)	2	**100**	-	-	-
GBC (Alshamlan vd., 2015)	4	100	-	-	-
IBPSO (Jain vd., 2018)	4	100	-	-	-
E_1 (Ghosh vd., 2019)	12	100	-	-	-
SLLE-SC^2 (Xu vd., 2018)	5	99.70	-	-	-
M-SVM (Baliarsingh vd., 2019)	7	98.82	-	-	-
PSO + Adaptive KNN (Kar vd., 2015)	3	97.06	-	95,89	-
IG-SGA (Salem vd. 2017)	3	97.06	-	-	-
SVM+RFE (Fu ve Fu-Liu, 2005)	4	97.06	-	-	-
SMMDA/KNN (Cui vd., 2013)	8	96.88	-	-	-

Yöntem	Gen Sayısı	Test (%)	LOOCV (%)	K5 (%)	K10 (%)
Bagboost (Dettling, 2004)	200	95.92	-	-	-
HPSO TS (Shen vd., 2008)	7	95.81	-	-	-
SLR (Algamal ve Lee, 2019)	7	95.51	-	-	-
Pure TS (Shen vd., 2008)	5	94.24	-	-	-
AEN-CMI (Wang vd., 2019)	26.85	91.05	-	-	-
GA/SVM (Peng vd., 2003)	6	-	100	-	-
GA/SVM (Li vd., 2008)	7	-	100	-	-
MMACO/SVM (Yu vd., 2009)	6,3	-	100	-	-
ACO/SVM (Yu vd., 2009)	8,6	-	100	-	-
SNR/SVM (Yu vd., 2009)	100	-	100	-	-
IG+GA/KNN (Yang vd., 2008)	203	-	100	-	-
GS2/KNN (Yang vd., 2006)	10/85	-	98.60	97.10	-
GS1/KNN (Yang vd.,, 2006)	60/100	-	98.60	97.90	-
SVM (Furey vd., 2000)	500	-	94.10	-	-
MBPSO (Mohamad vd., 2011)	2	-	-	**100**	
MSFCM+WNN (Zainuddin ve Ong, 2011)	10	-	-	-	100
GSA (Kumar vd., 2012)	10	-	-	100	-
SWKC (Shim vd., 2009)	14.2	-	-	(3Fold) 98.20	-
Geliştirilen yaklaşım	6	100	100	100	**100**
	3	100	100	98.64	99.15
	2	**100**	**100**	98.57	97.14

Elde edilen test sonuçları incelendiğinde geliştirilen algoritma, GALA ve E_u algoritmalarının en başarılı test sonucuna en az sayıda (2 gen) seçerek ulaştığı görülmektedir. LOOCV değerinde ise geliştirilen algoritma %100 başarı oranını sadece 2 gen ile ederek literatürdeki çalışmaların önüne geçmektedir. Geliştirilen algoritma K10 değerini paylaşan tek yaklaşım olan MSFCM+WNN yöntemini daha az gen sayısı (6 gen) ile %100 başarı oranına erişerek geçmiştir. Algoritma sadece K5 çapraz doğrulama değerinde MBPSO algoritması tarafından geçilmiştir.

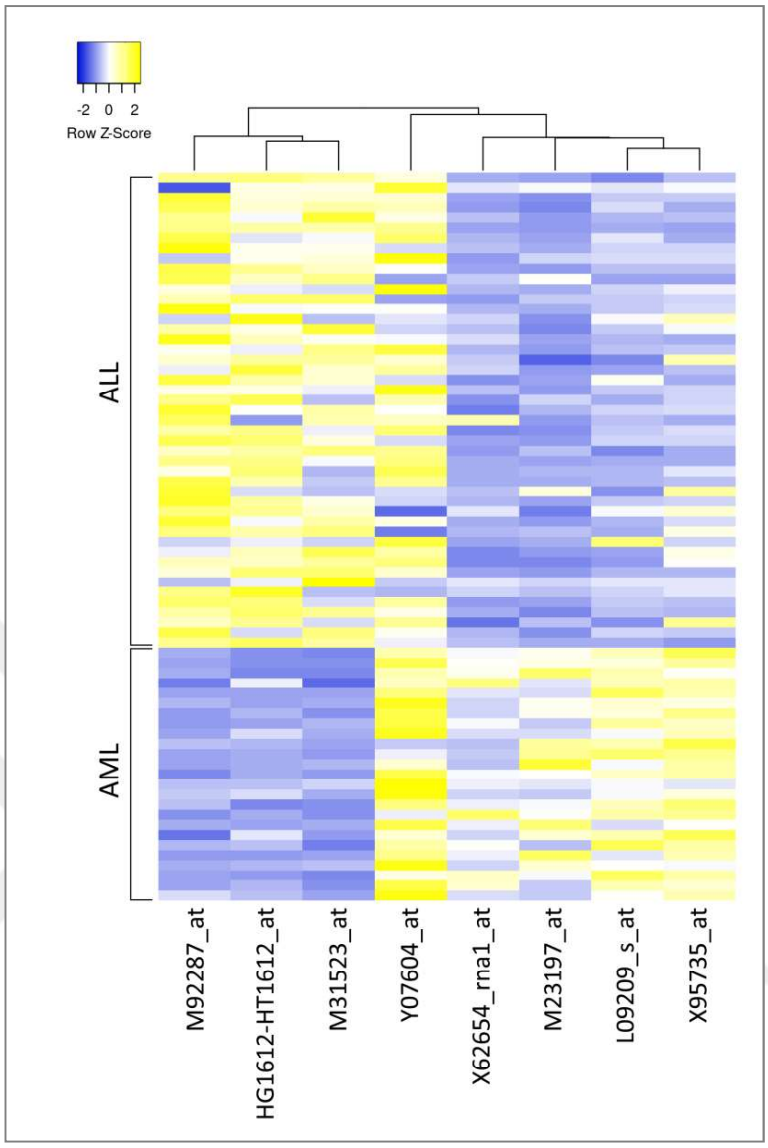
4.5. Biyolojik Bulgular

Gen-kanser ve gen-gen ilişkilerinin açığa çıkarılması kanser araştırmalarında önemli bir yere sahiptir. Bu nedenle geliştirilen yaklaşım ile seçilen ve Lösemi kanserinin türleri olan ALL ve AML hastalıklarını ayırt etmede başarılı olan genlerin biyolojik temelleri araştırılmak istenmiştir. Bu nedenle geliştirilen yaklaşımın yüz tekrar sonucunda en çok seçim yaptığı genler ve seçilme sıklıkları bulunmuştur. Bu genler ve seçim sıklıkları Şekil 4.51 verilmiştir.



Şekil 4.51. En sık seçilen 8 gen ve seçilme sıklığı

En çok seçilen genlerin farklı örneklere ait gen ekspresyon değerlerinin incelenebilmesi için bu genlere ait sıcaklık haritası oluşturulmuştur. Sıcaklık haritası oluşturulurken kümeleme metodu(clustering method) olarak ortalama bağlantı (average linkage), uzaklık ölçümü için ise Öklid uzaklığı kullanılmıştır. Sıcaklık haritası incelendiğinde seçilen genlerin ALL ve AML örneklerinde farklı ekspresyon seviyelerine sahip olduğu görülmektedir. Bu durumun bu iki hastalığın ayırt edilmesinde ki önemi, genler arasındaki korelasyon bağıntıları Şekil 4.52'de verilen sıcaklık haritası ile incelenebilmektedir.



Şekil 4.52. En sık seçilen 8 gen ile elde edilen sıcaklık haritası

Geliştirilen algoritma ile seçilen genlerin lösemi hastalığı ile olan ilişkisinin belirlenebilmesi için en çok seçilen genlerin gerçek isimleri HUGO gen isimleri veri tabanından, genlerin fonksiyonlarına ait bilgiler ise Avrupa Biyoinformatik Enstitüsü'nün (European Bioinformatics Institute) Gen Ontoloji veri tabanından araştırılırmıştır ve elde edilen bilgiler Çizelge 4.14'de özetlenmiştir.

Çizelge 4.14. Seçilen genlerin isimleri ve işlevleri

Gen kodu	**Sembolü**	**İsmi**	**İşlevleri**
X95735_at	ZYX	Zyxin	RNA bağlanması, metal iyon bağlanması, demir bağlaması
HG1612-HT1612_at	MARCKSL1	MARCKS like 1	Aktin bağlaması, kalmodulin bağlanması, protein bağlanması

Gen kodu	Sembolü	İsmi	İşlevleri
L09209_s_at	APLP2	Amyloid beta precursor like protein 2	Heparin bağlanması, özdeş protein bağlanması, protein bağlanması, serin tipi endopeptidaz inhibitör aktivitesi, geçiş metali iyon bağlanması
Y07604_at	NME4	NME/NM23 nucleoside diphosphate kinase 4	ATP bağlanması, kardiyolipin bağlanması, metal iyon bağlanması, nükleozid difosfat kinaz aktivitesi, protein bağlanması
M23197_at	CD33	CD33 molecule	Karbonhidrat bağlanması, protein bağlanması, protein fosfataz bağlanması, sialik asit bağlanması, sinyal reseptör aktivitesi
M92287_at	CCND3	Cyclin D3	Sikline bağımlı protein serin / treonin kinaz aktivitesi, sikline bağlı protein serin / treonin kinaz düzenleyici aktivite, protein bağlanması, protein kinaz aktivite katkısı, protein kinaz bağlanması, protein kinaz bağlanması
M31523_at	TCF3	Transcription factor 3	DNA bağlanması, transkripsiyon faktör aktivitesi, transkripsiyon baskılayıcı aktivitesi, E-box bağlanması, RNA polimeraz II proksimal promoter sekansına spesifik DNA bağlanması, bHLH transkripsiyon faktörü bağlanması, güçlendirici bağlanması, mitojenle aktive edilen protein kinaz bağlanması, protein bağlanması, protein heterodimerizasyonu, protein homodimerizasyonu
X62654_rna1_at	CD63	CD63 molecule	Protein bağlanması

ZYX geni yapılan birçok çalışmada bir onkogen olarak nitelendirilmiş ve Lösemi ile ilişkili olduğu ortaya koyulmuştur (Golub vd., 1999; Zhong vd., 2019).

Vargova vd. (2016) yaptıkları çalışmada MARCKSL genini inceleyerek elde ettikleri bulgular sonucunda bu genin onkojenik olduğunu ayrıca lenfoma ve lösemi hastalıkları için önemli bir biomarker olduğunu belirtmişlerdir. Yapılan farklı çalışmalarda da MARCKSL geninin Lösemi hastalığı ile bağlantılı olduğu paylaşılmıştır (Gutiérrez, 2007; Meerloo vd., 2015; Franke vd., 2016). APLP2 geninin genel olarak kanser hücrelerinde aşırı ekspresyon (Overexpressed) gösterdiği ve tümör hücrelerindeki artışla ilişkili olduğu ayrıca APLP grubundaki genlerin Lösemi hastalığının AML türüne ait hücre göçündeki (Cell Migration) artışta rolü olduğu daha önce paylaşılmıştır (Jiang vd., 2013;Pandey vd., 2016). Kracmarova vd. (2009) yapmış oldukları çalışmada yüksek gen ekspresyon değerine sahip NME4 geninin ileri seviyelerdeki AML kanseri ile bağlantılı olduğunu tespit etmişler ve NME grubundaki genlerin tümör seviyesinin ilerlemesinde rol aldığını vurgulamışlardır. CD33 genine ait

çalışmalar incelendiğinde AML'nin tedavisi için bu genin hedef alınarak yapıldığı birçok bilimsel çalışma bulunduğu görülmektedir (Bernstein, 2002; Walter vd., 2012; O'Hear vd., 2015). Bu çalışmalarda ayrıca AML hastalarının %90'ında bu gene ait yüksek ekspresyon değeri ölçüldüğü vurgulanmıştır. CCND3 geni için yapılan bilimsel çalışmalar incelendiğinde Lösemi hastalığının birçok türünde bu gende meydana gelen mutasyonların görüldüğü anlaşılmaktadır (Matsuo vd., 2018; Smith, vd., 2005). Lin vd. (2018) yaptıkları 20 yıla yayılan çalışmalarında TCF3 geninde meydana gelen gen düzenlemelerinin ALL hastalarının hayatta kalma oranında olumlu bir katkı sağladığını tespit etmişlerdir. CD63 geninin ALL hastalığına sahip hücrelerde, sağlıklı hücrelere kıyasla yüksek ekspresyon değerine sahip olduğu yapılan bir çalışmada gözlemlenmiştir (Mirkowska vd., 2013).

