AĞIRLIKLI ÇOKLU SINIFLANDIRICI KULLANARAK BİYOLOJİK VERİLERİN TAHMİNİ

TAYLAN İYİDOĞAN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EYLÜL 2013
ANKARA

Fen Bilimleri Enstitü onayı

_____

Prof. Dr. Necip Camuşcu
Müdür

Bu tezin Yüksek Lisans derecesinin tüm gereksinimlerini sağladığını onaylarım.

_____

Doç. Dr. Erdoğan Doğdu
Anabilim Dalı Başkanı

Taylan İYİDOĞAN tarafından hazırlanan AĞIRLIKLI ÇOKLU SINIFLANDIRICI KULLANARAK BİYOLOJİK VERİLERİN TAHMİNİ adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

_____

Yrd. Doç. Dr. Tansel ÖZYER
Tez Danışmanı

Tez Jüri Üyeleri

Başkan : Yrd. Doç. Dr. Mehmet Tan _____

Üye : Yrd. Doç. Dr. Çetin ÜRTİŞ _____

Üye : Yrd. Doç. Dr. Tansel ÖZYER _____

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Taylan İYİDOĞAN

Üniversitesi : **TOBB Ekonomi ve Teknoloji Üniversitesi**

Enstitüsü : **Fen Bilimleri**

Anabilim Dalı : **Bilgisayar Mühendisliği**

Tez Danışmanı : **Yrd. Doç. Dr. Tansel ÖZYER**

Tez Türü ve Tarihi : **Yüksek Lisans – Eylül 2013**

TAYLAN İYİDOĞAN

AĞIRLIKLI ÇOKLU SINIFLANDIRICI KULLANARAK BİYOLOJİK VERİLERİN TAHMİNİ

ÖZET

Kanser, günümüzde çok yaygın olarak rastlanılan tedavisi zor olan bir hastalıktır. Göğüs kanseri, akciğer kanseri, merkezi sinir sistemi kanseri ve lösemi başta olmak üzere birçok çeşidi bulunmaktadır. Bunlar, vücuttaki farklı gen dizilimlerindeki hatalardan dolayı oluşmaktadır. Bu hastalık türlerinin tedavi edilebilmeleri için erken teşhis önemlidir. Başka bir deyişle, genler dizilimlerinin doğru anlamlandırılması anlamına gelmektedir. Günümüzde mikro çip teknolojisi ile geniş çaplı gen sınıflandırılması mümkündür. DNA mikroçip, aynı anda, farklı seviyelerde genlerin durumlarını sunar. Geçmiş çalışmalar, mikroçip teknolojisinin tümör sınıflandırmasında da çok etkili olduğunu göstermektedir. İlgili veri setleri fazla özniteliğe sahip olmalarına rağmen, az örnek bulundururlar. Geçmiş yıllarda, farklı sınıflandırıcı yöntemleri geliştirilmiştir. Yöntemler, veri setlerine bağımlı olarak farklı doğruluk oranlarına sahip olabilirler. Bu da bir sınıflandırıcının, bütün veri setlerinde aynı etkinlikte çalışamayacağı anlamına gelmektedir. Bir veri seti için; birden çok sınıflandırıcı kullanıp bir topluluk oluşturarak, sınıflandırıcı doğruluğunu arttırmak ve yanlış sınıflandırıcı seçme olasılığını azaltmak hedeflenmiştir. Yaygın kullanımda sınıflandırıcı topluluğu, birçok sınıflandırıcının bireysel performanslarına bakılarak ortak bir karar mekanizması yaratmak ve daha önce tanımlanmamış örnekleri bu karar mekanizması ile sınıflandırılması için kullanılır. Sınıflandırıcı topluluk oluşturma, sadece mikroçip veri setlerini sınıflandırmanın yanı sıra, görüntü işleme, yapay zekâ ve tıp gibi birçok alanda da kullanılmaktadır.

Bu tez çalışmasında, literatürde bulunan yirmi dokuz sınıflandırıcı kullanaarak, veri setlerine özel sınıflandırıcı topluluklar oluşturulmuştur. Ayrıca oluşturulan topluluktaki her bir sınıflandırıcıya ağırlıklar atanmıştır. Bu sayede, sınıflandırmanın doğruluğunun arttırılması amaçlanmıştır. Sınıflandırıcı topluluğu oluşturmak için iki seviyeli genetik algoritma kullanılarak yeni bir metot geliştirilmiştir. İki seviyeli genetik algoritmanın ilk seviyesinde sınıflandırıcılar seçilmiş, ikinci seviyesinde ise, bu sınıflandırıcılara uygun ağırlıklar atanmıştır. Önerilen yöntem ile alınan sonuçlar, diğer yöntemlerle elde edilen sonuçlardan üstündür.

Anahtar Kelimeler: Genetik Algoritma, genetik değişim, mutasyon, sınıflandırıcı, sınıflandırıcı topluluğu, doğruluk, duyarlılık, özgüllük, f-skorlama, uygunluk fonksiyonu, seçim, hassaslık, sınıflandırıcı ağırlık, kromozom, popülasyon, doğru pozitif oranı, doğru negatif oranı, yanlış pozitif oranı, yanlış negatif oranı.

University : **TOBB Economics and Technology University**

Institute : **Institute of Natural and Applied Sciences**

Science Programme : **Computer Engineering**

Supervisor : **Assistant Prof. Dr. Tansel ÖZYER**

Degree Awarded and Date : **Master of Science – September 2013**

TAYLAN İYİDOĞAN

PREDICTION OF BIOLOGICAL DATA BY USING WEIGHTED ENSEMBLE CLASSIFIERS

ABSTRACT

Nowadays, cancer disease is rapidly spreading and it is difficult to treat. There are different types of cancer such as breast, lung, central nervous system, leukemia and many more. Each type consists of different sequence error in the gene syntax in the body. In order to treat them, early diagnosis is mandatory, It means that gene sequences must be interpreted in a smart and accurate way. Genome wide scale data classification has been triggered by the microarray technologies. A DNA microarray can have different expression levels of genes simultaneously. Past studies indicate that microarray technology can provide means for tumor classification. Related datasets contain small number of samples; the number of genes is extremely larger than the number of samples and these make knowledge discovery challenging in microarray. For this reason, different classification methods were proposed. However, classifiers' performance results may be dataset dependent. While the classification accuracy of one classifier can make high accuracy for some dataset, it may give poor accuracy for another. That means, there is no perfect classifier works for all datasets robustly. Regardless of selecting one classification approach, ensemble classifiers approach is proposed. It combines each individual classifier prediction in the pool to create joint decision-rules and classify unknown samples according to this decision rules. Ensemble classifiers is being used in many image processing, artificial intelligence and medical fields as well.

In this thesis, I worked on creating an ensemble classifiers method that chooses best classifier combination in the classifier pool. I focused on twenty nine different classifiers in literature to selection each having different weights based on their past performance. A two level genetic algorithm is used to create ensemble classifiers. In this proposed method, the first level determines the classifiers to use; the second level optimizes their weights for effectiveness. The results of this study are promising when compared to other methods.

Keywords: Genetic algorithm, crossover, mutation, classifier, classifier ensemble, accuracy, specificity, f-measure, fitness function, selection, sensitivity, classifier weight, chromosome, population, true positive rate, true negative rate, false positive rate, false negative rate.

TEŞEKKÜR

Bu tezin hazırlanmasında yardım ve katkılarıyla beni yönlendiren değerli hocam Yrd. Doç. Dr. Tansel Özyer'e, yüksek lisans eğitimim boyunca bana değerli katkılarda bulunan TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği bölümü hocalarıma ve en önemlisi beni bugünlere getiren değerli aileme teşekkürü borç bilirim.

İÇİNDEKİLER

ŞEKİLLERİN LİSTESİ

TABLOLARIN LİSTESİ

ALGORİTMALARIN LİSTESİ

KISALTMALAR

Kısaltma	Açıklama
GA	Genetik Algoritma
SÖD	Sınıf Özellikli Doğruluk
CPU	Merkezi İşlem Birimi
TP	Doğru Pozitif Oranı
TN	Doğru Negatif Oranı
FP	Yanlış Pozitif Oranı
FN	Yanlış Negatif Oranı
EA	Evrimsel Algoritmalar

SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile birlikte aşağıda sunulmuştur.

Simge	**Açıklama**
$maks$	Maksimum
μ_i	Destek
d	Sınıflandırıcı tahmini
w	Ağırlık
t	İterasyon
P	Olasılık
K	Kromozom
x	Sınıflandırılmayan örnek
S	Toplam sınıflandırıcı sayısı

1 GİRİŞ

DNA mikroçipleri, gen ifadelerini gözlemlemek için önemlidir. Farklı dokularda, farklı koşullar altında yapılmış olan ölçümler, bize kapsamlı analiz ve DNA yapılarının keşfedilmesi için fırsat vermektedir[16]. Doğadaki çevresel etmenler ve bireydeki genetik faktörler sebebi ile gen dizilimlerinde oluşan en ufak değişim kansere yol açabilir. Eğer erken teşhis edilmezse ölüm ile sonuçlanabilir. Bu yüzden de hastalığın teşhisi, seyrinin takibi tedavi için çok önemli dönüm noktalarıdır. DNA örneklerini taşıyan veri setleri, boyut bakımından fazla özniteliğe sahip olmalarının yanı sıra, az örnek bulundururlar. Araştırmacılar, büyük veri setleri içerisinden anlamlı bilgileri çıkarma ve sınıflandırma konularında zorluklar yaşamaktadırlar. Kanser verisinin sınıflandırılması, klinik tanılar için zorunlu hale gelmiştir. Zaman içerisinde hastalık tespiti amacıyla birçok metot ve yöntem geliştirilmiştir. Son zamanlarda DNA mikroçipler üstünde sınıflandırma, kümeleme [4] ve özellik seçim [19] konuları üzerinde yapılan çalışmalar sayesinde araştırmacılar, büyük veri setleri üzerinde, tek bir deneyde, binlerce genin hareketini izleyebilir hale gelmişlerdir [20],[21]. Bu da normal ve tümörlü hücrelerin farklı ifade seviyelerine göre sınıflandırılabilmesi, genler arası ilişkinin ortaya çıkarılabilmesi ve hastalığa sebep olan genlerin bulunabilmesi konularında yardım sağlamaktadır.

Günümüze farklı sınıflandırma algoritmaları önerilmiştir. Ancak, bu algoritmaların performansları genellikle veri setine bağlıdır. Sınıflandırıcıların çoğu, eğitim setiyle test verisinin benzer nitelikte olduğunu hesaba katar ama gerçek hayatta bu doğru olmayabilir. Veri setindeki sınıf bazındaki dağılım farklılığı ya da verilerin farklı kaynaklardan gelmesi, örnek sayısının az olması nedeniyle sınıflandırıcılar düşük performans gösterebilirler. Bu yüzden de farklı sınıflandırıcıları birleştirmek, Sonucun iyileştirmesi için yardımcıdır. Bir sınıflandırıcının yanlış sınıflandırdığı bir veri için, diğer sınıflandırıcılar yaptıkları farklı sınıflandırma ile bunu düzeltebilirler. Bu da, genel performansta bir artış sağlayabilir. Seçilen sınıflandırıcılar ne kadar çeşitli olursa, genel performans o kadar artacaktır.

Mikroçip veri seti sınıflandırma konusunda önerilen birçok yöntem bulunmaktadır. Bunlardan bazıları (Torbalama ve Arttırılma [17]), bilinmeyen örneklerin doğru sınıflandırılabilmesi için, sınıflandırıcı metotlarını birleştirmeye yönelik çalışmalardır. Sınıflandırıcı toplulukları, her bir sınıflandırıcının bireysel tahminlerinin bir şekilde birleştirilip, yeni örnekleri sınıflandırabilmek için oluşturulmuş bir modeldir. Bu topluluktaki amaç, sınıflandırma doğruluğunu tek başına en iyi çalışan sınıflandırıcının doğruluğunun üstüne çıkarmaktır. Sınıflandırıcının hangi veri setinde daha iyi sonuç vereceği bilinmediğinden, sınıflandırıcı topluluğu yaratmak ve her bir sınıflandırıcının karar vermedeki katkısını aynı zamanda eniyileme, düşük performans gösterebilecek bir sınıflandırıcı ile sonuç elde etmemizin önüne geçecektir.