Sonuç olarak geliştirilen yaklaşım ile en çok seçilen genlerin tamamının onkogen olduğu ayrıca lösemi hastalığının oluşumunda, tespit ya da tedavi edilmesinde etkin bir rol oynadığı daha önceki yapılan çalışmalarla da doğrulanmıştır.

5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında mikro dizi verilerinin üzerinde ön işlem, gen seçimi, sınıflandırma ve diğer analizlerin yapılabilmesi için web tabanlı bir arayüz tasarımı gerçekleştirilmiştir. Dünyanın her yerinden araştırmacıların sadece internet bağlantısına sahip bir cihaz ile bağlanarak kendi veri kümeleri üzerinde farklı algoritmaları kullanarak önişlem, filtreleme, gen seçimi ve sınıflandırma analizleri yapabileceği bir çalışma ortamı oluşturulmuştur. Buna ek olarak lösemi verilerinin içerisindeki anlamlı genlerin filtrelenmesi, veri boyutunun indirgenmesi, gen seçimi işlemlerinin gerçekleştirilmesi ve bu veri kümesine ait örneklerin başarılı bir şekilde sınıflandırılabilmesi için yeni bir Birlik-Hibrit gen seçim algoritması tasarlanmıştır. Geliştirilen algoritma filtreleme ve gen seçimi olmak üzere iki adımdan oluşmaktadır. İlk adımda Fisher korelasyon skoru, Willcoxon rütbeler toplamı ve Bilgi kazanımı algoritmalarından oluşan bir Birlik algoritma tasarlanarak gen filtreleme işlemi gerçekleştirilmektedir. İkinci adımda ise güçlendirilmiş bir genetik algoritma kullanılarak filtrelenmiş genlerin içerisinden en başarılı genler seçilmektedir.

Kanser araştırmalarında mikro dizi verileri önemli bir yere sahiptir. Mikro dizi verilerinin binlerce genin ekspresyon değerine sahip olması, kanser ile gen ilişkilerinin ortaya çıkarılmasında büyük bir avantaj sağlamaktadır. Ancak bu avantajın kullanılabilmesi için çok büyük bir gen havuzu içerisinden kanser ile ilişkili olan genlerin doğru bir şekilde seçilebilmesi gerekmektedir. Bu noktada gen filtreleme yaklaşımları hızlı bir çözüm önerisi getirmektedir. İstatistiksel olarak anlamlı genler bu yöntemler ile puanlanarak en yüksek puana sahip olan genlerin kanser ile ilişkili olduğu çıkarımı yapılmakta ve bu genler seçilmektedir. Bu adımda düşük puana sahip ancak kanser ile ilişkisi olan genler elenebilmektedir. Birlik yöntemler ile farklı puanlama yöntemlerine sahip algoritmalar bir arada kullanarak, seçilen gen havuzu tüm anlamlı genleri kapsayacak şekilde genişletilebilmektedir. Bu durumda seçilen gen sayısı ve yeterince anlam taşımayan genlerin seçilme ihtimali artmaktadır. Bunun sonucunda da veri kümesi içerisinde sınıflandırma başarısını olumsuz etkileyen bir gürültü oluşmaktadır. Bu problemin üstesinden gelebilmek için ikinci bir

adım olarak sarmal yaklaşımlar önerilse de veri kümesinin içerisindeki istenmeyen genler ve veri kümesi boyutu sarmal yaklaşımların başarısını olumsuz etkilemektedir. Bu sorunun çözümü için daha önceki çalışmalardan farklı olarak, filtreleme adımında belirlenen genlerin önem derecesinin belirlenmesi için yeni bir hesaplama yöntemi önerilmiştir. Gen seçim adımında kullanılan genetik algoritmanın çözüm kümesini araştırırken bu önem derecesini dikkate alabilmesi için yeni bir mutasyon işlevi geliştirilmiştir. Böylelikle gürültünün en aza indirilerek, sınıflandırma başarısının daha az gen ile sağlanması hem de genetik algoritmanın daha az işlem adımı ile sonuca ulaşması sağlanmıştır. Çalışma içerisinde önerilen güçlendirilmiş genetik algoritma gen seçim işlemlerini ortalama 173,95 iterasyonda, %100 başarı oranı ile gerçekleştirmeyi başarmıştır.

Çalışmanın odaklandığı bir başka konu ise genetik algoritmanın uygunluk fonksiyonun belirlenmesidir. Uygunluk fonksiyonunun belirlenmesi gen seçim işlemlerinin başarılı bir şekilde gerçekleşmesi için kritik bir öneme sahiptir. Genlerin sahip olduğu anlamı doğru ölçemeyen bir uygunluk fonksiyonu başarılı bir gen seçim işlemi gerçekleştiremeyecektir. Sadece bir sınıflandırma algoritması ile elde edilen sınıflandırma başarısı, genetik algoritmanın uygunluk değerinin hesaplanmasında yaygın olarak kullanılmaktadır. Sadece test işlemi gerçekleştirilen bu yöntemde, sınıflandırma algoritmasının veri kümesini ezberlemesi problemi (overtraining) ile sıklıkla karşılaşılabilmektedir. Bu sorunun önüne geçilebilmesi için iki farklı işlem gerçekleştirilmiştir. İlk olarak uygunluk değeri hesaplamasında tek bir sınıflandırma algoritması kullanmak yerine KNN, SVM ve Naive Bayes algoritmalarından oluşan birlik bir sınıflandırma algoritması kullanılmıştır. Ağırlıklı oylama ile sınıflandırma tahmini yapan bu algoritmanın elde etmiş olduğu başarı, uygunluk değeri hesaplanırken kullanılmıştır. İkinci olarak seçilen sınıflandırma algoritması ile tek bir test işlemi yerine LOOCV çapraz doğrulama işlemi gerçekleştirilmiştir. Bu sayede geliştirilen veri kümesini ezberleme ihtimaline karşı dayanıklı olması sağlanmıştır.

Tez kapsamında kullanılan Lösemi kanseri veri kümesi üzerinde yapılan işlemler sonucunda elde edilen bulgular incelendiğinde, geliştirilen algoritmanın literatürde önerilen yaklaşımlardan birçok kıstasta daha başarılı sonuçlar elde ettiği görülmüştür. Sadece 2 gen seçimi ile Lösemi kanserinin ALL ve AML türlerine ait örneklerin %100 test ve %100 LOOCV çapraz doğrulama başarısı ile sınıflandırılabilmesi sağlanarak literatüre önemli bir katkı sağlanmıştır. Ayrıca algoritma tarafından en sık sayıda seçilen genler ("ZYX, MARCKSL1, APLP2, NME4, CD33, CCND3, TCF3") tıp literatüründe daha önce yapılan çalışmalar ile karşılaştırıldığında lösemi kanserinin oluşumunda, teşhis, tedavi ve ilaç geliştirme aşamalarında bu genlerin etkin bir role sahip olduğunu doğrulamaktadır. Geliştirilen algoritma gen-kanser ve gen-gen ilişkilerinin ortaya çıkarılmasında ki bu başarısı ile tıp literatürüne de önemli bir katkı sağlanmıştır.

Mikro dizi veri türü her ne kadar bir büyük veri türü olarak görülse de aslında bu verilerin sahip olduğu yüksek karmaşıklık ve az sayıda örnek, bu veri kümelerini farklı bir kategoriye sokmaktadır. Her bir mikro dizi veri kümesinin karakteristik özelliğinin farklı olması tüm mikro dizi veri kümelerinde başarılı sonuçlar elde eden genel bir algoritma geliştirilmesini güçleştirmektedir. Buna rağmen geliştirilen algoritma Merkezi Sinir Sistemi Tümörü veri kümesinde 10 gen ile %98,33, Kolon Kanseri veri kümesinde 5 gen ile %95,16, Sars-Cov-2 veri kümesinde 10 gen ile %93,7, Prostat Kanseri veri kümesinde 5 gen ile 96,08 ve Yumurtalık Kanseri sadece 2 gen ile %100 LOOCV başarısı elde etmiştir. Elde edilen yüksek performans değerleri göz önünde bulundurulduğunda geliştirilen algoritmanın farklı özelliklere, gen ve örnek sayılarına sahip veri kümelerinde başarılı bir şekilde çalıştığı görülmektedir. Bu nedenle geliştirilen algoritmanın farklı kanser türlerine ait verilerde gen-kanser ve gen-gen ilişkilerinin açığa çıkarılması için yapılan çalışmalarda da kullanılması önerilmektedir.

Geliştirilen yaklaşım içerisinde kullanılan gen filtreleme yaklaşımlarının yapısı nedeni ile sadece iki sınıf parametresine sahip veri kümeleri ile çalışabilmektedir. Gelecekteki çalışmalarda farklı gen filtreleme yaklaşımlarının

algoritma içerisine dâhil edilerek geliştirilen algoritmanın çalışabildiği veri kümesi çeşitliliğinin artırılması amaçlanmaktadır. Eklenen algoritmaların arayüze dâhil edilmesi ile birçok farklı kombinasyonun denenebilmesi ve en iyi yöntemin seçilebilmesi için daha fazla seçeneğe sahip bir çalışma ortamı oluşturulması amaçlanmaktadır.

Birden fazla algoritmanın ve tekrarlı yaklaşımların bir arada kullanılması, veri kümesinin sahip olduğu yüksek karmaşıklık gen seçim işlemlerinin süresini oldukça uzatmaktadır. Her ne kadar gen seçimi sonrasında indirgenen veri boyutu sınıflandırma işlemlerinin süresini büyük ölçüde azaltsa da gelecekte yapılacak çalışmalarda gen seçimi işlemlerinde harcanan süresinin iyileştirilmesi de planlanmaktadır.

KAYNAKLAR

Adiwijaya, , Wisesty, U. N., Lisnawati, E., Aditsania, A., Kusumo, D. S., 2018. Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification. Journal of Computer Science, 14(11), 1521-1530.

Algamal, Z. Y., Lee, M. H., 2019. A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification. Advances in Data Analysis and Classification (13), 753-771.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Cell Biology, 96, 6745-6750.

Alshamlan, H. M., Badr, G. H., Alohali, Y. A., 2015. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational Biology and Chemistry, (56), 49-60.

Azzawi, H., Hou, J., Xiang, Y., Alanni, R., 2016. Lung cancer prediction from microarray data by gene expression programming. IET Systems Biology, 5(10), 168-178.

Baliarsingh, S. K., Ding, W., Vipsita, S., Bakshi, S., 2019. A memetic algorithm using emperor penguin and social engineering optimization for medical data classification. Applied Soft Computing, 105773(85), 14.

Barbulovic-Nad, I., Lucente, M., Sun, Y., Zhang, M., Wheeler, A. R., Bussman, M., 2006. Bio-Microarray Fabrication Techniques—A Review. Critical Reviews in Biotechnology, 26(4), 237-259.

Beasley, D., Bull, D., R., Martin, R., R., 1993. An Overview of Genetic Algorithms: Part 1, Fundamentals. University Computing, 15(4), 170-181.

Bernstein, I. D., 2002. CD33 as a Target for Selective Ablation of Acute Myeloid Leukemia. Clinical Lymphoma, 1(2), 9-11.

Blanton, H. (1997). An introduction to neural networks for technicians, engineers and other non PhDs. Proceedings of the 1997 Artificial Neural Networks in Engineering Conference. 8-10 October, St. Louis

Bolon-Canedo, V., Alonso-Betanzos, A., 2019. Ensembles for feature selection: A review and future trends. Information Fusion, (52), 1-12.

Cortes, C., Vapnik, V. N., 1995. Support-vector Networks. Machine Learning, 20(3), 273-297.

Cover, T., M., Hart, P., E., 1967. Nearest neighbor patter classification. IEEE Transactions on Information Theory, 13(1), 21-27.

Cui, Y., Zheng, C.-H., Yang, J., Sha, W., 2013. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Computers in Biology and Medicine, 7(43), 933-941.

Dash, M., Liu, H., 1997. Feature selection for classification. Intelligent Data Analysis, 3(1), 131-156.

Dashtban, M., Balafar, M., Suravajhala, P., 2018. Gene selection for tumor classification using a novel bio-inspired multiobjective approach. Genomics(110), 10-17.

Davis, A. S., Viera, A. J., Mead, M. D., 2014. Leukemia: An Overview for Primary Care. American Family Physician, 9(89), 731-738.

Dettling, M., 2004. BagBoosting for tumor classification with gene expression data. Bioinformatics (20), 3583-3593.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. Wiley Interscience, 688p, New York.

European Bioinformatic Institution, 2019. Gene Ontology Annotation (GOA). Access Date: 12.01.2019. https://www.ebi.ac.uk/GOA/

Ewart, K. V., Belanger, J. C., Williams, J., Karakach, T., Penny, S., Tsoi, S. C., 2005. Identification of genes differentially expressed in Atlantic salmon (Salmo salar) in response to infection by Aeromonas salmonicida using cDNA microarray technology. Developmental and Comparative Immunology, 29, 333-347.

Fatonah, N. S., Tjandrasa, H., Fatichah, C., 2018. Automatic Leukemia Cell Counting using Iterative Distance Transform for Convex Sets. International Journal of Electrical and Computer Engineering, 8(3), 1731-1740.

Franke, N. E., Kaspers, G. L., Assaraf, Y. G., Meerloo, J., 2016. Exocytosis of polyubiquitinated proteins in bortezomib-resistant leukemia cells: a role for MARCKS in acquired resistance to proteasome inhibitors. Oncotarget, (7), 74779.

Fu, L. M., Fu-Liu, C. S., 2005. Evaluation of gene importance in microarray data based upon probability of selection. BMC Bioinformatics, 67(6), 11.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, (16), 906-914.

Gao, L., Ye, M., Lu, X., Huang, D., 2017. Hybrid Method Based on Information Gain and Support Vector Machine for Gene Selection in Cancer Classification. Genomics Proteomics Bioinformatics, (15), 389-395.

Geary, C. G., 2000. The story of chronic myeloid leukaemia. Br J Haematol, 110(1), 2-11.

Ghosh, M., Adhikary, S., Ghosh, K. K., Sardar, A., Begum, S., Sarkar, R., 2019. Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & Biological Engineering & Computing, (57), 159-176.

GLOBOCAN, WHO., 2019a. All Cancers 2018, International Agency for Research on Cancer.

GLOBOCAN, WHO., 2019b. Leukaemia, International Agency for Research on Cancer.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 5439(286), 531-537.

Gu, Q., Li, Z., Han, J., 2012. Generalized fisher score for feature selection. arXiv, In Press.

Gutiérrez, N. O., 2007. Gene expression profiling of B lymphocytes and plasma cells from Waldenström's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals. Leukemia, (21), 541-549.

Haykin, S., 1999. Neural Networks: A comprehensive foundation. Prentica Hall, 842p, New Jersey.

Hesper, B., Hogeweg, P., 1970. Bioinformatica: een werkconcept. Kameleon, 1(6), 28-29.

Holland, J. H., 1975. Adaptation in Natural and Artificial Systems. MIT Press, 232p, Massachusetts.

Huang, J., Cai, Y., Xu, X. 2007. A hybrid genetic algorithm for feature selection wrapper based on mutual information. Patter Recognition Letters, 13(28), 1825-1844.

Hugo Gene Names, 2019. Gene Names. Access Date :12.01.2019 https://www.genenames.org

Jain, I., Jain, V. K., Jain, R., 2018. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Applied Soft Computing, (62), 203-215.

Jain, S., Shukla, S., Wadhvani, R., 2018. Dynamic selection of normalization techniques using data complexity measures. Expert Systems With Applications, 106, 252-262.

Jiang, L., Yu, G., Meng, W., Wang, Z., Meng, F., Ma, W., 2013. Overexpression of amyloid precursor protein in acute myeloid leukemia enhances extramedullary infiltration by MMP-2. Tumour Biology, (34), 629-636.

Jolliffe, I., T., 2020. Principal Component Analysis, Second Edition. Springer, 487p, New York.

Kar, S., Sharma, K. D., Maitra, M., 2015. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications, 1(42), 612-627.

Karakach, T. K., Flight, R. M., Douglas, S. E., Wemtzell, P. D., 2010. An introduction to DNA microarrays for gene expression analysis. Chemometrics and Intelligent Laboratory Systems, 104(1), 28-52.

Kim, K-J., Cho, S-B., 2008. An Evolutionary Algorithm Approach to Optimal Ensemble Classifiers for DNA Microarray Data Analysis. IEEE Transactions on Evolutionary Computation, 3(12), 377-388.

Kracmarova, A., Cermak, J., Brdicka, R., Bruchova, H., 2009. High expression of ERCC1, FLT1, NME4 and PCNA associated with poor prognosis and advanced stages in myelodysplastic syndrome. Leukemia & Lymphoma, 7(49), 1297-1305.

Kumar, G. P., Victoire, T. A.-A., Renukadevi, P., Devaraj, D., 2012. Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. Expert Systems with Applications, (39), 1811-1822.

Lee, C-P., Leu, Y., 2011. A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing, (11), 208-213.

Li, S., Wu, X., Hu, X., 2008. Gene selection using genetic algorithm and support vectors machines. Soft Computing, 7(12), 693-698.

Liao, C., Li, S., Luo, Z., 2006. Gene Selection for Cancer Classification using Wilcoxon Rank Sum Test and Support Vector Machine. International Conference on Computational Intelligence and Security, 3-6 November, Guangzhou, 368-373.

Lin, A., Cheng, F. W., Chiang, A. K., Luk, C-W., Li, R. C., 2018. Excellent outcome of acute lymphoblastic leukaemia with TCF3-PBX1 rearrangement in Hong Kong. Pediatric Blood & Cancer, 12(65), 5.

Liu, H., Liu, L., Zhang, H., 2010. Ensemble gene selection for cancer classification. Pattern Recognition, (43), 2763-2772.

Liu, K-H., Li, B., Zhang, J., Du, J-X. 2009. Ensemble component selection for improving ICA based microarray data prediction models. Pattern Recogniton, 7(42).

Man, K., F., Tang, K. S., Kwong, S., 1996. Genetic Algorithms: Concepts and Applications. IEEE Transactions on Industrial Electronics, 43(5), 519-533.

Masoudi-Sobhanzadeh, Y., Motieghader, H., Masoudi-Nejad, A., 2019. FeatureSelect: a software for feature selection based on machine learning approaches. BMC Bioinformatics, 170(20), 1-17.

Matsuo, H., Yoshida, K., Fukurama, K., Nakatani, K., Noguchi, Y., Takasaki, S., 2018. Recurrent CCND3 mutations in MLL-rearranged acute myeloid leukemia. Blood Advances, 21(2), 2879-2889.

Meerloo, J. V., Niewerth, D., Horton, T. M., Chng, W. J., Bi, C., Menezes, R. X., 2015. Marcks Marks Resistance to Proteasome Inhibitors: Exocytosis of Polyubiquitinated Proteins in Bortezomib-Resistant Leukemia Cells. Blood, 23(126), 3712.

Mick, E., Kamm, J., Pisco, A. O., Ratnasiri, K., Babik, J. M., Calfee, C. S., 2020. Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2. In Press

Mirkowska, P., Hofmann, A., Sedek, L., Slamova, L., Mejstrikova, E., Szczepanski, M., 2013. Leukemia surfaceome analysis reveals new disease-associated features. Blood, 25(121), 149-159.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, 432p, New York.

Mohamad, M. S., Omatu, S., Deris, S., Yoshioka, M., 2011. A Modified Binary Particle Swarm Optimization for Selecting the Small Subset of Informative Genes From Gene Expression Data. IEEE Transactions on Information Technology in Biomedicine, 6(15), 813-822.

Mollaee, M., Moattar, M. H., 2016. A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybernetics and Biomedical Engineering, 3(36), 521-529.

Motieghader, H., Najafi, A., Sadeghi, B., Masoudi-Nejad, A., 2017. A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Informatics in Medicine Unlocked, (9), 246-254.

O'Hear, C., Heiber, J. F., Schubert, I., Fey, G., Geiger, T. L. 2015. Anti-CD33 Chimeric Antigen Receptor Targeting Of Acute Myeloid Leukemia. Haematologica, (100), 336-344.

Pandey, P., Sliker, B., Peters, H. L., Tuli, A., Herskovitz, J., 2016. Amyloid precursor protein and amyloid precursor-like protein 2 in cancer. Oncotarget, 15(7), 19430-19444.

Peng, S., Xu, Q., Ling, X. B., Peng, X., Du, W., Chen, L., 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Letters, 2(555), 358-362.

Petricoin III, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., 2002. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 359, 572-577.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., 2002. Prediction of central nervous system embryonal tumor outcome based on gene expression. Letter to Nature, 415, 436-442.

Ruiz, R., Riquelme, J. C., Aguilar-Ruiz, J. S., Garcia-Torres, M., 2012. Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches. Expert Systems with Applications, (39), 11094-11102.

Salem, H., Attiya, G., El-Fishawy, N., 2017. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, (50), 124-134.

Scholz, M., 2006. Approaches to analyse and interpret biological profile data. University of Potsdam, Max Planck Institute of Molecular Plant Physiology, Ph.D. Thesis, 93p, Potsdam, Germany.

Shannon, C. E., 1948. A Mathematical Theory of Communication. (27), 623-656.

Sharbaf, F. V., Mosafer, S., Moattar, M. H., 2016. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics, (107), 231-238.

Shen, Q., Shi, W. M., Kong, W., 2008. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Computational Biology and Chemistry, (32), 53-60.

Shim, J., Sohn, I., Kim, S., Lee, J. W., Green, P. E., Hwang, C., 2009. Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine. Computational Statistics & Data Analysis, 5(53), 1736-1742.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2), 203-209.

Smith, M. L., Arch, R., Smith, L-L., Bainton, N., Neat, M., Taylor, C., 2005. Development of a human acute myeloid leukaemia screening panel and consequent identification of novel gene mutation in FLT3 and CCND3. British Journal of Haematology, (128), 318-323.

T.C. Sağlık Bakanlığı, 2019. Türkiye Kanser İstatistikleri 2016, Ankara T.C. Sağlık Bakanlığı Halk Sağlığı Genel Müdürlüğü.

Türkiye İstatistik Kurumu (TÜİK), 2019. Ölüm Nedeni İstatistikleri. Erişim Tarihi :06.05.2020. http://tuik.gov.tr/PreHaberBultenleri.do?id=30626

Vargova, J., Vargova, K., Dusilkova, N., Kulvait, V., Pospisil, V., Zavadil, J., 2016. Differential expression, localization and activity of MARCKS between mantle cell lymphoma and chronic lymphocytic leukemia. Blood Cancer Journal, 475(6).