Bu tez çalışmasında, sınıflandırıcı topluluğu oluşturmak için yeni bir yöntem önerilmiştir. Bu yöntem iç içe iki seviyeli genetik algoritma [1] kullanarak, veri seti için aşağıdakiler hedeflenmiştir:

- Büyük sınıflandırıcı kümesinden en iyi sınıflandırıcı kümesini seçmek,
- Küme içindeki sınıflandırıcıların test verisi üzerindeki katkılarına optimal ağırlık değeri verilmesi.

Genetik algoritmanın ilk seviyesinde sınıflandırıcı kromozom seçilirken, ikinci seviyedeki genetik algoritmada bu sınıflandırıcılara ait ağırlıklar seçilmektedir. Sınıflandırıcılar, farklı veri setleri için performans değişikliği gösterebilirler. Bu yüzden sınıflandırıcılara farklı ağırlıklar atayarak, sınıflandırma doğruluğunun arttırılması hedeflenmiştir. Önerilen yöntemin başarısını ölçmek için yedi farklı veri seti kullanılmıştır (Tablo 4).

Tez çalışması altı bölümden oluşmaktadır. Bölüm 1 de tez konusunu oluşturan problemin tanımını derinleştiren ve bu problemin çözümü için önerilen metodun kabaca hangi aşamalardan oluştuğunu anlatan giriş bölümü bulunmaktadır. Bölüm 2 de Genetik Algoritma tanıtımı ve işleyişinden bahsedilmiş, ayrıca GA ve sınıflandırıcı toplulukla ilgili yapılan çalışmalar incelenmiştir. Bölüm 3 de önerilen metodun GA ile ilişkisinden bahsedilmiş ve metot ayrıntılı olarak anlatılmıştır.

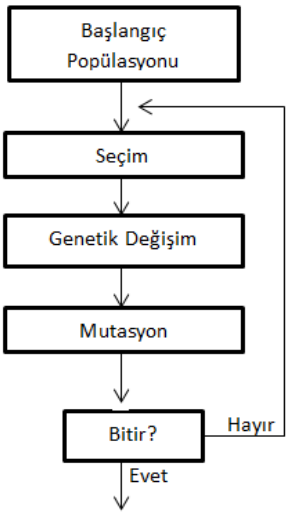
Bölüm 4 de deneylerde kullanılmış veri setlerinden bahsedilmiş ve deneylerin hangi ortamda ve nasıl alındıkları anlatılmıştır. Bölüm 5 de deneyler ve elde edilen sonuçlara yer verilmiştir. Tez çalışması 6. bölüm ile yapılan çalışmaya ait izlenimlerin aktarılması ve gelecek çalışmaların ne doğrultuda olacağı belirtilerek sonlandırılmıştır.

2 İLGİLİ ÇALIŞMALAR

Bu bölümde genetik algoritmanın detaylı tanımı, sınıflandırıcı topluluklar ve genetik algoritma teknolojilerinin birleşimi ile ilgili geçmiş yıllarda yapılan çalışmalardan ve sınıflandırıcı birleşim metotlarından bahsedilmiştir.

2.1 GENETİK ALGORİTMA

Genetik algoritma doğal evrim sürecini taklit eden sezgisel bir arama yöntemidir. İlk olarak Holland tarafından 1975 yılında yayınlanmıştır[1]. Bu sezgisel arama yöntemi, genelde en uygun şekle sokma ve arama algoritmalarına yararlı çözümler sunmak için kullanılır. Genetik algoritmalar, evrimsel algoritmalarda (EA) üst sıralarda yer alırlar ve optimizasyon problemlerine sonuçlar üretmek için kullanılırlar. Bu çözümler üretilirken, miras alma, mutasyon, seçim ve genetik değişim gibi başlıkları kullandıkları için doğanın gelişiminden ilham almıştır denilebilir. Genetik algoritmanın çalışma şekli (Şekil 2.1.1) de verilmiştir. Ayrıca sözde kodu (Algoritma 1) de gösterilmiştir.



Şekil 2.1.1 - Genetik algoritmanın en basit işleyişi

Algoritma 1 : GenetikAlgoritma()

```
1: Çıktılar:
2:   p % Sonuç popülasyonu
3:
4: başlangıçPopülasyonSeç()
5:  foreach birey : Popülasyon
6:      uygunlukFonksiyonHesapla(birey)
7:  end for
8:  while (bitir = false)
9:    GenetikDeğişimVeMutasyonYap()
10:     foreach birey : Popülasyon
11:         uygunlukFonksiyonHesapla(birey)
12:     end for
13:   p = popülasyonaUymayanBireyleriYenileriİleDeğiştir()
14:   bitir = bitirmeFonksiyonu()
15:  end while
16: return p
```

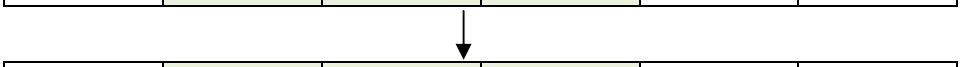
2.1.1 Genetik Değişim

Genetik değişim, yeni bir nesil oluşturmanın bir alt metodudur. Kromozomları genetik değişime sokmak zorunda olmamızın sebebi, topluluk içerisindeki çeşitliliği ve kromozom sayısını arttırmaktır. Genetik değişim sonrasında, topluluk içerisindeki kromozom sayısı yaklaşık iki katına çıkar. Bu da, yeni bir nesil oluştururken en iyi kromozomları seçme şansı verir. Genetik değişim yaygın anlamda 3 farklı yöntem ile yapılmaktadır. Bu yöntemler aşağıda açıklanmıştır.

2.1.2 Tek Noktalı Genetik Değişim

Tek noktalı genetik değişimde seçilen iki kromozom arasında, kromozomun boyunu geçmeyecek bir yer belirlenir ve bu belirlenen noktanın başta kalan kısımları yer değiştirilir (Şekil 2.1.2).

K 1	1	2	3	4	5
K 2	6	7	8	9	10



K 3	6	7	8	4	5
K 4	1	2	3	9	10

Şekil 2.1.2 – Tek Noktalı Genetik Değişim

Burada *K,* kromozomları ifade ederken, diğerleri kromozomdaki genlerin değerleridir. Bu örnekteki tek noktalı genetik değişimde, örnek olarak üçüncü indis seçilmiş ve bu seçimdeki iki kromozomun ilk üç indisindeki genler yer değiştirmiştir.

2.1.3 İki Noktalı Genetik Değişim

Bu genetik değişim metodunu tek noktalı genetik değişim metodundan ayıran özellik, kromozomların boyu geçmeyecek şekilde 2 tane indis belirlenir ve bu indisler arasında kalan genler yer değiştirir. (Şekil 2.1.3).

K 1	1	7	8	9	5
K 2	6	2	3	4	10

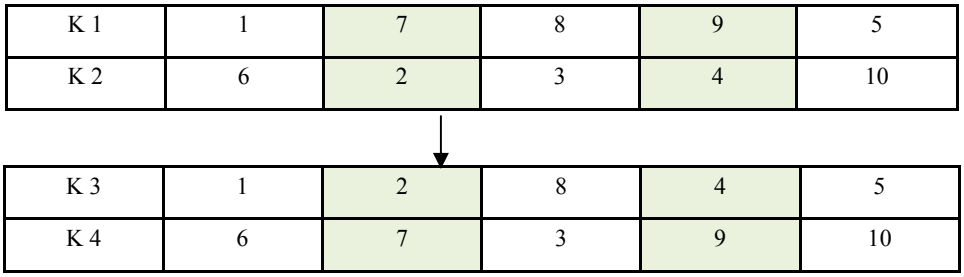
K 3	1	2	3	4	5
K 4	6	7	8	9	10

Şekil 2.1.3 – İki Noktalı Genetik Değişim

Burada kromozom içerisinden 1. ve 4. indis seçilmiştir. Arada kalan genler yer değiştirdiklerinde ortaya iki farklı kromozom çıkmaktadır.

2.1.4 Standart Genetik Değişim

Bu genetik değişim metodu, üst bölümde anlatılan 2 metottan daha farklıdır. Standart genetik değişimde, her bir gen ayrı ayrı ele alınır. Kromozomun bütün genleri baştan sona taranarak, bütün indisler için rastgele bir değer üretilir. Bu rastgele değer, belirlenen eşik değerini geçiyorsa, genler arası değişim gerçekleşir. Eğer eşik sınırı geçilmemişse, o zaman değişim yapılmaz ve bir sonraki gene geçilir (Şekil 2.1.4).



Şekil 2.1.4 – Standart Genetik Değişim

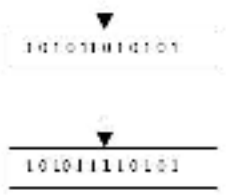
Bu örnekte, 2. ve 4. indiste üretilen rastgele sayılar eşik değerini geçtikleri için bu genler yer değiştirmişlerdir.

Bir kromozom topluluğunda çeşitlilik arttıkça, o topluluktaki uygunluk fonksiyon sonuçları da değişiklik gösterir. Fakat kromozomlar, bazen birbirlerine oldukça benzeyebilir. Böyle bir durumda popülasyonda birbirine çok benzeyen kromozomlar oluşabilir. Bu da en iyi kromozomu bulmada zorluklar yaratabilir çünkü popülasyon yerel maksimumda takılabilir. Bunu engellemek için de mutasyon kullanılır. Ayrıca genetik değişim sonrası oluşan kromozomların uygunluk fonksiyon sonuçları ebeveynlerden daha iyi çıkmayabilir. Hem topluluğa uymayan kromozomları ayıklamak, hem de yeni nesil popülasyon büyüklüğünü azaltmak için seçim işlemi gerçekleşmektedir.

2.1.5 Mutasyon

Mutasyon, dışarıdan bir müdahale ile kromozomun yapısını değiştirmektir. Popülasyonlar gelişimlerini sağlarlarken yerel maksimumda takıldığı zaman popülasyon daha fazla gelişemez hale gelir. Mutasyonun uygulanmasının önemli

olmasının sebebi ise, bu yerel maksimumlardan kurtulmak ve kromozomlardaki çeşitliliği arttırmaktır. Mutasyonun çalışma şekli (Şekil 2.1.5) de gösterilmiştir.



Şekil 2.1.5- Mutasyon çalışma örneği

2.1.6 Seçim

Seçim, topluluktaki kromozom sayısını azaltmak için uygulanan metottur. Genetik değişim ve mutasyon sonunda kromozom sayısı arttığı için, yeni oluşan toplulukta bu kromozomları azaltmak gerekmektedir. Eğer kromozomlarda azaltma olmazsa, her yeni toplulukta sürekli kromozom artışı olacaktır. Bunu engellemek ve daha iyi topluluklar yaratmak için değişim ve mutasyon sonucunda yaklaşık iki katına çıkmış olan kromozomlardan en iyileri seçilir ve seçilen kromozomlar ile yola devam edilir. En iyi kromozomları seçmenin yolu ise uygunluk fonksiyonudur. Bu fonksiyona göre, her bir kromozomu değeri hesaplanır. Yeni bir topluluk için ise en iyi değerlere sahip olan kromozomlar seçilir.

2.1.7 Uygunluk Fonksiyonu

Uygunluk fonksiyonu en iyi kromozomlara karar veren metottur. Yeni bir nesil seçilmeden önce her bir kromozom uygunluk fonksiyonundan geçirilir ve en iyi değerlere sahip olan kromozomlar yeni nesil için seçilir.

2.1.8 Sınıflandırıcı

Sınıflandırıcı, bir veri setindeki nitelikleri haritalama işini üstlenen fonksiyonlardır. Bu nitelikler yardımı ile yeni bir örnek veri sınıflandırılabilir. Bir sınıflandırıcı topluluğunun çıktısı ise, tek bir sınıflandırıcıya bağlı kalmamakla birlikte, bütün sınıflandırıcıların, tek tek sonuçlarının birleştirilmesi ile ortaya çıkmaktadır. Bu

yüzden bu topluluğun ürettiği çıktının doğru olması, aslında içindeki her bir sınıflandırıcının bireysel performansı ile alakalıdır.

Bir sınıflandırıcı topluluğunun iyi sonuçlar vermesi iki gereklilikle sağlanır. Bunlardan birincisi, toplulukta farklı sınıflandırıcıların bulunmasıdır. Benzer sınıflandırıcılar, benzer hatalar yaparlar ve bu da, topluluğun performansını yükseltemez. İkincisi ise sınıflandırıcıların tahminlerindeki doğruluklardır. Bir toplulukta eğer kötü tahmin yapan bir sınıflandırıcı olursa, bu sınıflandırıcı genel performans da bozukluklara yol açacaktır. Buradaki önemli faktör, yanlış tahmin edilen sonuçlardaki, sınıflandırıcı tahminleridir. Eğer topluluktaki sınıflandırıcılar sürekli aynı örneği yanlış tahmin ederlerse, o zaman genel performans düşecektir. Bir diğer yandan, eğer topluluktaki sınıflandırıcılar sürekli birbirlerinden farklı tahmin yaparlarsa, bu da genel performansın düşmesine yol açacaktır. Bu yüzden bir topluluktaki sınıflandırıcılar birbirleri ile ne kadar tutarlı olurlarsa, genel performans o kadar iyi olacaktır.

2.1.9 Sınıflandırıcı Birleştirme

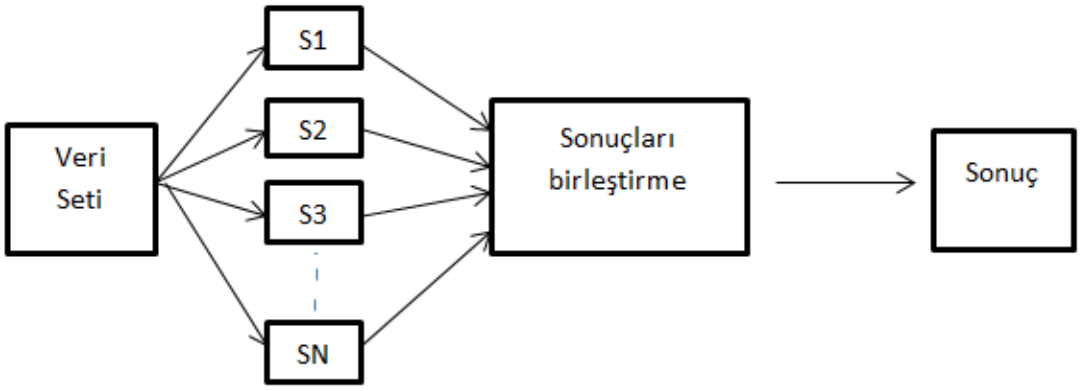
Sınıflandırıcıları birleştirme kavramı aslında gayet basittir. Aynı veri setini kullanarak birçok sınıflandırıcı için model ürettikten sonra, bu model çıktılarını birleştirmektir. Genellikle bu birleştirme iki farklı oylama ile yapılmaktadır. Birincisinde tüm sınıflandırıcıların aynı ağırlıkları vardır ve çıktı olarak en çok hangi sınıf en fazla oy almışsa, sonuç o sınıf kabul edilir. Diğer oylamada ise sınıflandırıcılara farklı ağırlıklar atanır ve sınıflandırıcıların çıktıları bu ağırlıklarla hesaplanır. Burada da en fazla oy alan sınıf, sonuç olarak kabul edilir.

2.1.10 Sınıflandırıcı topluluk oluşturmanın yapısı

Sınıflandırıcı toplulukları oluşturmak iki kısımda incelenebilir. Bunlardan birincisi sınıflandırıcıları seçmek, diğer ise bu sınıflandırıcıların tahmin ettikleri sonuçları birleştirmektir. Sınıflandırıcıları seçmek için sınıflandırıcıların bireysel olarak tahmin ettiği sonuçlara bakılır. Bunların çeşitliliği ve birbirleri ile olan tutarlılığı, topluluktaki genel performansı arttıracaktır. Topluluktaki sınıflandırıcılar seçildikten

sonra, bunların tahminlerinin birleştirilmesi gerekmektedir. Bu şekilde, bu topluluktan tek bir sınıflandırıcı seçmek yerine bütün sınıflandırıcıları kapsayan bir model yaratılır. Bu aşamalar (Şekil 2.1.6) da gösterilmiştir.

Önceki çalışmalarda Bu aşamaların verimli geçirilmesi için farklı metotlar önerilmiştir. Farklı sınıflandırıcı modeller seçilmesi bunlardan biridir. Ayrıca sınıflandırıcıların eğitildikleri veri seti ne kadar tutarlı olursa, çıkardıkları modeller de o kadar tutarlı olur. Bu sayede en iyi model çıkartılabilir. Bir diğer durum ise sınıflandırıcılar farklı veri setlerinde farklı performanslar sergileyebilirler. Bunun için, sınıflandırıcıları seçmeden önce, eğitim setinden, farklı alt veri setleri ile, sınıflandırıcı modellerin performansları değerlendirilerek seçim yapılır. Son olarak ise sınıflandırıcıların çıkardıkları tahminler için farklı kombinasyonlar kullanılmıştır.



Şekil 2.1.6 – Sınıflandırıcı Topluluğu modeli örneği

2.1.11 Sınıflandırıcı birleşim seçenekleri

Sınıflandırıcılar, bir veri seti için eğitildikten sonra, çıkardıkları tahmin sonuçları farklı şekilde birleştirilebilir. Bazı birleşim seçenekleri sınıf etiketleri üstünden, bazıları ise sınıflandırıcıların, o sınıflara verdikleri destek ile anlatılmaktadır [2]. Burada iki temel birleşim yönteminden bahsedilmiştir. Birincisinde, sınıflandırıcılar test edilen örneklerde, tek bir sınıf tahmin ederler. İkincisinde, tahmin için tek bir sınıf değil, bütün sınıfları sıralamaya koyalar.

2.1.12 Sınıf etiketleri birleşimi

Topluluktaki sınıflandırıcıların yaptıkları tahminler, sınıf etiketleridir. Her bir sınıflandırıcı tek bir sınıfın etiketini tahmin eder. Daha sonra ise bunlar üç şekilde birleştirilir [3].

1) **Çoğunluk oylaması**

Bu yaklaşımda, her bir sınıflandırıcının eşit seviyede oyu vardır ve en çok oyu olan tahmin, o örneğin sonucu olarak kabul edilir. Çoğunluk oylaması hesaplaması 2.1.1'de gösterilmiştir.

$$\sum_{k=1}^{K} d_{t,j}(x) = maks_{j=1}^{S} \sum_{k=1}^{K} d_{kj} \quad (2.1.1)$$

Bu denklemde *K,* toplam sınıflandırıcı sayısı, her bir *k* değeri ise o sınıflandırıcıyı temsil eder. d_{kj} ise k sınıflandırıcısının yaptığı tahmindir. *S*, veri setindeki toplam sınıf sayısıdır.

2) **Ağırlıklı çoğunluk oylaması**

Bu yaklaşımda, sınıflandırıcıların farklı seviyelerde oyları vardır. Eğer bir sınıflandırıcının diğerlerine göre daha performanslı sonuç verildiğine dair bir kanıt olursa, o sınıflandırıcı, diğerlerine göre daha yüksek oya sahip olur. Bu durumda ortaya çıkan sonuç, daha yüksek performanslı ve doğru olur Ağırlıklı çoğunluk oylaması hesaplaması 2.1.2'de gösterilmiştir.

$$\sum_{k=1}^{K} a_k d_{t,j}(x) = maks_{j=1}^{S} \sum_{k=1}^{K} a_k d_{kj} \quad (2.1.2)$$

Bu denklemde *K,* toplam sınıflandırıcı sayısını ifade ederken, her bir *k* değeri o sınıflandırıcıyı temsil eder. *A* ise o sınıflandırıcıya ait ağırlığı temsil eder. d_{kj} ise *k* sınıflandırıcısının yaptığı tahmindir. *S*, veri setindeki toplam sınıf sayısını belirler.

3) **Sınıf etiketlerine verilen destek ile birleşim**

Bu yaklaşımda, sınıflandırıcı sonuçları tek bir sınıf etiketi olmayıp, bütün etiketlere verdiği destek ile hesaplanır. Burada, sınıflandırıcının en çok destek verdiği sınıf, o sınıflandırıcı için sonuç kabul edilmez. Test edilen örnek için sonuç, bütün sınıflandırıcılardan en çok desteği alan sınıftır.

$$\mu_j(x) = \sum_{t=1}^{T} d_{tj}(x) \tag{2.1.3}$$

Burada, $d_{t,j}(x)$, örnek x için t sınıflandırıcısının j sınıfına verdiği destek, $\mu_j(x)$, j sınıfı için x örneğine verilen toplam destek, *T*, toplam sınıflandırıcı sayısıdır.

2.2 SD-EnClass

Sajid ve diğerleri tarafından önerilen SD-EnClass metodu[4] sınıflandırıcıları seçme ve sonuçlarının birleştirilmesi için farklı bir yöntem sunmaktadır. Bu metot, iki aşamadan oluşmaktadır. İlk aşamasında SD-EnClass metodunu uygulamaktadırlar.

SD-EnClass metodu iki aşamalı bir metottur. İlk aşamasında veri seti içerisindeki gereksiz gördükleri eğitim verilerini siler. Bu verileri silmek için SD-Prune-Redundant algoritmasını kullanırlar. Bu algoritma veri setindeki bütün örnekleri okur ve bu örnekler içerisinde tüm örneklerin niteliklerine değer verir. Daha sonra ise birbirlerine benzeyen örnekleri çıkarır. Bu sayede veri seti üzerinde bir filtreleme yapmış olur. Veri seti azaltıldıktan sonra önerdikleri SD-EnClass metodunu uygularlar.

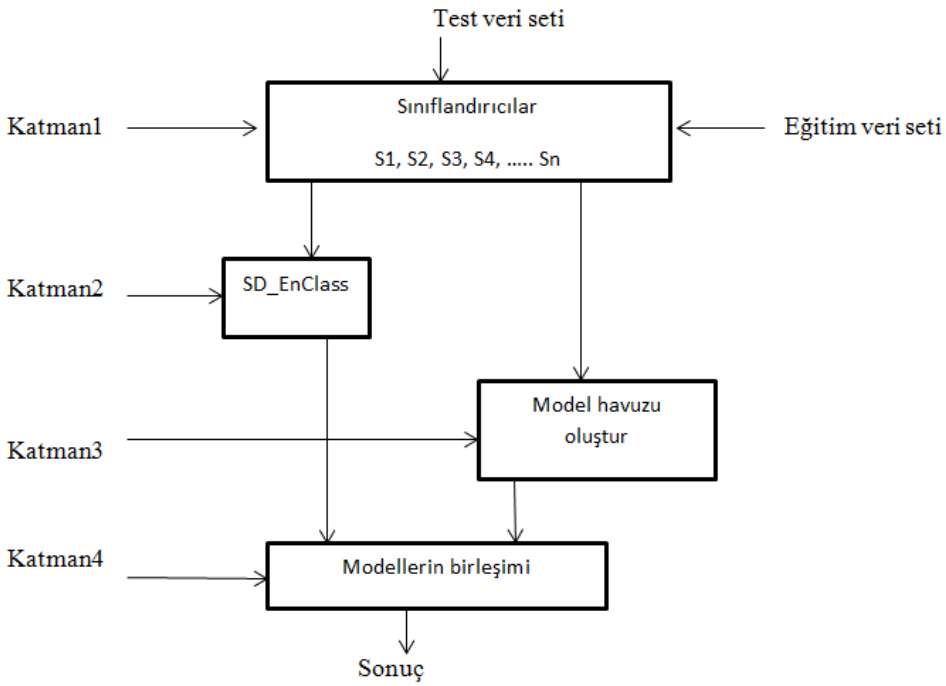
SD-EnClass metodu için üç sınıflandırıcı seçmişlerdir. Bunlar J48, IBK ve NaiveBayes sınıflandırıcılarıdır. Bu metotta temel amaç, 3 sınıflandırıcı sonuçlarını maksimum doğrulukla birleştirmektir. Önerdikleri metodun ilk aşamasında öncelikle her bir veri seti içerisindeki sınıf için uzman sınıflandırıcılar bulurlar. Bu sınıflandırıcılar "Sınıf özellikli doğruluk" hesaplaması ile tespit edilir. Bu hesaplama 2.2.1'de gösterilmiştir.

$$S\ddot{O}D = \frac{Sınıf\ için\ toplam\ tahmin\ edilen\ örnek\ sayısı}{Sınıf\ için\ doğru\ tahmin\ edilen\ örnek\ sayısı} \tag{2.2.1}$$

Her bir sınıf için uzman sınıflandırıcılar seçildikten sonra test aşamasında sınıflandırıcıların tahminlerinden üretilen model basamakları aşağıda verilmiştir.