Walter, R. B., Appelbaum, F. R., Estey, E. H., Bernstein, I. D., 2012. Acute myeloid leukemia stem cells and CD33-targeted immunotherapy. Blood, 26(119), 6198-6208.

Wang, K. W., Wang, W. J., Li, M. A., 2018. A brief procedure for big data analysis of gene expression. Animal Models and Experimental Medicine, 1(8), 189-193.

Wang, Y., Chen, X., Jiang, W., Li, L., Li, W., Yang, L., 2011. Predicting human microRNA precursors based on an optimized feature subset generated by GA–SVM. Genomics, 2(98), 73-78.

Wang, Y., Yang, X.-G., Lu, Y., 2019. Informative gene selection for microarray classification via adaptive elastic net with conditional mutual information. Applied Mathematical Modelling, (71), 286-297.

Xiong, M., Fang, X., Zhao, J., 2001. Biomarker identification by feature wrappers. Genome Res, 11(11), 1878-1887.

Xu, J., Mu, H., Wang, Y., Huang, F., 2018. Feature Genes Selection Using Supervised Locally Linear Embedding and Correlation Coefficient for Microarray Classification. Computational and Mathematical Methods in Medicine, 11.

Yang, C-S., Chuang, L-Y., Ke, C-H., Yang, C-H., 2008. A hybrid feature selection method for microarray classification. International Journal Of Computer Science, (35), 285-290.

Yang, K., Cai, Z., Li, J., Lin, G., 2006. A stable gene selection in microarray data analysis. BMC Bioinformatics, 228(7), 16.

Yu, H., Gu, G., Liu, H., Shen, J., Zhao, J., 2009. A Modified Ant Colony Optimization Algorithm for Tumor Marker Gene Selection. Genomics, Proteomics & Bioinformatics, 4(7), 200-208.

Zainuddin, Z., Ong, P., 2011. Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. Expert Systems with Applications, (38), 13711-13722.

Zhong, C., Yu, J., Li, D., Jiang, K., Tang, Y., Yang, M., 2019. Zyxin as a potential cancer prognostic marker promotes the proliferation and metastasis of colorectal cancer cells. Journal of Cellular Physiology, 9(234), 15775-15789.

EKLER

EK A. Geliştirilen algoritmanın farklı veri kümelerinde elde ettiği sonuçlar

EK A. Geliştirilen algoritmanın farklı veri kümelerinde elde ettiği sonuçlar

Bu kısımda geliştirilen algoritmanın Merkezi Sinir Sistemi Tümörü (MSST) (Pomeroy vd., 2002), Kolon Kanseri (Alon vd., 1999), Sars-Cov-2 (Mick vd., 2020), Prostat Kanseri (Singh vd., 2002), Yumurtalık Kanseri (Petricoin vd., 2002) veri kümelerine ait gen seçim işlemleri sonucunda elde ettiği sonuçlar sunulmaktadır (Çizelge EK A.1).

Çizelge EK A.1. Geliştirilen algoritmanın diğer kümelerde elde ettiği performans değerleri

Veri Kümesi	Toplam Gen Sayısı	Seçilen Gen Sayısı	Örnek Sayısı	Başarılı Sınıflandırma	LOOCV		
					En Yüksek (%)	Ortalama (%)	En Düşük (%)
Merkezi Sinir Sistemi Tümörü,	7126	10	60	59	98,3333	95,3333	91,6667
Kolon Kanseri	2000	5	62	59	95,1613	94,5161	93,5484
Sars-Cov-2(COVID 2019)	15900	10	238	223	93,6975	91,8907	90,7563
Prostat Kanseri	2135	5	102	98	96,0784	95,9803	95,0980
Yumurtalık Kanseri	15154	2	253	253	100	100	100

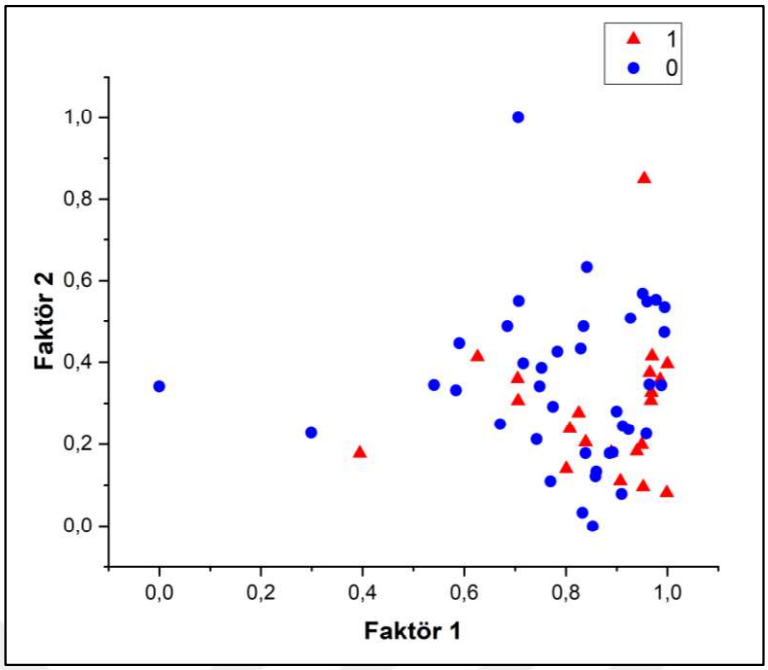
EK A.1. Merkezi sinir sistemi tümörü

Çizelge EK A.2. Filtreleme sonuçları (MSST)

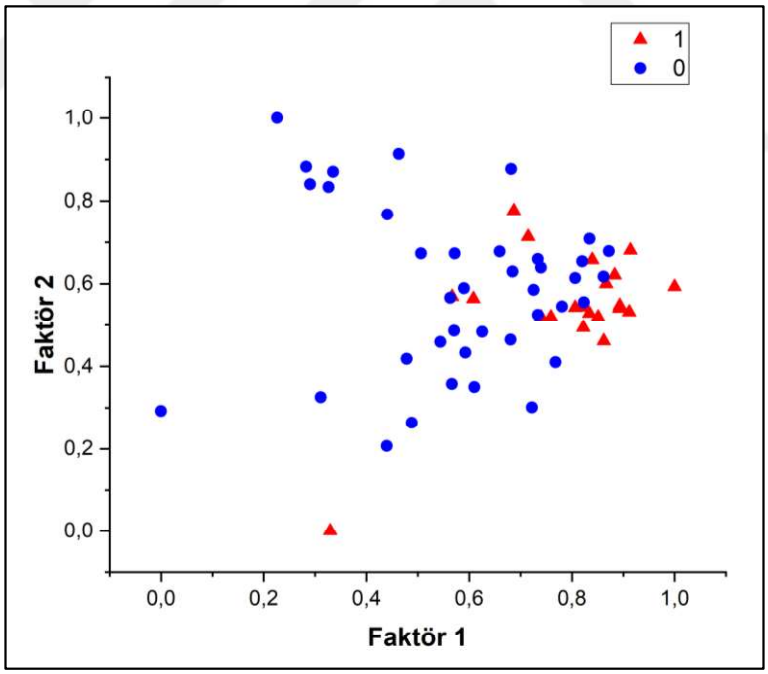
		FKS (LOOCV)			BK (LOOCV)			WRT (LOOCV)		
Gen Sayısı	**Toplam Örnek Sayısı**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**
500	60	48	80,0000	0,8242	46	76,6667	0,7875	46	76,6667	0,7766
250	60	47	78,3333	0,8114	47	78,3333	0,8004	47	78,3333	0,7894
200	60	47	78,3333	0,8004	**49**	**81,6667**	**0,8260**	48	80,0000	0,8022
150	60	47	78,3333	0,8333	47	78,3333	0,8004	46	76,6667	0,7766
100	60	46	76,6667	0,7985	46	76,6667	0,7875	45	75,0000	0,7527
75	60	49	81,6667	0,8370	46	76,6667	0,7875	45	75,0000	0,7527
50	60	50	83,3333	0,8608	47	78,3333	0,8004	45	75,0000	0,7527
25	60	51	85,0000	0,8516	44	73,3333	0,7399	45	75,0000	0,7308
10	60	44	73,3333	0,7509	48	80,0000	0,8022	49	81,6667	0,7821
9	60	49	81,6667	0,8370	46	76,6667	0,7656	45	75,0000	0,7088
8	60	48	80,0000	0,8132	47	78,3333	0,7674	44	73,3333	0,6850
7	60	47	78,3333	0,7894	48	80,0000	0,8022	48	80,0000	0,7473
6	60	**51**	**85,0000**	**0,8516**	46	76,6667	0,7766	46	76,6667	0,7106
5	60	45	75,0000	0,7308	41	68,3333	0,6795	49	81,6667	0,7821
4	60	42	70,0000	0,6593	43	71,6667	0,7161	48	80,0000	0,7363
3	60	44	73,3333	0,6960	46	76,6667	0,7656	51	85,0000	0,8077
2	60	40	66,6667	0,6227	42	70,0000	0,6703	**49**	**81,6667**	**0,7711**

Çizelge EK A.3. Seçilen genlerin performans değerleri (MSST)

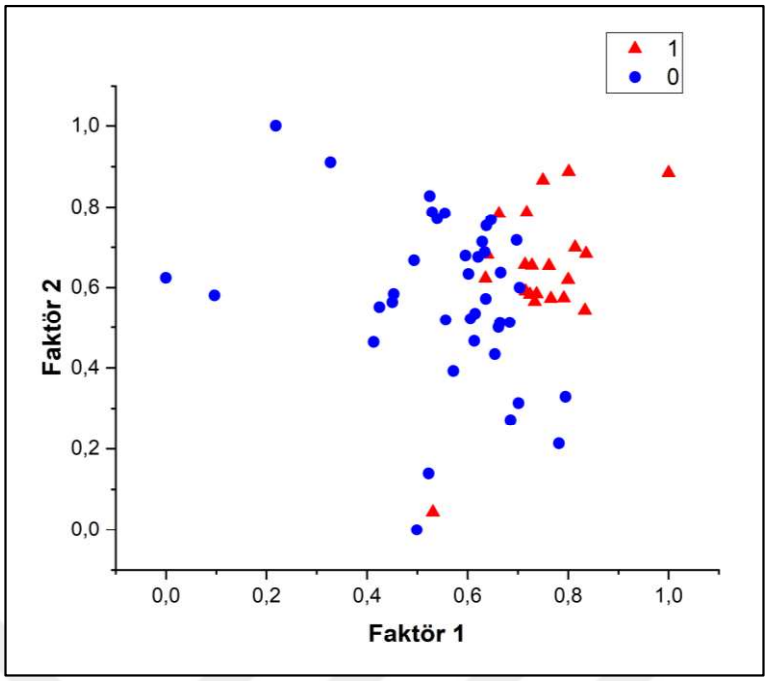
		LOOCV			K 5 (%)			K 10 (%)		
Gen Sayısı	**Toplam Örnek Sayısı**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**	**En Başarılı**	**En Başarısız**	**Ortalama**	**En Başarılı**	**En Başarısız**	**Ortalama**
10	60	57	95	0,9396	100	75	88,3333	100	50	88,3333
10	60	56	93,3333	0,9158	100	83,3333	91,6667	100	83,3333	91,6667
10	60	56	93,3333	0,9158	100	75	88,3333	100	66,6667	91,6667
10	60	58	96,6667	0,9634	100	83,3333	93,3333	100	83,3333	95
10	60	57	95	0,9286	100	83,3333	93,3333	100	83,3333	95
10	60	55	91,6667	0,9139	100	75	88,3333	100	66,6667	88,3333
10	60	57	95	0,9396	100	91,6667	95	100	83,3333	95
10	60	59	98,3333	0,9762	100	83,3333	90	100	66,6667	95
10	60	57	95	0,9396	100	83,3333	88,3333	100	83,3333	93,3333
10	60	58	96,6667	0,9744	100	66,6667	86,6667	100	66,6667	91,6667