1) Bütün sınıflandırıcılar aynı sınıfı tahmin etmişlerse, sonuç tahmin edilen sınıftır.
2) Çoğunluk aynı tahmini yapmışsa (3 üstünden 2) o zaman sonuç aşağıdakilerden biridir
    a. Eğer S3 kendi uzman olduğu sınıfı tahmin etmiş, S1 ve S2 farklı tahmin etmişse, o zaman sonuç S3 kararıdır.
    b. Eğer S3 kendi uzman olduğu sınıfı tahmin etmiş, S1 veya S2 de kendi uzman oldukları sınıfı tahmin etmiş ise sonuç, SÖD'si yüksek olan sınıflandırıcının tahmin ettiği sonuçtur.
3) Bütün sınıflandırıcılar birbirlerinden farklı tahmin yapmışlar ise
    a. Eğer herhangi bir sınıflandırıcı kendi uzman olduğu sınıfı tahmin etmişse, sonuç odur.
    b. Eğer iki sınıflandırıcı kendi uzman oldukları sınıfları tahmin etmişler ise SÖD'si yüksek olan sınıflandırıcının tahmin ettiği doğrudur.
    c. Eğer bütün sınıflandırıcılar kendi uzman oldukları sınıfları tahmin etmişler ise SÖD'si yüksek olan sınıflandırıcının tahmin ettiği doğrudur.

Oluşturulmuş olan bu modelin performansı ve bireysel sınıflandırıcıların performanslarına bakıldığında, bu modelin ortalama seviyede kaldığı görülmektedir. Bunun çözümü olarak da yaptıkları SD-EnClass modelinin yanına fazladan sınıflandırıcı koyarak doğruluk arttırılmaya çalışılmıştır. Bunun için en başta seçtikleri sınıflandırıcılardan 3 katı model elde edip, bu modellerden de en iyi SÖD değerine sahip olan 2 tanesini seçmişlerdir. Sonuç olarak, toplamda 3 model elde etmişlerdir. Bu şekilde toplam önerilen metot 4 katmanlı olmuştur. Bu katmanlar Şekil 2.2.1 da gösterilmiştir.



Şekil 2.2.1 – SD-EnClass metodu çalışma şekli

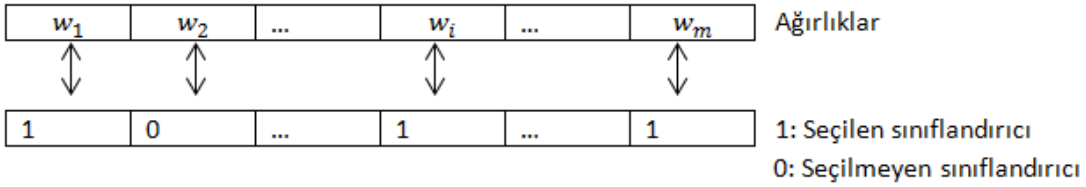
Katman1 de sınıflandırıcıların modelleri oluşturulmuştur. Katman2 de SD-EnClass metodu uygulanarak bu modeller birleştirilip yeni bir model oluşturulmuştur. Katman3 de, katman1 deki sınıflandırıcılardan model havuzları oluşturulmuştur. Ve son olarak katman4 de ise katman2 ve katman3 deki en iyi iki model birleştirilmiştir.

2.3 İki kodlu kromozomlu GA

Tian-zhong ve diğerleri tarafından önerilen bu metotta [5] öncelikle sınıflandırıcılar bir filtreden geçirilmektedir. Belirli bir doğruluk oranının üstünde olan sınıflandırıcılar aday sınıflandırıcı olarak seçilmektedir. Bu şekilde sınıflandırıcı kümesinde ilk elemeyi yapmış olur.

Aday sınıflandırıcılar için daha sonra GA kullanılarak en iyi topluluk kurulmaya çalışılmıştır. Bu topluluğu kurarlarken her bir sınıflandırıcıya bir ağırlık atamışlardır. Önerilen GA metodunda, çoklu sınıflandırıcılar içerisinden hem en iyi sınıflandırıcıları, hem de bu sınıflandırıcıların ağırlıklarını seçmeye çalışmışlardır. Bunun için yapılan GA uygulamasında iki kodlu kromozom kullanılmıştır. Bu

kromozomlardan bir tanesi ağırlıkları temsil ederken, diğeri sınıflandırıcıları temsil etmektedir. 2 kodlu kromozom örneği Şekil 2.3.1 de gösterilmiştir.



Şekil 2.3.1 – 2 kodlu kromozom örneği

İki kodlu kromozomda genetik değişim, kodların ayrı şekilde yapılması ile gerçekleşmektedir. Hem ağırlıklar, hem de sınıflandırıcılar için ayrı genetik değişim yapılır. Bunun için 2 noktalı genetik değişim kullanmışlardır. Mutasyonu da kodlar farklı şekilde yaparlar. Sınıflandırıcı kod için rastgele 4 tane sınıflandırıcı seçip, değerlerini değiştirmişlerdir ("0" "1" e, "1" de "0" a çevrilmiştir). Ağırlıklar içinse daha farklı bir metot kullanılmıştır. Her bir ağırlık için rastgele "0" veya "1" üretilmiş ve bu değere göre 2.3.1'deki işlem yapılmıştır.

$$w_k' = \begin{cases} w_k + \Delta(t, 1 - w_k), & n = 0 \\ w_k + \Delta(t, 1 + w_k), & n = 1 \end{cases} \tag{2.3.1}$$

Burada "n" değeri "0" veya "1" olan rastgele bir sayıdır. $\Delta(t, y)$ fonksiyonu ise [0,y] arasında bir değer dönmektedir. Bu fonksiyon 2.3.2'de gösterilmiştir.

$$\Delta(t, y) = y\left(1 - r^{\left(1 - \frac{t}{M}\right)^P}\right) \tag{2.3.2}$$

Burada *t,* iterasyon numarası, *r,* [0,1] arasında rastgele bir sayı, *M,* maksimum iterasyon numarası ve mutasyon parametresi *P* ise ilk olarak tanımlanmış bir değerdir.

Bu fonksiyonlar sayesinde GA çalışma süresince yerel maksimumda takılma sorununu aşmaya çalışmışlardır. Uygunluk fonksiyonu ise sınıflandırıcı topluluğun doğruluğu olarak belirlenmiştir. İterasyon sayısı belirledikleri bir sayıya ulaşınca algoritmayı sonlandırmışlardır.

2.4 GA ve Sınıflandırıcı Topluluk

Benisha Fida ve diğerleri tarafından önerilen [6] GA ile oluşturulan sınıflandırıcı toplulukta SVM sınıflandırıcısının 3 farklı çeşidi kullanılmıştır. Bunlar doğrusal çekirdek, polinom çekirdek ve radyal tabanlı çekirdektir. Bu sınıflandırıcıların ağırlıkları GA ile belirlenmiştir. Uygunluk fonksiyonu olarak doğruluk, duyarlılık ve özgüllük fonksiyonları kullanılmıştır. Önerilen metodun çalışma şekli Şekil 2.4.1 de verilmiştir.

Doğruluk: Sınıflandırıcıların doğruluk performansının değerlendirilmesinde en çok kullanılan fonksiyonlardan biridir. 2.4.1'de gösterildiği gibi, doğruluk fonksiyonu, doğru sınıflandırılan örnek sayısının toplam örnek sayısına bölünmesi ile gerçekleşir.

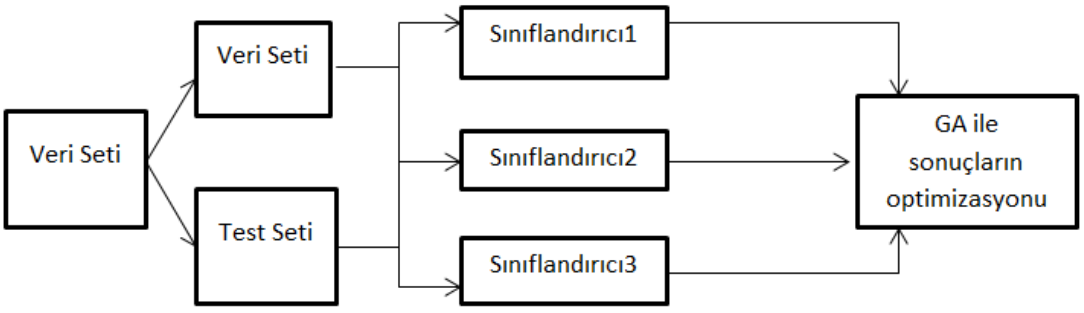
$$Doğruluk = \frac{Doğru\ sınıflandırılan\ örnek\ sayısı}{Toplam\ örnek\ sayısı} \tag{2.4.1}$$

Duyarlılık: Duyarlılık fonksiyonu 2.4.2'de gösterildiği gibi doğru pozitif değerinin, doğru negatif ve yanlış negatif değerlerinin toplamına bölünmesi ile hesaplanır.

$$Duyarlılık = \frac{TP}{TP+FN} \tag{2.4.2}$$

Özgüllük: Özgüllük fonksiyonu 2.4.3'de gösterildiği gibi doğru negatif değerinin, doğru negatif ve yanlış pozitif değerlerinin toplamına bölünmesi ile hesaplanır.

$$Özgüllük = \frac{TN}{TN+FP} \tag{2.4.3}$$



Şekil 2.4.1 – GA ve sınıflandırıcı topluluk çalışma şekli

2.5 GA ve Farklı Sınıflandırıcıları Birleştirme

Dina A. Salem ve diğerleri tarafından 2012 yılında önerilen bu metot da [7] mikroçip veri seti üzerinden gen seçimi için genetik algoritma ile 3 farklı sınıflandırıcıyı birleştirmişlerdir. Bu metot da başlangıç topluluğu 3 farklı şekilde oluşturulmuştur. Bunlar filtreleme, sarma ve gömme yöntemleridir. Filtreleme yöntemi basitçe sıralama yöntemidir. Veri setindeki her bir gene sıra verilir ve içlerinden özel bir eşiği geçen ve en fazla bilgiye sahip olan genler seçilir. Sarma yöntemi ise sınıflandırıcı üzerinden gelişmektedir. Bu teknikte bütün veri seti üzerinden farklı ufak gen setleri alınır ve bu setler ile sınıflandırıcı modeller kullanılarak test edilir. En iyi sınıflandırıcı doğruluğunu (2.4.1) veren gen seti seçilir ve diğerleri elenir. Bu teknikte sınıflandırıcı modelleri "kara kutu" olarak ifade edilir ve hiç değiştirilmez. Gömme yöntemi ise sarma yönteminden çok farklı değildir. Fakat bu yöntemde sınıflandırıcı "kara kutu" olarak ifade edilmez çünkü her bir alt gen seti için yeni bir sınıflandırıcı model oluşturulur ve bu model üzerinden test gerçekleştirilir. En iyi gen seti seçimi sarma yöntemi ile aynıdır. En fazla sınıflandırıcı doğruluğunu veren set, gen seti olarak seçilir. Bu teknikler, genetik algoritmanın başlangıç topluluğunu oluşturmak için kullanılan tekniklerdir.