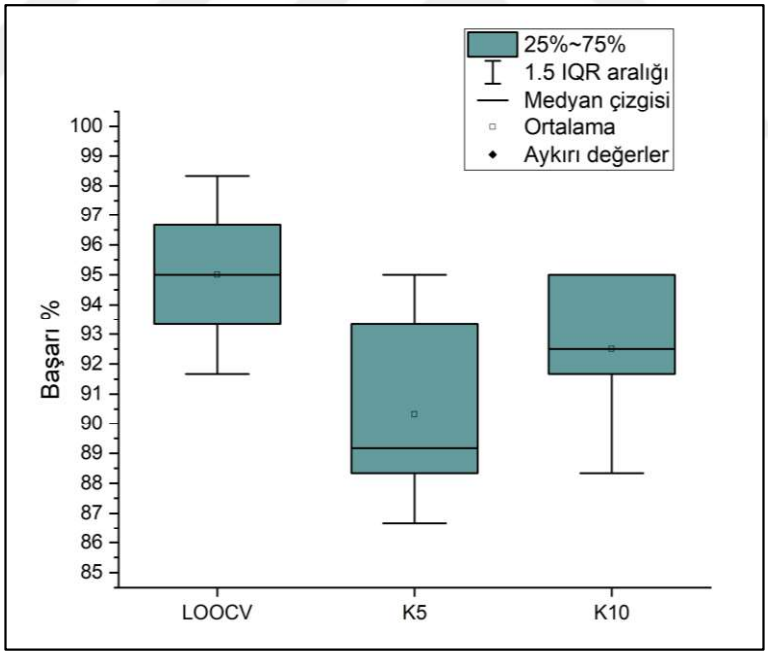
Şekil EK A.1. Tüm genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (MSST)



Şekil EK A.2. Birlik Algoritma ile filtrelenen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (MSST)



Şekil EK A.3. Seçilen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (MSST)



Şekil EK A.4. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (MSST)

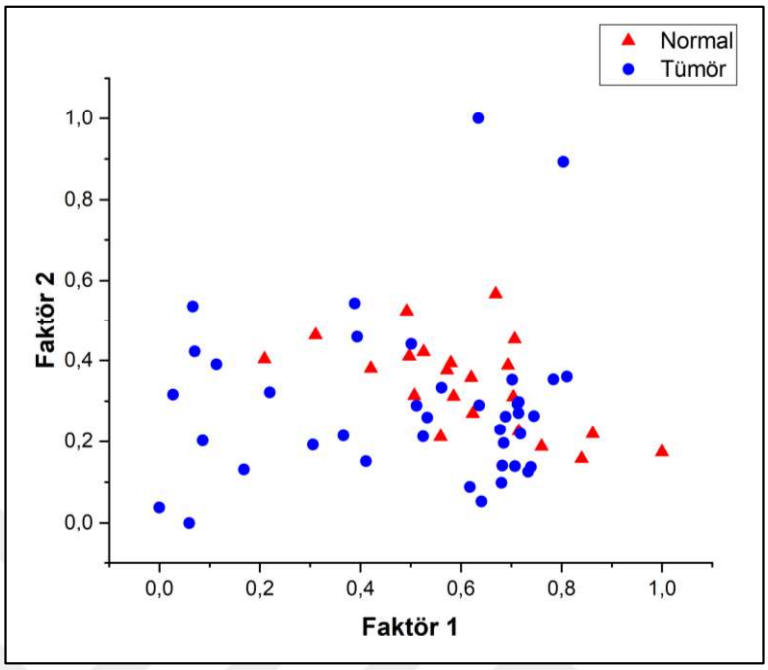
EK A.2. Kolon kanseri

Çizelge EK A.4. Filtreleme sonuçları (Kolon Kanseri)

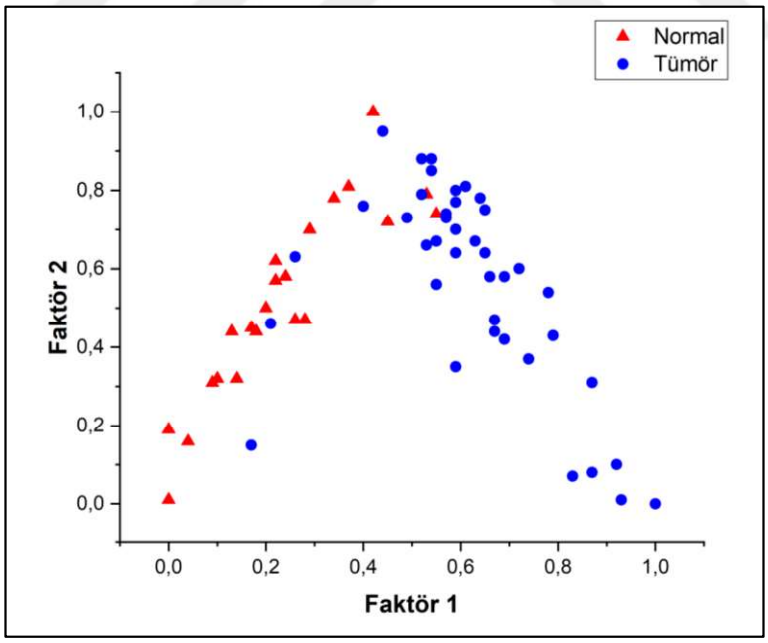
		FKS (LOOCV)			BK (LOOCV)			WRT (LOOCV)		
Gen Sayısı	Toplam Örnek Sayısı	Başarılı Sınıf.	Başarı (%)	AUC	Başarılı Sınıf.	Başarı (%)	AUC	Başarılı Sınıf.	Başarı (%)	AUC
500	62	52	83,871	0,8341	52	83,871	0,8341	53	85,4839	0,8466
250	62	53	85,4839	0,8466	52	83,871	0,8239	54	87,0968	0,8591
200	62	53	85,4839	0,8466	52	83,871	0,8239	52	83,871	0,8239
150	62	53	85,4839	0,8466	52	83,871	0,8239	53	85,4839	0,8364
100	62	54	87,0968	0,8591	52	83,871	0,8239	53	85,4839	0,8364
75	62	55	88,7097	0,8818	53	85,4839	0,8364	54	87,0968	0,8591
50	62	55	88,7097	0,8818	54	87,0968	0,8489	54	87,0968	0,8489
25	62	53	85,4839	0,8466	**54**	**87,0968**	**0,8489**	**55**	**88,7097**	**0,8716**
10	62	53	85,4839	0,8364	52	83,871	0,8034	54	87,0968	0,8591
9	62	52	83,871	0,8239	53	85,4839	0,8261	52	83,871	0,8239
8	62	52	83,871	0,8443	52	83,871	0,8034	52	83,871	0,8239
7	62	51	82,2581	0,8318	52	83,871	0,8136	54	87,0968	0,8693
6	62	55	88,7097	0,8818	53	85,4839	0,8364	53	85,4839	0,8568
5	62	55	88,7097	0,8716	53	85,4839	0,8261	52	83,871	0,8239
4	**62**	**56**	**90,3226**	**0,8943**	53	85,4839	0,8261	52	83,871	0,8239
3	62	47	75,8065	0,7614	51	82,2581	0,7909	54	87,0968	0,8693
2	62	44	70,9677	0,683	53	85,4839	0,8261	50	80,6452	0,7886

Çizelge EK A.5. Seçilen genlerin performans değerleri (Kolon Kanseri)

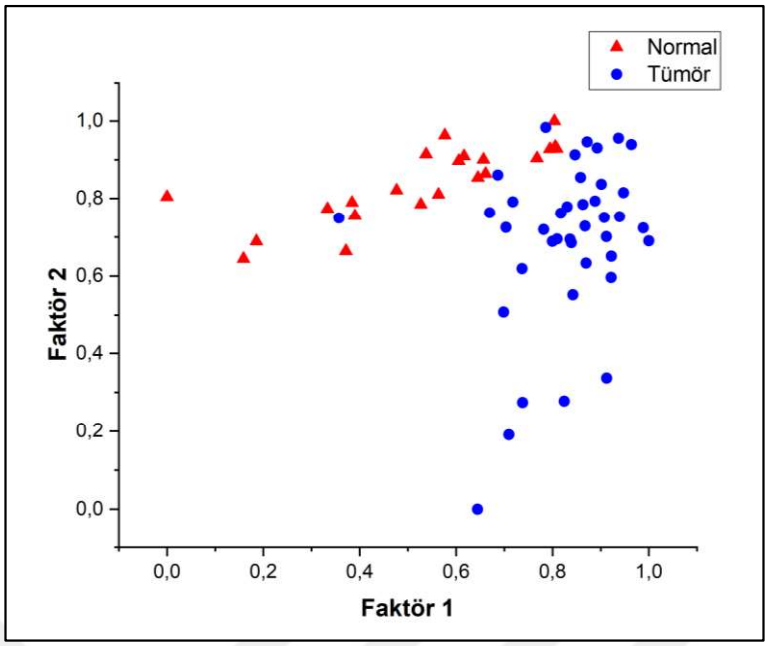
		LOOCV			K 5 (%)			K 10 (%)		
Gen Sayısı	Toplam Örnek Sayısı	Başarılı Sınıf.	Başarı (%)	AUC	En Başarılı	En Başarısız	Ortalama	En Başarılı	En Başarısız	Ortalama
5	62	58	93,5484	0,9295	100	83,3333	88,7179	100	66,6667	88,8095
5	62	58	93,5484	0,9398	100	75	90,2564	100	66,6667	90
5	62	59	95,1613	0,9625	100	69,2308	90,5128	100	50	90,2381
5	62	59	95,1613	0,9625	100	69,2308	90,5128	100	50	90,2381
5	62	58	93,5484	0,9295	100	83,3333	88,7179	100	66,6667	88,8095
5	62	58	93,5484	0,9295	100	83,3333	88,7179	100	66,6667	88,8095
5	62	59	95,1613	0,9625	100	69,2308	90,5128	100	50	90,2381
5	62	59	95,1613	0,9625	100	69,2308	90,5128	100	50	90,2381
5	62	59	95,1613	0,942	100	83,3333	93,5897	100	83,3333	93,5714
5	62	59	95,1613	0,9625	100	69,2308	90,5128	100	50	90,2381



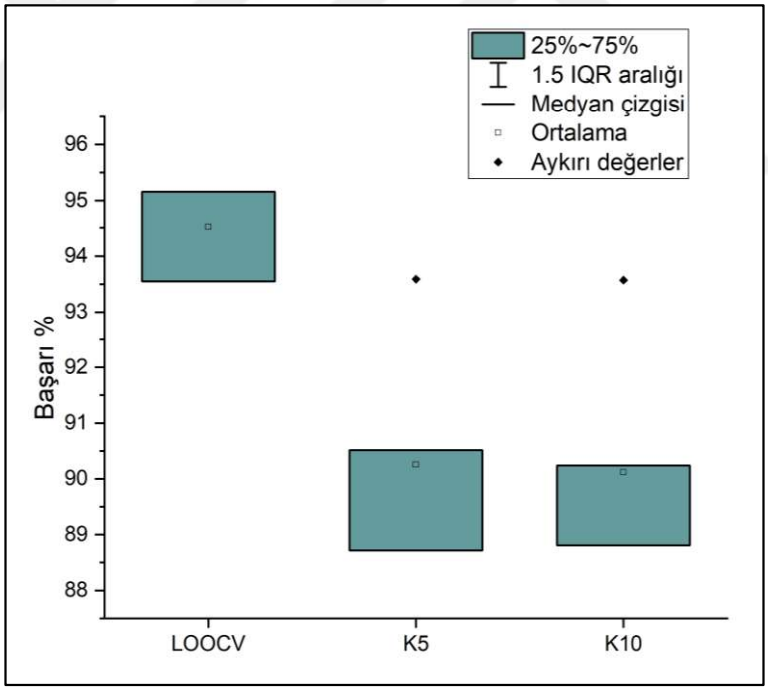
Şekil EK A.5. Tüm genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Kolon Kanseri)