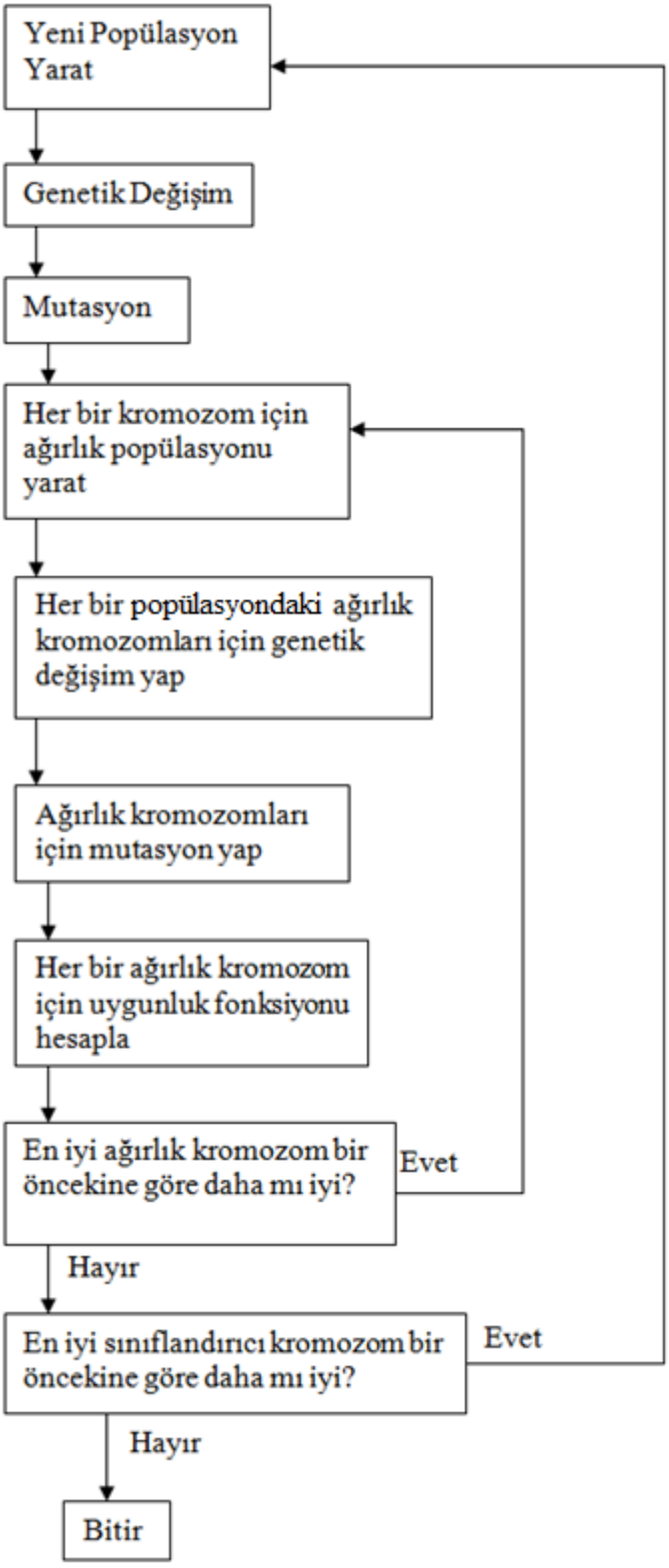
Genetik algoritmanın uygunluk fonksiyonu olarak 3 tane sınıflandırıcı kullanılmıştır. Bunlar LDA, KNN ve SVM sınıflandırıcılarıdır. Veri seti olarak da Lenfoma veri setini kullanmışlardır. Bu veri setinden çıkan sonuçlara bakıldığında GA-SVM ve GA-KNN tam doğruluk verirken GA-LDA %97.06 oranında doğruluk vermiştir.

3 AĞIRLIKLI ÇOKLU SINIFLANDIRICI KULLANARAK BİYOLOJİK VERİLERİN TAHMİNİ

Bu bölümde önerilen yöntem, detaylı olarak açıklanmıştır. 2 seviyeli genetik algoritmanın incelenmesi derinleştirilmiş, başlangıç popülasyonu oluşturmaya değinilmiş, sınıflandırıcı ve ağırlık kromozomu tanıtılmış, bu kromozomlarda genetik değişim ve mutasyonun nasıl yapıldığı açıklanmıştır. Daha sonra uygunluk fonksiyonu ve yeni nesil için kromozomların nasıl seçildiği anlatılmıştır.

3.1 Genel Bakış

Ağırlıklı çoklu sınıflandırıcı kullanmak, genetik algoritma tabanlı bir çalışmadır. Bu çalışma iç içe 2 aşamalı genetik algoritmadan oluşur. İlk aşamada sınıflandırıcılar seçilirken ikinci aşamada ise bu sınıflandırıcılara uygun ağırlıklar belirlenmeye çalışılmıştır. Bu çalışmada her bir kromozom aslında 2 kromozomun birleşiminden meydana gelir. İlk kromozom sınıflandırıcıların kimlikleri ile ifade edilirken, ikinci kromozom ise bunların ağırlıklarını ifade eder. Yeni bir nesil oluşurken, uygunluk fonksiyonu iki kromozomu da hesaba katarak belirlenir. Her bir sınıflandırıcı kromozom için ağırlık kromozomu hesaplandıktan sonra uygunluk fonksiyon sonucu, bu iki kromozomun sonucu olur. Bir kromozomun sonucu ne kadar iyi ise, bir sonraki nesilde bulunma olasılığı o kadar artar. Uygunluk fonksiyon sonucu olarak *f-skorlama* değeri kullanılmıştır. Bu değer karışıklık matrisi (Tablo 3) kullanılarak bulunur. Karışıklık matrisi sayesinde *Duyarlılık* ve *Hassaslık* hesaplanır. Daha sonra ise bunlardan *f-skorlama* değeri hesaplanarak kromozomun uygunluk fonksiyon sonucu belirlenir. Bütün kromozomların sonuçları hesaplandıktan sonra bir sonraki nesil için seçim işlemi gerçekleştirilir. Burada bir sonraki nesile geçecek kromozomlar seçilir. Bu sayede her bir yeni nesilde ulaşılmak istenen noktaya biraz daha yaklaşılır. Programın çalışma aşamaları Şekil 3.1.1 ve Şekil 3.1.2 de gösterilmiştir.



Şekil 3.1.1 – Programın çalışma aşamaları

1. İlk Popülasyonu yarat
   İlk Popülasyon kullanıcı tarafından alınan kromozom sayısı ve kromozom uzunluğuna göre yaratılır. Bu kromozom sayısı her yeni popülasyonda korunur.
   1.1 Herbir kromozom için
      1.1.1 Yeni Ağırlık popülasyonu yarat
      1.1.2 Ağırlık kromozomlan için genetik değişim yap
      1.1.3 Ağırlık kromozomlan için mutasyon yap
      1.1.4 Uygunluk fonksiyonu ile kromozomların sonuçlarını hesapla
      1.1.5 En iyi kromozomları seç.
2. Yeni popülasyonu oluştur.
      2.1 Bütün sınıflandırıcı kromozomlar için genetik değişim yap.
      2.2 Bütün sınıflandırıcı kromozomlar için mutasyon yap.
      2.3 Bütün sınıflandırıcı kromozomlar için ağırlıkları hesapla.
      2.3.1 Bütün sınıflandırıcı kromozomlar için 1.1-1.1.5 adımlarını tekrarla.
      2.4 Bütün kromozomlar için uygunluk fonksiyonlarını hesapla.
         2.4.1 Bütün sınıflandırıcılar için tahmin edilmiş değerleri al.
         2.4.2 Test veri seti için sınıflandırıcı sonuçlarını al.
         2.4.3 Sınıflandırıcı sonuçları ile ağırlıkları çarp ve her bir test verisi için topla
         2.4.4 Her bir test verisi için sınıflandırıcı sonuçlarını hesapla
         2.4.5 Her bir sınıflandırıcı için F-skorlama değerlerini hesapla
3. Kromozomlar üstünden seçim işlemini gerçekleştir.
      3.1 Kromozomların F-skorlama değerlerine göre sırala
      3.2 İlk yaratılan sayı kadar en iyi F-skorlama değerlerine sahip kromozomları seç
4. Eğer en iyi uygunluk fonksiyonuna sahip kromozom bir önceki ile aynı veya daha kötü ise, bitir. Eğer değilse 1. 2. Adıma geri dön.

Şekil 3.1.2 – Programın ayrıntılı çalışma aşamaları

3.2 Yeni Popülasyon Yaratma

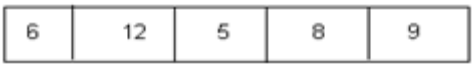
Bu aşamada kromozomlar popülasyonu yaratılmaktadır. Her bir kromozom kullanıcı tarafından belirlenen gene sahiptir. Yaptığımız çalışmada her bir veri seti için bu sayı 1'den başlayarak 29'a kadar değişmektedir. Her bir sayı için çalışma tekrarlanmaktadır. Bu şekilde hangi sayıda maksimum değer alınabileceği gösterilmiştir. Veri setlerine göre, maksimum değer alınan gen sayıları değişiklik göstermektedir.

3.3 Kromozom

Yaptığımız çalışmada 1 kromozomda iki farklı gen listesi vardır. Bunlardan ilki sınıflandırıcıları tutan, diğeri ise bun sınıflandırıcıların ağırlıklarını tutan gen listesidir.

3.4 Sınıflandırıcı Gen Listesi

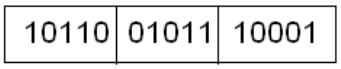
Sınıflandırıcı gen listesi içlerinde sınıflandırıcıların kimliklerini tutarlar (Şekil 3.4.1)



Şekil 3.4.1 - Sınıflandırıcı kromozom örneği

3.5 Ağırlık Gen Listesi

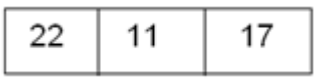
Ağırlık gen listesi, içlerinde sınıflandırıcıların ağırlıklarını tutarlar. Bu sayede her bir sınıflandırıcının bir ağırlığı olmuş olur. Ağırlık gen listesinde her bir gen 5 bit ile ifade edilir (Şekil 3.5.1).



Şekil 3.5.1 - Ağırlık kromozom örneği

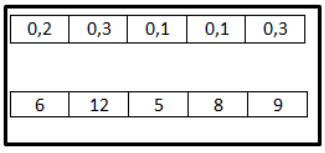
Bu gen listesi, genetik değişim ve mutasyondan geçirildikten sonra, 10'luk tabana çevrilir ve her bir ağırlık popülasyonundaki gen listesi için sınıflandırıcı ile birlikte test yapılır. Eğer ağırlık popülasyonu yeni bir nesil oluşturacak ise, 10'luk sistemdeki genler yeniden 2'lik sisteme dönüştürülür (Şekil 3.5.2).

Sınıflandırıcı gen listesi için ağırlık genleri seçildikten sonra bu ağırlıklar 10'luk tabanına dönüştürülüp saklanırlar. En iyi ağırlıklar bulunduktan sonra bütün ağırlıklar [0,1] arasına normalize edilir.



Şekil 3.5.2 - Ağırlık kromozomunun 10'luk tabandaki hali.

Sonuç olarak, yaptığımız çalışma için, tek bir kromozom gösterildiği gibi saklanır (Şekil 3.5.3).



Şekil 3.5.3 - Kromozom örneği

3.6 Genetik Değişim

Genetik değişim iki kromozom arasında geçen bir operasyondur. Bu operasyon sonucunda iki yeni kromozom daha oluşur. Değişim yapılacak kromozomlar topluluk içerisinden rastgele seçilir ve o popülasyon için bir daha değişime seçilmezler. Önerilen yöntem için, yaptığımız genetik değişimde seçilen 2 kromozomun, değişim işleminden geçirilebilmesi için oran 0.9 olarak belirlenmiştir (Tablo 1). Değişim metodu olarak da standart genetik değişim seçilmiştir. Burada her bir gene sırayla bakılır. Her bir gen çifti için rastgele [0,1] arası bir sayı üretilir. Eğer bu sayı eşik değerinin (0.5) üstünde ise değişim yapılır. 0.5 seçilmesinin sebebi ise değişime ortak bir oran verebilmektir. Eşik değeri ne kadar az tutulursa, değişim o kadar fazla olacaktır. Bir diğer yandan bu değer ne kadar büyük tutulursa değişim o kadar az olacaktır. Yeni bireylerin çeşitliliğini artırmak için eşik değerini orta değer seçmek en ideal durumdur.