Şekil EK A.6. Birlik Algoritma ile filtrelenen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Kolon Kanseri)



Şekil EK A.7. Seçilen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Kolon Kanseri)



Şekil EK A.8. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (Kolon Kanseri)

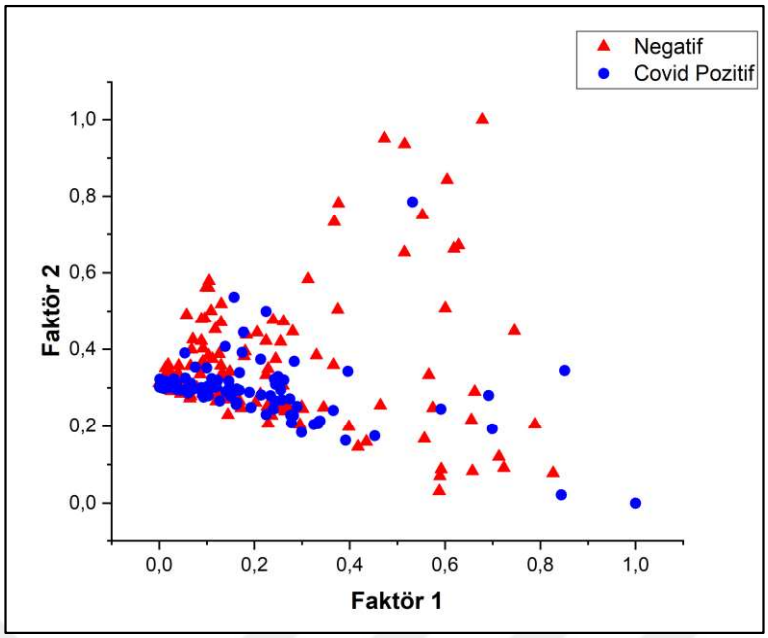
EK A.3. Sars-Cov-2 (COVID-19)

Çizelge EK A.6. Filtreleme sonuçları (Sars-Cov-2)

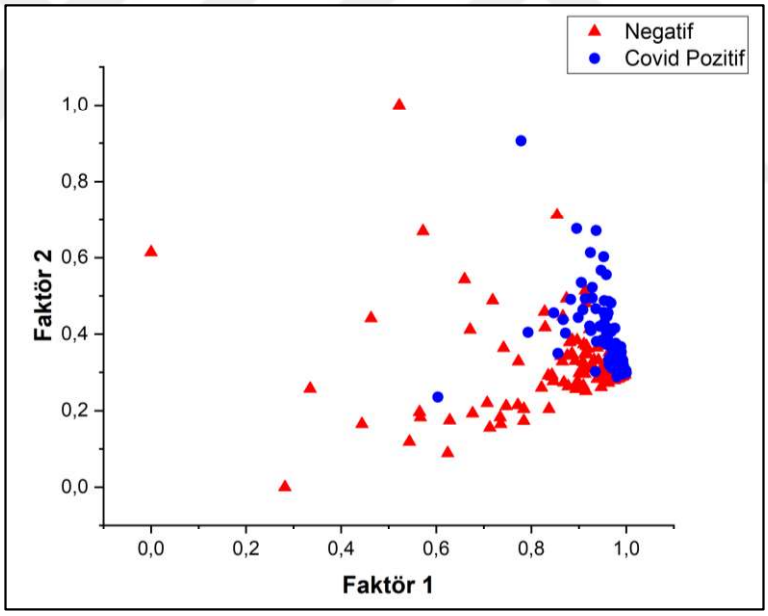
		FKS (LOOCV)			BK (LOOCV)			WRT (LOOCV)		
Gen Sayısı	**Toplam Örnek Sayısı**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**
500	238	179	75,2101	0,7619	182	76,4706	0,7575	183	76,8908	0,7665
250	238	179	75,2101	0,7656	179	75,2101	0,7416	182	76,4706	0,7483
200	238	185	77,7311	0,779	185	77,7311	0,768	**192**	**80,6723**	**0,7978**
150	238	187	78,5714	0,786	186	78,1513	0,777	188	78,9916	0,7802
100	238	189	79,4118	0,7911	195	81,9328	0,8119	186	78,1513	0,7714
75	238	**198**	**83,1933**	0,8279	197	82,7731	0,8207	184	77,3109	0,7719
50	238	161	67,6471	0,692	196	82,3529	0,8098	182	76,4706	0,7557
25	238	168	70,5882	0,7145	**199**	**83,6134**	**0,8258**	175	73,5294	0,724
10	238	163	68,4874	0,6786	194	81,5126	0,8066	178	74,7899	0,7418
9	238	161	67,6471	0,6698	192	80,6723	0,7941	187	78,5714	0,7767
8	238	156	65,5462	0,6432	197	82,7731	0,8152	180	75,6303	0,7395
7	238	162	68,0672	0,6641	195	81,9328	0,8045	187	78,5714	0,773
6	238	156	65,5462	0,6377	197	82,7731	0,8133	183	76,8908	0,7573
5	238	152	63,8655	0,6257	193	81,0924	0,7994	182	76,4706	0,7465
4	238	145	60,9244	0,5903	196	82,3529	0,8098	184	77,3109	0,7589
3	238	157	65,9664	0,6449	163	68,4874	0,6657	175	73,5294	0,724
2	238	158	66,3866	0,6373	169	71,0084	0,6921	185	77,7311	0,7624

Çizelge EK A.7. Seçilen genlerin performans değerleri (Sars-Cov-2)

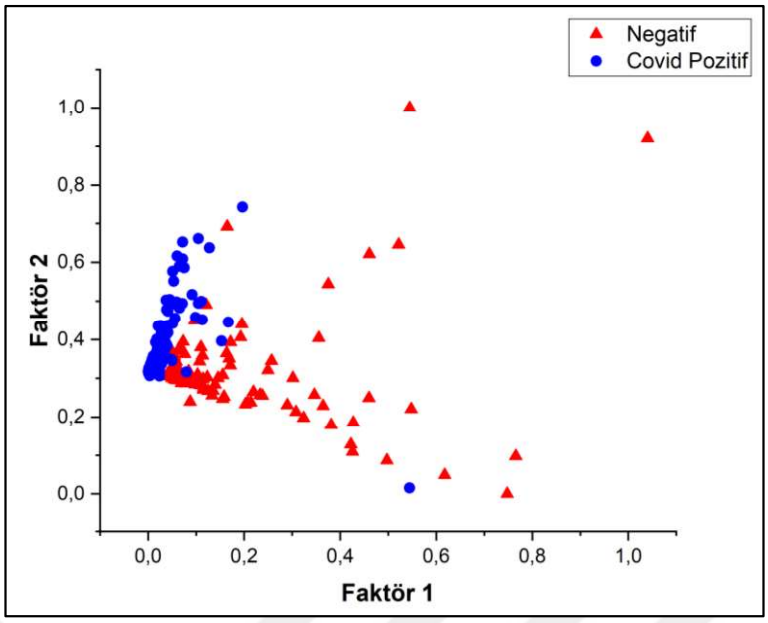
		LOOCV			K 5 (%)			K 10 (%)		
Gen Sayısı	**Toplam Örnek Sayısı**	**Başarılı Sınıf.**	**Başarı (%)**	**AUC**	**En Başarılı**	**En Başarısız**	**Ortalama**	**En Başarılı**	**En Başarısız**	**Ortalama**
10	238	218	91,5966	0,9139	93,75	85,1064	88,6348	95,8333	83,3333	91,1594
10	238	220	92,437	0,9209	93,75	80,8511	89,0337	100	78,2609	90,2717
10	238	218	91,5966	0,9121	93,617	85,1064	88,2447	100	83,3333	90,3261
10	238	216	90,7563	0,9051	91,6667	78,7234	86,1082	100	79,1667	88,2428
10	238	217	91,1765	0,9068	93,75	80,8511	87,3759	100	78,2609	89,0399
10	238	219	92,0168	0,9156	93,617	85,1064	88,6613	100	79,1667	89,8732
10	238	217	91,1765	0,9105	91,6667	85,1064	88,6525	100	79,1667	89,4928
10	238	223	93,6975	0,9368	91,6667	87,234	89,477	100	87,5000	92,029
10	238	219	92,0168	0,9137	91,6667	82,9787	89,0514	100	82,6087	90,2899
10	238	220	92,437	0,9209	91,6667	78,7234	85,6826	100	78,2609	89,4384



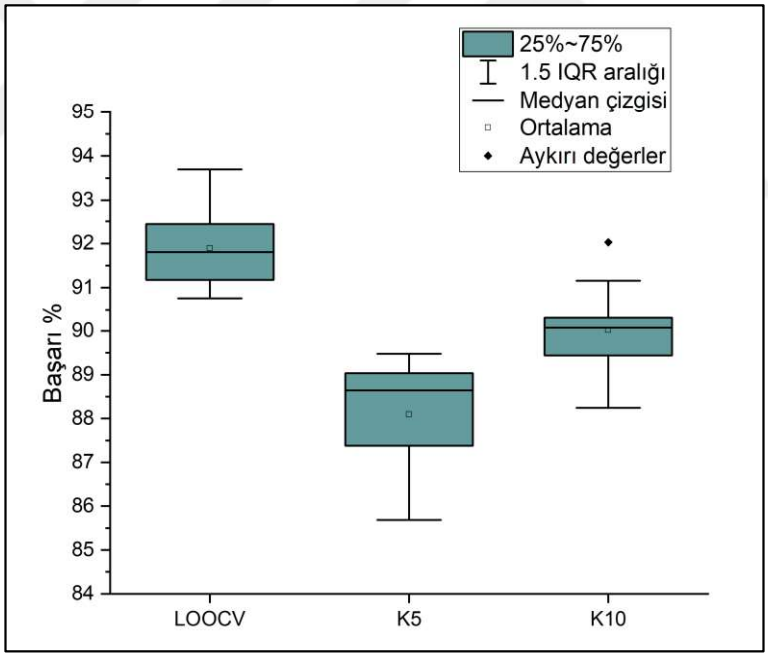
Şekil EK A.9. Tüm genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Sars-Cov-2)



Şekil EK A.10. Birlik Algoritma ile filtrelenen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Sars-Cov-2)



Şekil EK A.11. Seçilen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Sars-Cov-2)



Şekil EK A.12. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (Sars-Cov-2)

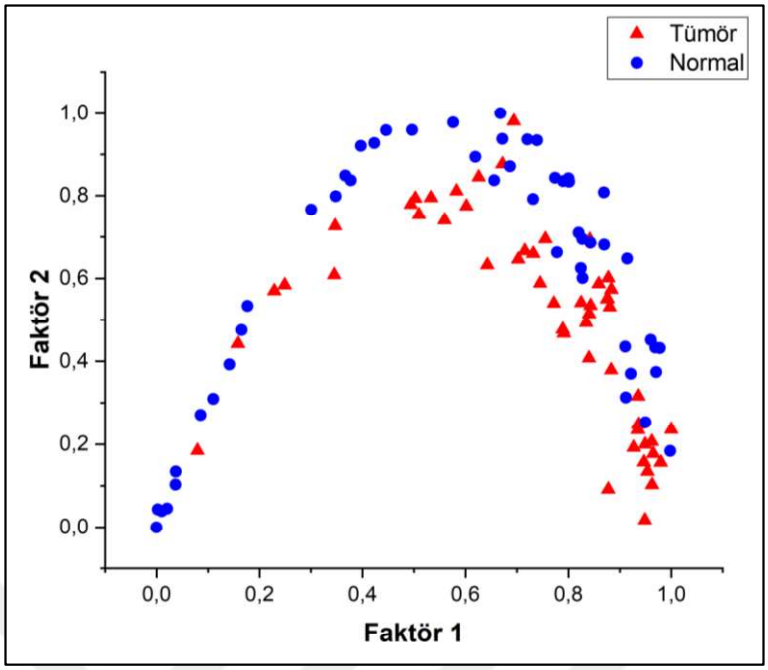
EK A.4. Prostat kanseri

Çizelge EK A.8. Filtreleme sonuçları (Prostat Kanseri)

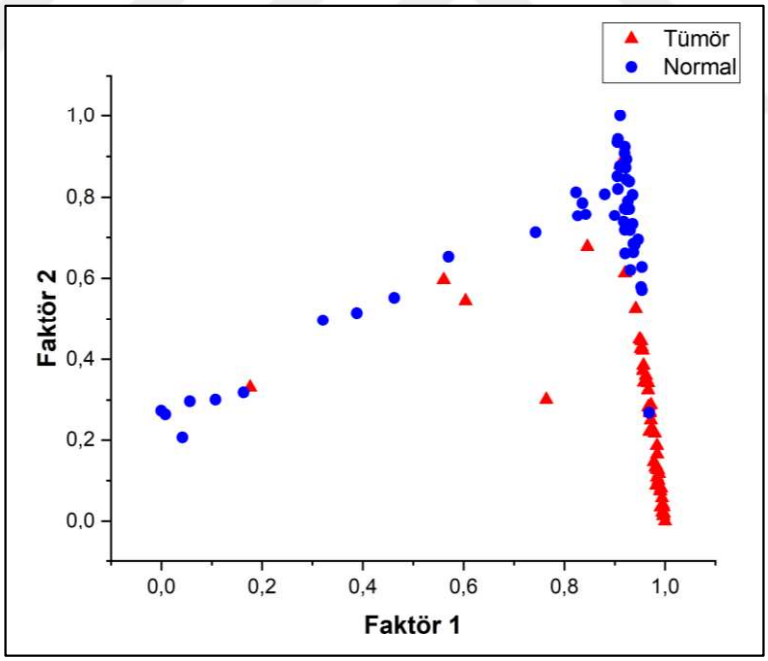
		FKS (LOOCV)			BK (LOOCV)			WRT (LOOCV)		
Gen Sayısı	Toplam Örnek Sayısı	Başarılı Sınıf.	Başarı (%)	AUC	Başarılı Sınıf.	Başarı (%)	AUC	Başarılı Sınıf.	Başarı (%)	AUC
500	102	90	88,2353	0,8827	93	91,1765	0,9123	92	90,1961	0,9031
250	102	92	90,1961	0,9019	95	93,1373	0,9319	92	90,1961	0,9027
200	102	94	92,1569	0,9223	95	93,1373	0,9319	93	91,1765	0,9127
150	102	92	90,1961	0,9023	93	91,1765	0,9123	92	90,1961	0,9027
100	102	94	92,1569	0,9219	93	91,1765	0,9123	92	90,1961	0,9027
75	102	95	93,1373	0,9319	95	93,1373	0,9319	92	90,1961	0,9027
50	102	93	91,1765	0,9123	94	92,1569	0,9219	**93**	**91,1765**	**0,9127**
25	102	94	92,1569	0,9223	94	92,1569	0,9223	86	84,3137	0,8435
10	102	94	92,1569	0,9223	95	93,1373	0,9323	48	47,0588	0,5308
9	102	94	92,1569	0,9223	95	93,1373	0,9323	48	47,0588	0,5308
8	102	95	93,1373	0,9323	94	92,1569	0,9223	47	46,0784	0,5408
7	102	95	93,1373	0,9323	94	92,1569	0,9223	48	47,0588	0,5308
6	**102**	**95**	**93,1373**	**0,9323**	94	92,1569	0,9223	48	47,0588	0,5308
5	102	94	92,1569	0,9223	94	92,1569	0,9223	48	47,0588	0,5308
4	102	94	92,1569	0,9223	**95**	**93,1373**	**0,9323**	48	47,0588	0,5308
3	102	93	91,1765	0,9123	92	90,1961	0,9023	45	44,1176	0,5596
2	102	93	91,1765	0,9123	89	87,2549	0,8735	46	45,098	0,55

Çizelge EK A.9. Seçilen genlerin performans değerleri (Prostat Kanseri)

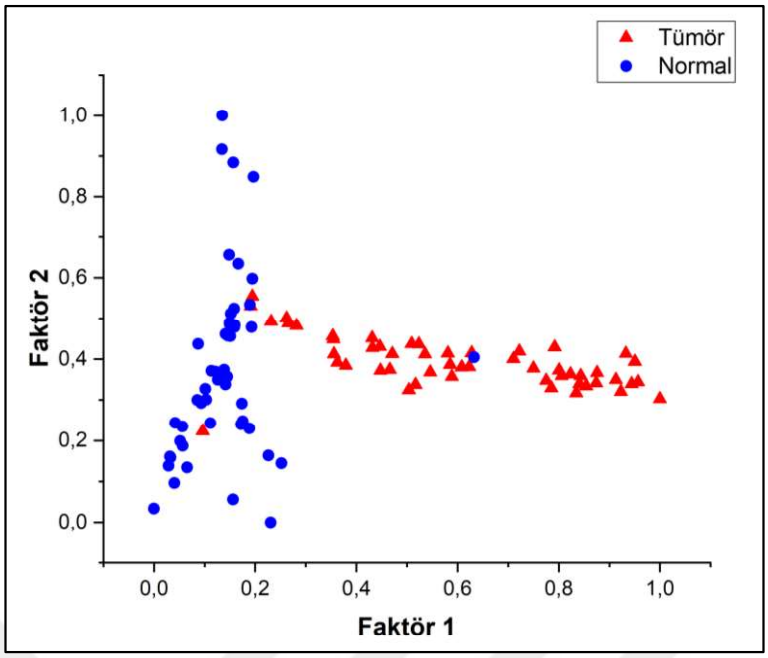
			LOOCV			K 5 (%)			K 10 (%)	
Gen Sayısı	Toplam Örnek Sayısı	Başarılı Sınıf.	Başarı (%)	AUC	En Başarılı	En Başarısız	Ortalama	En Başarılı	En Başarısız	Ortalama
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	97	95,098	0,9515	100	90	94,1429	100	81,8182	95,2727
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818
5	102	98	96,0784	0,9612	100	90,4762	96,0952	100	81,8182	96,1818



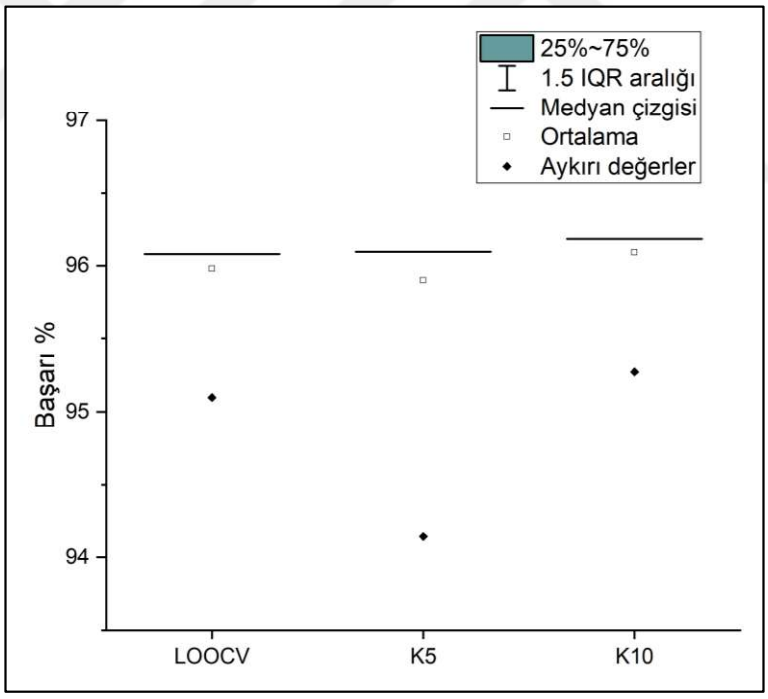
Şekil EK A.13. Tüm genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Prostat Kanseri)



Şekil EK A.14. Birlik Algoritma ile filtrelenen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Prostat Kanseri)



Şekil EK A.15. Seçilen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Prostat Kanseri)



Şekil EK A.16. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (Prostat Kanseri)

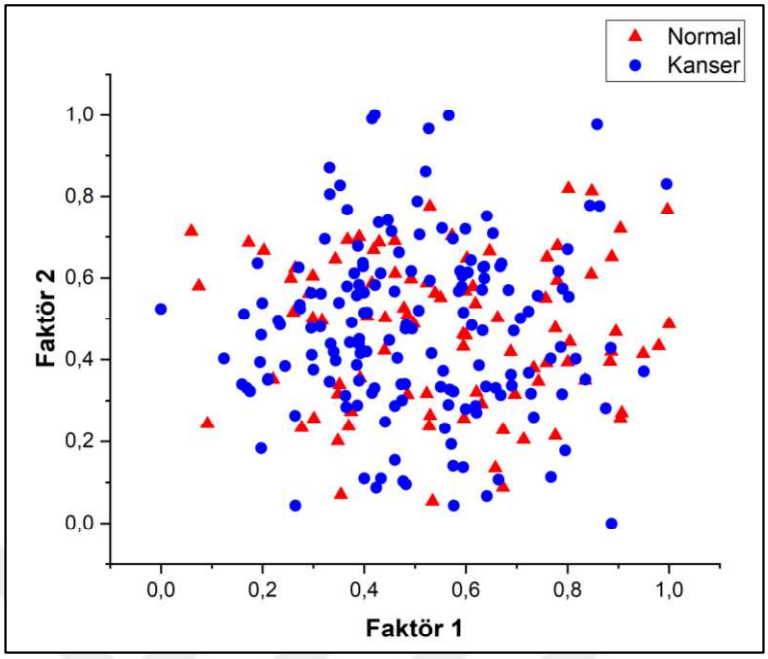
EK A.5. Yumurtalık kanseri

Çizelge EK A.10. Filtreleme sonuçları (Yumurtalık Kanseri)