Tablo 1 – Genetik değişim olasılıkları

Genetik Değişim olasılığı	0.9
Gen değişim olasılığı	0.5

Değişim sayesinde popülasyondaki kromozom sayısı yaklaşık iki katına çıktığı ve çeşitlilik arttığı için, daha iyi bir seçim gerçekleştirebilmeye olanak sağlayacaktır. Bu aşamadan sonra kromozomlara mutasyon uygulanarak çeşitlilik arttırılmaya çalışılır.

3.7 Mutasyon

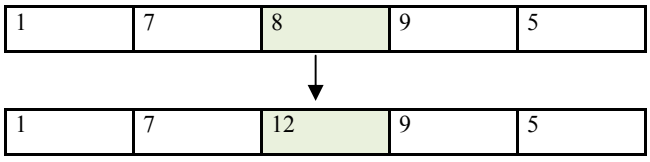
Mutasyon, tek bir kromozom içerisine dışarıdan etki eden değişikliktir. Bu aşamada her bir sınıflandırıcı kromozom için, kromozomdaki tüm genler dolaşılır. Her yeni gene geçildiğinde rastgele bir sayı üretilir. Bu rastgele sayı eğer 0.8'in altında ise bu gen için mutasyon yapılmaz (Algoritma 2). Eğer 0.8'in üstünde ise, o gendeki sınıflandırıcı, listede olmayan başka bir sınıflandırıcı ile değiştirilir (Şekil 3.7.1).

Algoritma 2 : Mutasyon(k)

```
1: Girdiler:
2:   k % Mutasyon yapılacak kromozom
3:
4: Çıktılar:
5:   k % Mutasyon yapılmış kromozom
6:
7: for i=0 to KromozomBoyutu
8:                    rasgeleSayi = rastgeleSayiUret(0,1)
9:                    İf(rastgeleSayi > 0.8)
10:                     k = siniflandiriciDegistir(k,i)
11:                   end if
12: end for
13: return k
```

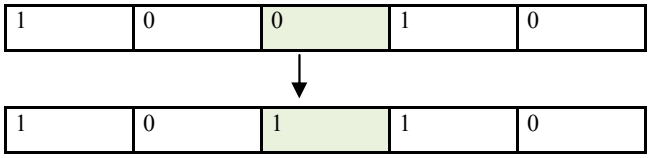
Algoritma 3 : sınıflandırıcıDeğiştir(k)

```
1: Girdiler:
2:   k % Mutasyon yapılacak kromozom
3:  i % Mutasyon yapılacak gen numarası
4:
5: Çıktılar:
6:   k % Mutasyon yapılmış kromozom
7:
8:  yeniSiniflandiriciKimlik =kromozomdaOlmayanSiniflandiriciBul(k)
9:  k = siniflandiriciDegistir(k,yeniSiniflandirici,i)
10: return k
```



Şekil 3.7.1 – Sınıflandırıcı kromozom mutasyonu

Ağırlık kromozomlarının mutasyona uğraması da aynı işlemle olmaktadır. Fakat ağırlık kromozomlarındaki genler 10'luk taban yerine 2'lik tabanda tutulduğu için, Bu kromozomlardaki mutasyon, gendeki 0 değerini 1 yapmak, veya 1 değerini 0 yapmak olacaktır (Şekil 3.7.2).



Şekil 3.7.2 – Ağırlık kromozom mutasyonu

3.8 Her bir Sınıflandırıcı Kromozom için Ağılık Popülasyonu Yaratma

Genetik algoritma ile çoklu sınıflandırıcılar hesaplanırken, her bir sınıflandırıcı kromozom için ayrı bir genetik algoritma uygulanır ve bu kromozomdaki sınıflandırıcıların ağılıkları hesaplanır. Bunun için her bir sınıflandırıcı kromozoma bağlı bir ağırlık popülasyonu oluşturulmaktadır. Bu popülasyondaki kromozom sayısı 100 olarak belirlenmiştir. Ayrıca kromozomlardaki gen sayıları ise sınıflandırıcı kromozomdaki gen sayısı ile eşit tutulmuştur. Popülasyon oluşturulurken her bir sınıflandırıcı gen için 5 bitlik rastgele sayılar atanmıştır (Algoritma 4).

Algoritma 4 : AgirlikPopulasyonYarat(populasyonBoyut,kromozomGenBoyut)

1: **Girdiler:**
2: *populasyonBoyut k % Yaratılacak olan topluluk büyüklüğü*
3: *kromozomGenBoyut % Kromozom gen sayısı*
4:
5: **Çıktılar:**
6: *kList % Kromozom popülasyonu*
7:
8: kList[popülasyonBoyut]
9: **for** i=0 **to** popülasyonBoyut

```
10:   agirlikKromozom[kromozomGenBoyut * 5]
11:           for j=0 to (kromozomGenBoyut * 5)
12:              bit = rastgeleBitYarat()
13:               agirlikKromozom[j] = bit
14:           end for
15:   kList [i] = agirlikKromozom
16: end for
17: return kList
```

Sonuç olarak sınıflandırıcı kromozom uzunluğunun 5 katı kadar bir liste oluşur.

3.9 Ağırlık Kromozomu Genetik Değişim

Ağırlık kromozomlarında genetik değişim yapılması, sınıflandırıcı kromozomların değişim yapılması ile aynı metodu kullanmaktadır. Kullanılan olasılıklar Tablo 2' de belirtilmiştir.

Tablo 2 – Ağırlık kromozom genetik değişim parametreleri

Genetik Değişim olasılığı	0.9
Gen değişim olasılığı	0.5

3.10 Ağırlık Kromozom Mutasyon

Ağırlık kromozomlarının mutasyon işlemi sınıflandırıcı kromozomların mutasyonundan biraz daha farklıdır. Bu aşamada mutasyon işlemi ikili tabanda olduğu için bütün genler için rastgele [0,1] arası sayı üretilir. Eğer bu sayı 0.8 den büyük ise o gendeki bit tersi ile değiştirilir (Algoritma 5).

Algoritma 5 : MutasyonAgirlik(k)

```
1: Girdiler:
2:   k % Mutasyon yapılacak kromozom
3:
4: Çıktılar:
5:   k % Mutasyon yapılmış kromozom
6:
7: for i=0 to KromozomBoyutu
8:    rasgeleSayi = rastgeleSayıYarat()
9:        if(rastgeleSayi > 0.8)
10:           k = bitDegistir(k,i)
11:       end if
12: end for
13: return k
```

Bir ağırlık kromozomu genetik değişim ve mutasyon işleminden sonra içindeki değerler ikilik tabandan 10'luk tabana dönüştürülür (Algoritma 6).

Algoritma 6 : OnlukSistemeDonustur(agirlikPopülasyon)

```
1: Girdiler:
2:   agirlikPopülasyon % Dönüştürülecek olan kromozom popülasyonu
3:
4: Çıktılar:
5:   agirlikPopülasyonD % Onluk sisteme dönüştürülmüş kromozom popülasyonu
6:
7: for i=0 to agirlikPopülasyonBoyut
8:     kromozom = agirlikPopülasyon [i]
9:    kromozomOnluk [kromozomBoyut/5]
10:          for j=0 to kromozomBoyut; j=j+5
11:              onlukSayi = onlukSistemeCevir(kromozom, j , j+5)
12:              kromozomOnluk [j/5] = onlukSayi
```

```
13:         end for
14:    agirlikPopülasyonD [i] = kromozomOnluk
14: end for
15: return agirlikPopülasyonD
```

3.11 Ağırlık kromozomu genetik değişim ve mutasyon hatası

Ağırlık kromozomlarında değişim ve mutasyon işlemleri ikilik tabanda yapıldıkları ve daha sonra bu ikilik tabandaki genler 10'luk tabana çevrildikleri için, bu çevirme işleminde bazı gen değerleri sıfır değerini alabilir. Bu hiç istenilmeyen bir durumdur çünkü bir sınıflandırıcının ağırlığının olmaması, o sınıflandırıcının orda olmadığı anlamına gelir. Böyle bir durumdan kurtulmak için 10'luk sistemde, sıfır gen değerine sahip kromozomlarda, bu gen değeri tekrar 5 bite dönüştürülür ve içlerinden rastgele bir tanesinin değeri 1 yapılır. Bu şekilde sıfır ağırlıklı hiçbir sınıflandırıcı bulunmaz (Algoritma 7).

Algoritma 7 : MutasyonVeGenDeğişimHatası(agirlikPopülasyon)

```
 1: Girdiler:
 2:   agirlikPopülasyon % Onluk sistemdeki ağırlık popülasyonu
 3:
 4: Çıktılar:
 5:   agirlikPopülasyon % Hata giderilmiş onluk sistemdeki ağırlık popülasyonu
 6:
 7: for i=0 to agirlikPopülasyonBoyut
 8:     kromozom = agirlikPopülasyon[i]
 9:         for j=0 to kromozomBoyut
10:             if (kromozom[j] == 0) then
11:                 ikilikListe [5]
12:                 rastgeleSayı = rastgeleSayiUret(0,5)
13:                 ikilikListe = bitDegistir(ikilikListe,rastgeleSayı)
14:                 kromozom[j] = = onlukSistemeCevir(ikiliListe)
```

```
15:             end if
16:         end for
17:         agirlikPopülasyon [i] = kromozom
18: end for
19: return agirlikPopülasyon
```

3.12 Uygunluk Fonksiyonu

Yaptığımız çalışmada uygunluk fonksiyonu, daha önce belirlenmiş olan test verisi içindeki örnekleri tek tek dolaşmaktadır. Her bir test verisi için sınıflandırıcıların verdikleri sonuçlar, ağırlıkları ile çarpılmaktadır. Daha sonra bu sonuçlar toplanır ve en yüksek değere sahip olan sonuç, o test verisinin o satırı için sonuç kabul edilir. Bu durum bütün test setindeki satırlar için yapılır. En son bu sonuçlar üzerinden *F-skorlama* değerleri hesaplanır (Algoritma 8). Bu değer, o sınıflandırıcı kromozomun ve ağırlık kromozomunun sonucudur. Bu sonuç daha sonra yeni popülasyon oluşturulurken en iyilerin seçimi için kullanılacaktır.

Algoritma 8 : UygunlukFonksiyonu(sK, aK, sonucL)

```
 1: Girdiler:
 2:  sK % Sınıflandırıcı kromozom
 3: aK % Ağırlık kromozom
 4: T % Test veri seti
 5: sonucL % Sınıflandırıcı sonuç sözlüğü
 6:
 7: Çıktılar:
 8:  f-skorlama % Fonksiyon sonu kromozomun f-skorlama değeri
 9:
10: sonucListesi
11: testSonucListesi
12: for j=0 to T satır sayısı
13:     testDatası = T [j]
```

```
14:       sonucListesiTemizle()
15:           for k=0 to sKBoyut
16:               sonuc = sınıflandır(sK[k] , testDatası)
17:               sonucListesi.add(sonuç,aK[k])
18:           end for
19:       enCokOyAlanSınıf = baskınSınıfıHesapla(sonucListesi)
20:       testSonucListesi.add(enCokOyAlanSınıf)
21:  end for
22:  f-skorlama = f-skorlamaHesapla(sonucListesi,testDatası,sonucL)
23: return fÖlcüm
```

Algoritma 9 : BaskınSınıfıHesapla(sonucListesi)

```
 1: Girdiler:
 2:    sonucListesi % Sınıflandırıcıların bulduğu sınıflar ve ağırlıkları
 3:
 4: Çıktılar:
 5:   enCokOyAlanSınıf % En çok oy alan sınıf
 6:
 7:  sonucSınıfListesi
 8: for i=0 to sonucListesiBoyut
 9:     if sonucSınıfListesi içinde sonucListesi[i]  yoksa then
10:         sonucSınıfListesi.add(sonucListesi[i])
11:      end if
12:     agirlikGuncelle(sonucListesi[i].sonuc, sonucListesi[i].ağırlık)
13:  end for
14: return maksAğırlığaSahipSınıf
```

3.12.1 Uygunluk Fonksiyonu Hesaplaması

Uygunluk fonksiyonu hesaplamasında, ilk olarak doğru pozitif oranı, doğru negatif oranı, yanlış pozitif oranı ve yanlış negatif değerleri hesaplanır. Daha sonra bu verilerden Duyarlılık (2.4.2) ve Hassaslık (3.12.1) değlerleri hesaplanır. En son olarak da *f-skorlama* (3.12.2) değeri hesaplanarak fonksiyon sonucu belirlenmiş olunur. Aşağıda bu oranları daha detaylı açıklamaları bulunmaktadır.

Doğru pozitif oranı (TP): Gerçek sonucun pozitif, ve tahmin edilen sonucun pozitif olduğu örnek sayısıdır.
Doğru negatif oranı (TN): Gerçek sonucun negatif, ve tahmin edilen sonucun negatif olduğu örnek sayısıdır.
Yanlış Pozitif Oranı (FP): Gerçek sonucun negatif, ve tahmin edilen sonucun pozitif olduğu örnek sayısıdır.
Yanlış Negatif Oranı (FN): Gerçek sonucun pozitif, ve tahmin edilen sonucun negatif olduğu örnek sayısıdır.
Test verisi sınıflandırıldıktan sonra ortaya karmaşıklık matrisi (Tablo 3) ortaya çıkar.

	Gerçek Sonuçlar	
Tahmin edilen sonuçlar	TP	FP
	FN	TN

Tablo 3 – Karmaşıklık matrisi

Bu matris içerisinde tp, fp, fn, tn değerlerini bulundurur. Bu değerlerden ise *f-skorlama* hesaplanır.

$$Hassaslık = \frac{TP}{TP+FP} \tag{3.12.1}$$

$$F - skorlama = \frac{2\ x\ Duyarlılık\ x\ Hassaslık}{Duyarlılık+Hassaslık} \tag{3.12.2}$$

F-skorlama değeri 1'e ne kadar yakınsa, üretilmiş olan sınıflandırıcı ve ağırlık kromozomları o kadar iyi demektir.

3.13 En iyi kromozomlarının seçimi

Kromozomlar seçilirken uygunluk fonksiyonu sonucunda ortaya çıkan f-skorlama değerlerine bakılır. Bu değer 1'e ne kadar yakınsa kromozomun sınıflandırma doğruluğu o kadar iyi demektir. Kromozomlar için yeni bir popülasyon oluşturulurken de, en iyi f-skorlama değerine sahip olan kromozomlar seçilir (Algoritma 10).

Algoritma 10 : Seçim(kList)

```
1: Girdiler:
2:   kList % Toplulukdaki bütün kromozomlar
3:
4: Çıktılar:
5:   sKList %  Bir sonraki nesil için seçilen kromozomlar
6:
7:  kList = sıralaUygulukFonk(kList)
8:  sKlist = enIyiKromozomlarıSec(kList)
9:    return sKlist
```

Bu aşamada ayrıca, bir önceki popülasyonda, en iyi kromozomun f-skorlama değeri ile kıyaslama yapılır. Bu kıyaslama sonucunda eğer en iyi kromozom 5 kez tekrarlanmışsa yeni bir popülasyon yaratılmaz ve algoritma bitirilir. Eğer 5 kez tekrarlanmamışsa, seçilmiş olan 100 ağırlık kromozomu ikilik sisteme dönüştürülerek tekrardan genetik değişim ve mutasyon işlemine sokulur.

Algoritma 11 : Bitirme(yeniPopülasyon, Pe)

```
1: Girdiler:
2:   popülasyon % Genetik değişim ve mutasyon yapılmış kromozom topluluğu
3:  Pe % Bir önceki popülasyondaki en iyi kromozom sonucu
4:
5: Çıktılar:
6:   bitir %  Bir sonraki nesil için seçilen ağırlık kromozomları
7:
8: sayım
9: if yeniPopülasyonEnİyiKromozomSonuc == Pe then
10:         sayım = sayım +1
11: else
12:          sayım = 0
13: end if
14: if sayım == 5 then
15:             bitir = true
16: else
17:             bitir = false
18: end if
19: popülasyon.enIyiKromozomEkle(popülasyon)
20: return bitir
```

Ağırlıkların belirlenmesi aşamasında, her bir ağırlık kromozomu için, algoritma bitirildikten sonra en iyi sonuca sahip kromozom, uygunluk fonksiyonu içerisinde tanımlanmış olan sınıflandırıcı kromozoma atanır. Artık o sınıflandırıcı kromozomun ağırlık kromozomu ve bu ikilinin sonucu belirlenmiş olur.

Seçim algoritması hem sınıflandırıcı hem de ağırlık kromozom popülasyonunu azaltmak için kullanılır.

4 UYGULAMA VE VERİ SETLERİ

Bu bölümde deneylerin hangi ortamda alındığı, kullanılan veri setlerinin açıklamaları ve bu setlerin özelliklerinden bahsedilmiştir. Daha sonra deneyler için veri setlerinin nasıl düzenlendiğinden bahsedilmiştir. Veri setleri toplamda 10 parçaya bölündükleri için, bölüm 3 de anlatılan yöntem 10 farklı alt veri seti için çalıştırılmaktadır. Bu yüzden ortaya 10 farklı sonuç çıkmaktadır. Bu bölümde son olarak, bu sonuçların nasıl birleştirildiği anlatılmaktadır.

4.1 Uygulama

Uygulama Java dilinde, WEKA[8] kütüphanesi kullanılarak, Windows 7 platformu üzerinde geliştirilmiştir. Tüm sonuçlar i7 3.40 GHz işlemcisine sahip, 16 GB RAM bilgisayar üzerinde alınmıştır.

4.2 Kullanılan Veri Setleri

Uygulamanın test edilmesi aşamasında Microçip veri setleri[13] kullanılmıştır. Bu veri setleri aşağıda daha ayrıntılı bir şekilde açıklanmıştır. Kullanılan veri setlerin genel açıklamaları ise Tablo 4 de gösterilmiştir.

Lenfoma: Bağışıklık sisteminin urlarıdır. Lenf düğümlerinde çıkan, ve lenfositlerden oluşan urların tümüne lenfoma denir [9]

Akciğer: Akciğer dokularındaki hücrelerin kontrolsüz çoğaldığı bir hastalıktır. [10]

MLL: Üç sınıflı lösemi veri setidir.[11]

Göğüs: Göğüs kanseri meme hücrelerinde başlayan kanser türüdür. Akciğer kanserinden sonra, dünyada görülme sıklığı en yüksek olan kanser türüdür.[12]

Yumurtalık: Yumurtalıklarda oluşan kanser türüdür.[14]

ALL-MLL-4: 4 Sınıflı Lösemi veri setidir. [18]

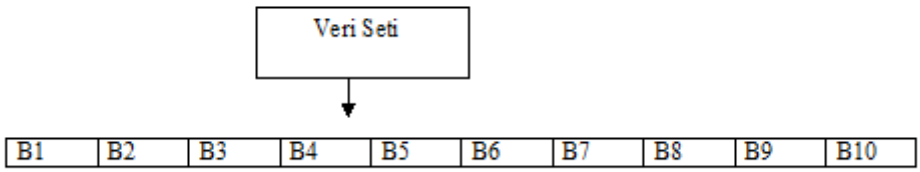
Mavi Hücreli Tümör: Küçük, yuvarlak, mavi hücreli tümör veri setidir.[15]

Veri Seti Adı	Örnek Sayısı	Gen Sayısı	Sınıf Sayısı
Lenfoma	62	4026	3
Akciğer	42	7129	5
MLL	72	12582	3
Yumurtalık	253	15154	2
ALL-MLL-4	72	7129	4
Göğüs	97	24481	2
Mavi hücre Tümör	83	2308	4

Tablo 4 – Veri setleri özellikleri

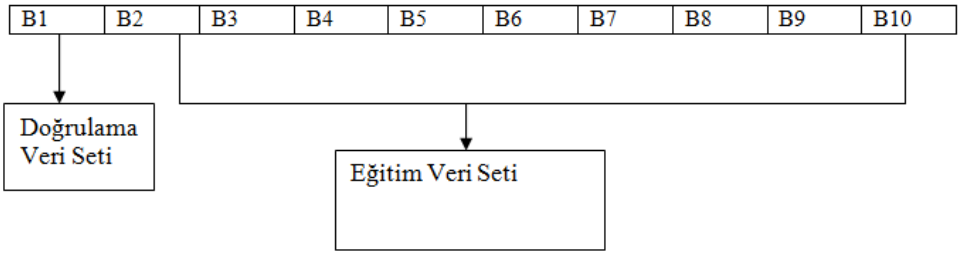
4.2.1 Veri Seti düzenlemesi

Deneyler alınırken öncelikle veri seti 10 parçaya bölünür (Şekil 4.2.1).



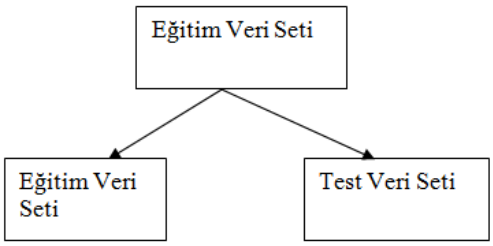
Şekil 4.2.1 – Veri setinin bölünmesi

Burada her bir "B", veri seti içerisindeki %10'luk kısma karşılık gelmektedir. Veri seti 10 parçaya bölündükten sonra, sırasıyla, her bir "B" test verisi olarak seçilir (Şekil 4.2.2). Kalan veri seti ise eğitim seti olarak seçilir. Yaratılan sınıflandırıcı model, doğrulama verisi ile test edilir. Bu sayede yapılan modelin doğruluğu ölçülmüş olur.



Şekil 4.2.2 – Eğitim ve Doğrulama veri seti seçimi

Sınıflandırıcıların ve ağırlıklarının modeli oluşturulurken de bir test verisine ihtiyaç duyulduğu için eğitim veri setini %90'ı eğitim veri seti, %10'u test veri seti şeklinde ikiye bölünür (Şekil 4.2.3).

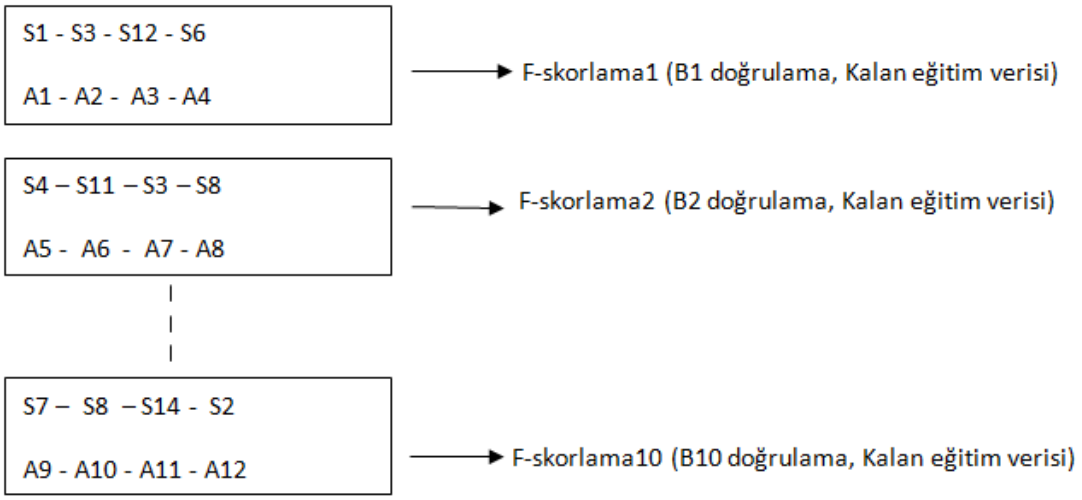


Şekil 4.2.3 – Eğitim veri seti bölünmesi

Model oluşturulurken B1…B9 doğrulama veri seti olarak sırayla seçilir. Modeller oluşturulurken bir veri seti toplamda 10 kez işleme sokulur. Her bir farklı "B" için, farklı modeller yaratılır. Böylelikle bir veri seti için farklı modeller ve bunların sonuçlarını karşılaştırma imkânı olur.