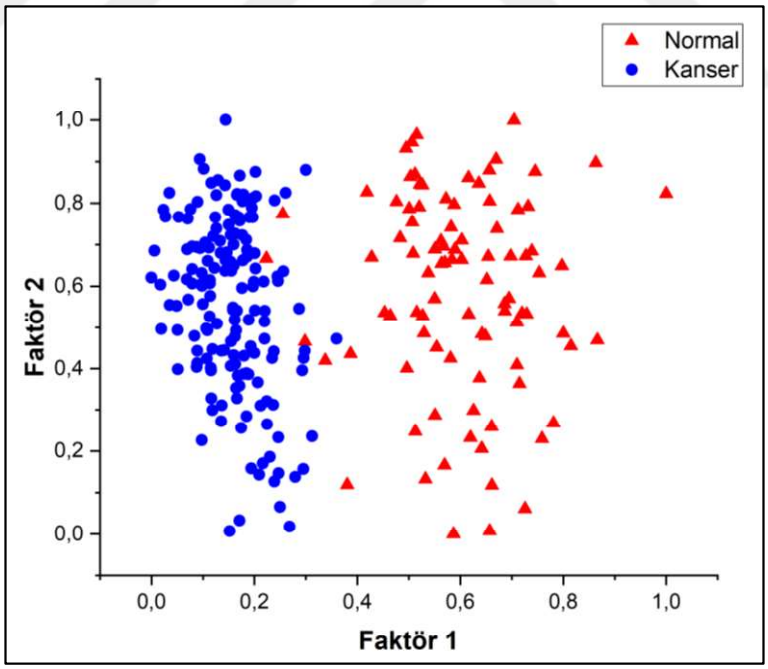
		FKS (LOOCV)			BK (LOOCV)			WRT (LOOCV)		
Gen Sayısı	Toplam Örnek Sayısı	Başarılı Sınıf.	Başarı (%)	AUC	Başarılı Sınıf.	Başarı (%)	AUC	Başarılı Sınıf.	Başarı (%)	AUC
500	253	248	98,0237	0,9725	247	97,6285	0,967	247	97,6285	0,967
250	253	249	98,419	0,978	248	98,0237	0,9725	248	98,0237	0,9725
200	253	249	98,419	0,978	247	97,6285	0,967	248	98,0237	0,9725
150	253	249	98,419	0,978	247	97,6285	0,967	249	98,419	0,978
100	253	250	98,8142	0,9835	248	98,0237	0,9725	252	99,6047	0,9945
75	253	250	98,8142	0,9835	**250**	**98,8142**	**0,9835**	**249**	**98,419**	**0,9804**
50	253	**253**	**100**	**1**	249	98,419	0,978	248	98,0237	0,9773
25	253	249	98,419	0,9804	249	98,419	0,9804	244	96,4427	0,9578
10	253	247	97,6285	0,9718	241	95,2569	0,9461	247	97,6285	0,9743
9	253	246	97,2332	0,9712	241	95,2569	0,9461	247	97,6285	0,9743
8	253	248	98,0237	0,9773	242	95,6522	0,9492	247	97,6285	0,9743
7	253	249	98,419	0,9804	244	96,4427	0,9578	247	97,6285	0,9743
6	253	248	98,0237	0,9773	244	96,4427	0,9578	247	97,6285	0,9743
5	253	247	97,6285	0,9743	245	96,8379	0,9609	247	97,6285	0,9743
4	253	249	98,419	0,9804	246	97,2332	0,9664	247	97,6285	0,9743
3	253	244	96,4427	0,9578	244	96,4427	0,9578	247	97,6285	0,9743
2	253	244	96,4427	0,9578	244	96,4427	0,9578	247	97,6285	0,9743

Çizelge EK A.11. Seçilen genlerin performans değerleri (Yumurtalık Kanseri)

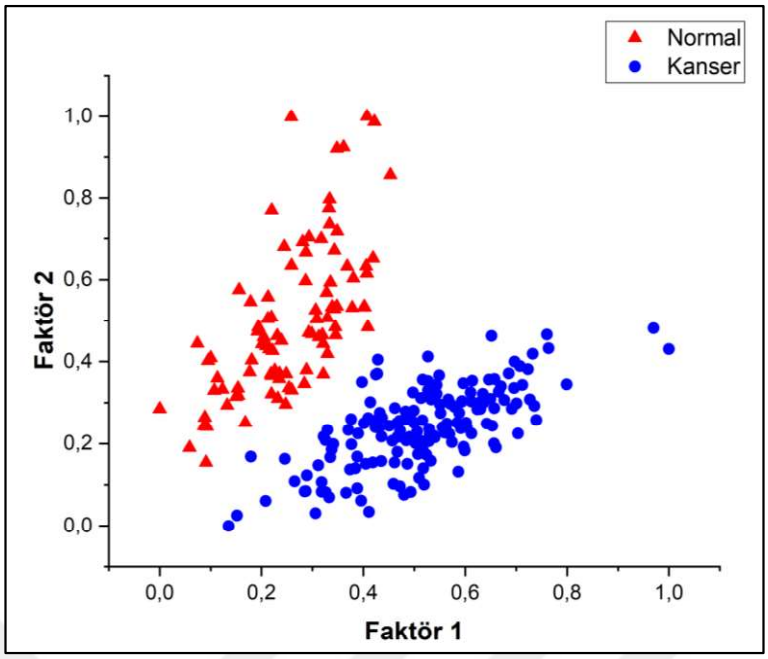
		LOOCV			K 5 (%)			K 10 (%)		
Gen Sayısı	Toplam Örnek Sayısı	Başarılı Sınıf.	Başarı (%)	AUC	En Başarılı	En Başarısız	Ortalama	En Başarılı	En Başarısız	Ortalama
2	253	253	100	1	100	98	99,6	100	100	100
2	253	253	100	1	100	98	99,6	100	100	100
2	253	253	100	1	100	100	100	100	100	100
2	253	253	100	1	100	98	99,6	100	100	100
2	253	253	100	1	100	100	100	100	100	100
2	253	253	100	1	100	100	100	100	100	100
2	253	253	100	1	100	100	100	100	100	100
2	253	253	100	1	100	100	100	100	100	100
2	253	253	100	1	100	96	98,8078	100	96	99,6
2	253	253	100	1	100	100	100	100	100	100



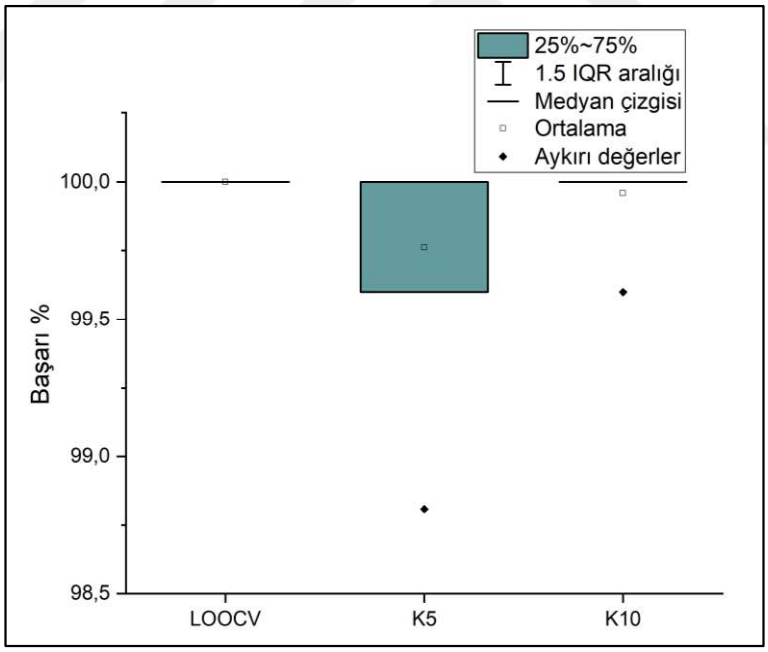
Şekil EK A.17. Tüm genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Yumurtalık Kanseri)



Şekil EK A.18. Birlik Algoritma ile filtrelenen genlerin iki faktörlü temel bileşen analizi sonrası örnek dağılım grafiği (Yumurtalık Kanseri)



Şekil EK A.19. Seçilen genlerin ekspresyon değerlerinin dağılım grafiği üzerinde gösterilmesi (Yumurtalık Kanseri)



Şekil EK A.20. Seçilen genlerin sınıflandırılması sonrasında elde edilen çapraz doğrulama değerlerinin kutu grafiği üzerinde gösterimi (Yumurtalık Kanseri)

ÖZGEÇMİŞ

Adı Soyadı : Mehmet BİLEN

Doğum Yeri ve Yılı : Uluborlu, 1986

Yabancı Dili : İngilizce, İspanyolca

E-posta : mbilen@mehmetakif.edu.tr

Eğitim Durumu

Lise : Isparta Anadolu Meslek Lisesi, 2004

Lisans : SDÜ, Teknik Eğitim Fakültesi, Bilgisayar Sistemleri Öğretmenliği, 2009

Yüksek Lisans : SDÜ, Mühendislik Fakültesi, Bilgisayar Mühendisliği, 2014

Mesleki Deneyim

GÜ Kelkit Aydın Doğan MYO	2011-2012
MAKÜ Çavdır MYO	2012-2020
MAKÜ Gölhisar Uyg. Bil. Yüksek Okulu	2020-…….. (halen)

SCI, SCI-E, SSCI ve AHCI tarafından taranan dergilerde yayımlanan makale:

Bilen, M., Işık A. H., Yiğit T., 2020. A new hybrid and ensemble gene selection approach with an enhanced genetic algorithm for classification of microarray gene expression values on leukaemia cancer. International Journal of Computational Intelligence Systems, 13(1), 1554-1566. DOI: https://doi.org/10.2991/ijcis.d.200928.001

Uluslararası hakemli dergilerde yayımlanan makaleler:

Bilen, M., Işık A. H., Yiğit T, 2019. The Use of Artificial Neural Networks Optimized with Fire Fly Algorithm in Cancer Diagnosis. Gazi University Journal of Science, 32(3), 823-831.

Bilen, M., Işık A. H., Yiğit T, 2019. Development of Web Based Courseware for Artificial Neural Networks. Gazi University Journal of Science, 32(4), 1138-1148.

Bilen, M., IŞIK A. H., YİĞİT T, 2017. İnteraktif ve Web Tabanlı Genetik Algoritma Eğitim Yazılımı. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 21(3), 928-934

Uluslararası bilimsel toplantılarda tam metin olarak sunulan ve bildiri kitaplarında basılan bildiriler:

Bilen, M., Işık A. H., Yiğit T, 2019. Mikro Dizi Verilerinde Sınıflandırma ve Gen Seçim İşlemleri için iki Adımlı Melez bir Yaklaşım. Uluslararası Mühendislikte Yapay Zekâ ve Uygulamalı Matematik Konferansı, 20-22 Nisan, Antalya, 180-187.

Bilen, M., Işık A. H., Yiğit T, 2017. Expert system software for domestic animals. International Conference on Computer Science and Engineering – IEEE, 5-8 October, Antalya, 130-134.

Bilen, M., Işık A. H., Yiğit T, 2015. A Hybrid Artificial Neural Network-Genetic Algorithm Approach for Classification of Microarray Data. 23rd Signal Processing and Communications Applications Conference, 16-19 May, Malatya, 339-342.

Bilen, M., Işık A. H., Yiğit T, 2015. Sismik Darbelerin Sınıflandırılarak Deprem Tehlikesinin Tahmin Edilmesi. Uluslararası Burdur Deprem ve Çevre Sempozyumu, 7-9 Mayıs, Burdur, 41-48.

Bilen, M., Işık A. H., Yiğit T, 2015. KNN Algoritması Tabanlı Mobil Devam Takip Yazılımı. Akademik Bilişim Konferansı, 4-6 Şubat, Eskişehir, 256-261.

Uysal İ., Bilen M., Çiçek A., 2018. Estimation of Cryotherapy with Artificial Neural Network. 1st International Health Science and Life Congress, 02-05 May, Burdur, 39-42.

Uysal İ., Bilen M., Ulukuş S., 2018. Estimation of Fertility Rates with Linear Regression. European Conference on Science, Art & Culture, 19-22 April, Antalya, 545-547.

Uysal İ., Bilen M., Ulukuş S., 2018. Analysis of Classification Algorithms with Rapidminer. European Conference on Science, Art & Culture, 19-22 April, Antalya, 517-520.

Uysal İ., Bilen M., Ulukuş S., 2017. RapidMiner ile Biyolojik Verilerin Sezgisel Algoritmalar Kullanılarak Sınıflandırılması. Uluslararası Türk Dünyası Mühendislik ve Fen Bilimleri Kongresi, 7-10 Aralık, Antalya, 44-48.