4.2.2 Algoritma Sonucu Hesaplanan değerlerin işlenmesi

Bir veri seti için bir model yaratıldıktan sonra, bu model doğrulama verisi ile test edilir ve *f-skorlama* değeri hesaplanır. Bu işlem 10 farklı doğrulama veri seti ile tekrarlanır ve toplamda 10 farklı *f-skorlama* değerinin ortalaması alınır (Şekil 4.2.4). Bu sayede veri seti için oluşturulan modellerin doğrulukları gösterilmiş olur.



Şekil 4.2.4 – Alınan f-skorlama değerinin işlenmesi

Burada her bir "S", sınıflandırıcıları temsil ederken, her bir "A" ise bu sınıflandırıcıların ağırlıklarını temsil ederler. Bir veri seti için sonuç, 10 iterasyon sonucunda elde edilen *f-skorlama* değerlerinin ortalamasıdır (4.2.1). Bu ortalama, o veri seti için ne kadar doğru modeller oluşturulduğunu gösterir.

$$\frac{\sum_{i=1}^{10} f\text{-}skorlama[i]}{10} \tag{4.2.1}$$

Bu denklemde "*f-skorlama*"; algoritma sonucunda oluşturulmuş modellerin, doğrulama veri seti ile test edildikten sonra ortaya çıkan değeridir.

5 DENEYLER

Yaptığımız çalışmanın başarısının karşılaştırılabilmesi için, bir üst bölümde anlatılan 7 veri seti kullanılmıştır. Bu veri setlerini kullanarak alınan sonuçlar ve her bir sınıflandırıcının tek başına aldığı sonuçlar karşılaştırılmıştır. Karşılaştırmada *Özgüllük*, *Duyarlılık*, *F-skorlama*, *Doğruluk* fonksiyonları ve 29 adet sınıflandırıcı kullanılmıştır. Sınıflandırıcılara atanan kimlik ve adlar Tablo 5'de gösterilmiştir.

Deneylerde her bir veri seti için ayrı tablo hazırlanmıştır. Bu tablolardan ilki farklı sınıflandırıcı kombinasyonlarının sonuçlarının bulunduğu tablodur. Burada en iyi 29 sınıflandırıcı seçimine kadar deneyler tekrarlanmıştır. Ayrıca bu tabloda her bir satır için kromozom büyüklüğü, sınıflandırıcı sayısı kadar seçilmiştir. Böylelikle en iyi kaç sınıflandırıcı seçileceği görülmeye çalışılmıştır. Diğer bir tablo ise sınıflandırıcı havuzundaki bütün sınıflandırıcıların bireysel performanslarını göstermektedir. En sona ise bu sonuçların maksimum değerlerini gösteren tablo eklenmiştir.

Tablo 5 – Sınıflandırıcılar ve kimlik bilgileri

Sınıflandırıcı Kimlik	Sınıflandırıcı Adı
1	Bayes Net
2	Complement Naive Bayes [22]
3	Dmnb Text [23]
4	Jrip [24]
5	Naive Bayes Multinomial [25]
6	Naive Bayes Multinomial Updateable [25]
7	Naive Bayes Updateable [26]
8	LibSvm [41]
9	Rbf Network
10	Simple Logistic [27]
11	Smo [28]
12	Ib1 [29]
13	Ibk [29]
14	Lwl [30]
15	Conjunctive Rule
16	Decision Table [31]
17	Oner [32]
18	Part [33]
19	Ridor [34]
20	Zeror
21	Bf Tree [35]
22	Decision Stump
23	Ft [36]

24	J48	[37]
25	Lad Tree	[38]
26	Lmt	[39]
27	Random Forest	[40]
28	Random Tree	
29	Rep Tree	

Göğüs Kanseri

Göğüs kanseri ile yapılan deney sonuçları Tablo 6, Tablo 7 ve Tablo 8'de verilmiştir.

Tablo 6 – Göğüs kanseri veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.661	0.71	0.693	0.685
2	0.747	0.75	0.748	0.748
3	0.764	0.804	0.788	0.784
4	0.853	0.825	0.836	0.839
5	0.827	0.825	0.825	0.826
6	0.836	0.816	0.823	0.826
7	0.793	0.784	0.787	0.788
8	0.813	0.825	0.82	0.819
9	0.8	0.773	0.783	0.787
10	0.832	0.846	0.84	0.839
11	0.785	0.795	0.79	0.79
12	0.816	0.805	0.809	0.81
13	0.824	0.815	0.818	0.82
14	0.831	0.826	0.828	0.829
15	0.836	0.825	0.829	0.831
16	0.781	0.784	0.782	0.782
17	0.814	0.836	0.826	0.825
18	0.873	0.866	0.868	0.87
19	0.831	0.835	0.833	0.833
20	0.812	0.806	0.808	0.809
21	0.807	0.827	0.818	0.817
22	0.815	0.85	0.835	0.832
23	0.815	0.848	0.834	0.832
24	0.82	0.86	0.842	0.84
25	0.809	0.846	0.83	0.828
26	0.806	0.847	0.83	0.827
27	0.783	0.828	0.809	0.806
28	0.83	0.845	0.838	0.837
29	0,769	0,772	0,77	0,77

Tablo 7 – Göğüs kanseri bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.831	0.830	0.832	0.831
2	0.690	0.700	0.697	0.695
3	0.666	0.639	0.643	0.653
4	0.820	0.816	0.818	0.819
5	0.690	0.700	0.697	0.695
6	0.456	0.556	0.521	0.506
7	0.720	0.757	0.743	0.739
8	0.446	0.554	0.517	0.500
9	0.693	0.739	0.723	0.716
10	0.825	0.845	0.839	0.835
11	0.837	0.849	0.843	0.843
12	0.785	0.798	0.793	0.791
13	0.785	0.798	0.793	0.791
14	0.807	0.813	0.811	0.810
15	0.691	0.684	0.683	0.687
16	0.802	0.781	0.785	0.791
17	0.744	0.720	0.729	0.732
18	0.788	0.813	0.805	0.800
19	0.806	0.803	0.804	0.805
20	0.446	0.554	0.517	0.500
21	0.803	0.812	0.810	0.808
22	0.798	0.812	0.805	0.805
23	0.817	0.837	0.831	0.827
24	0.761	0.783	0.777	0.772
25	0.798	0.808	0.804	0.803
26	0.817	0.838	0.831	0.828
27	0.794	0.798	0.796	0.796
28	0.827	0.823	0.826	0.825
29	0.779	0.793	0.789	0.786

Tablo 8 – Göğüs kanseri en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	0.873	0.866	0.868	0.870
Sınıflandırıcı	0.837	0.849	0.843	0.843

Göğüs kanseri veri seti ile yapılan deney sonuçlarına bakıldığında, en iyi f-skorlama değerini veren sınıflandırıcı sayısının 18 olduğu görülmektedir (Tablo 6). Bunun anlamı, havuzdan seçilen 18 sınıflandırıcı, havuzdaki en iyi kombinasyonu sağlayan sınıflandırıcılardır.

Tablo 7'de bu veri seti ile ilgili bireysel sınıflandırıcıların sonuçları görülmektedir. Bu sonuçlara bakıldığında ise en iyi sonucu veren sınıflandırıcı 11 numaralı sınıflandırıcıdır. Diğerlerine göre daha iyi bir ortalama sağlamıştır. En iyi sonuçların karşılaştırmaları Tablo 8'de gösterilmektedir. Bu tabloya bakıldığında Göğüs kanseri veri seti için özgüllük, duyarlılık, doğruluk ve f-skorlama metotlarının hepsinde önerilen yöntem bireysel sınıflandırıcılara göre daha iyi sonuç vermiştir.

ALL-MLL-4 Kanser

ALL-MLL-4 kanser veri seti ile yapılan deney sonuçları Tablo 9, Tablo 10 ve Tablo 11 da gösterilmiştir.

Tablo 9 – ALL-MLL-4 kanser veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.705	0.707	0.706	0.706
2	0.972	0.96	0.965	0.966
3	0.89	0.86	0.873	0.875
4	0.848	0.878	0.864	0.863
5	0.963	0.932	0.946	0.947
6	0.892	0.932	0.913	0.912
7	0.986	0.942	0.963	0.964
8	0.908	0.901	0.904	0.905
9	1.0	0.985	0.992	0.992
10	0.916	0.958	0.938	0.937
11	0.911	0.946	0.929	0.929
12	0.928	0.946	0.937	0.937
13	0.964	0.985	0.974	0.975
14	0.886	0.916	0.902	0.901
15	0.92	0.917	0.918	0.919
16	0.951	0.973	0.962	0.962
17	0.987	0.973	0.979	0.98
18	0.916	0.958	0.938	0.937
19	0.951	0.973	0.962	0.962
20	0.97	0.948	0.958	0.959
21	0.884	0.932	0.909	0.908
22	0.951	0.958	0.954	0.955
23	0.975	0.96	0.966	0.967
24	0.939	0.96	0.949	0.95
25	0.964	0.985	0.974	0.975
26	0.964	0.985	0.974	0.975
27	0.939	0.96	0.949	0.95
28	0.987	0.973	0.979	0.98
29	0,949	0,914	0,93	0,931

Tablo 10 – ALL-MLL-4 kanser veri seti bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.921	0.890	0.903	0.906
2	0.904	0.841	0.867	0.872
3	0.899	0.720	0.785	0.810
4	0.876	0.873	0.880	0.875
5	0.933	0.885	0.903	0.909
6	0.526	0.486	0.466	0.506
7	0.921	0.931	0.928	0.926
8	0.526	0.487	0.466	0.506
9	0.944	0.906	0.922	0.925
10	0.993	0.944	0.967	0.969
11	0.975	0.960	0.967	0.968
12	0.919	0.917	0.919	0.918
13	0.919	0.917	0.919	0.918
14	0.909	0.819	0.857	0.864
15	0.845	0.762	0.800	0.803
16	0.885	0.835	0.854	0.860
17	0.872	0.771	0.809	0.822
18	0.910	0.876	0.894	0.893
19	0.972	0.888	0.921	0.930
20	0.520	0.480	0.458	0.500
21	0.976	0.923	0.944	0.949
22	0.782	0.776	0.788	0.779
23	0.997	0.971	0.983	0.984
24	0.913	0.878	0.897	0.896
25	0.998	0.968	0.982	0.983
26	0.993	0.944	0.967	0.969
27	0.895	0.900	0.899	0.897
28	0.878	0.750	0.800	0.814
29	0.836	0.855	0.848	0.845

Tablo 11 – ALL-MLL-4 veri seti en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	1.000	0.985	0.992	0.992
Sınıflandırıcı	0,998	0.971	0,983	0,984

ALL-MLL-4 kanseri veri seti ile yapılan deney sonuçlarına bakıldığında**,** en iyi f-skorlama değerini veren sınıflandırıcı sayısının 9 olduğu görülmektedir (Tablo 9). Bunun anlamı, havuzdan seçilen 9 sınıflandırıcı havuzdaki en iyi kombinasyonu sağlayan sınıflandırıcılardır. Tablo 10'da bu veri seti ile ilgili bireysel sınıflandırıcıların sonuçları görülmektedir. Bu sonuçlara bakıldığında ise en iyi

sonucu veren sınıflandırıcı 23 numaralı sınıflandırıcıdır. Diğerlerine göre daha iyi bir ortalama sağlamıştır. En iyi sonuç karşılaştırmaları Tablo 11'de gösterilmektedir. Bu tabloya bakıldığında ALL-MLL-4 kanser veri seti için duyarlılık, doğruluk ve f-skorlama metotlarının hepsinde önerilen yöntem bireysel sınıflandırıcılara göre daha iyi sonuç vermiştir. Özgüllük ölçümünde ise her iki değer de maksimum sonuç vermiştir.

Akciğer Kanser

Akciğer kanser veri seti ile yapılan deney sonuçları Tablo 12, Tablo 13 ve Tablo 14'de gösterilmiştir.

Tablo 12 – Akciğer kanser veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.704	0.798	0.761	0.751
2	0.744	0.866	0.816	0.805
3	0.834	0.887	0.863	0.86
4	0.796	0.902	0.856	0.849
5	0.9	0.925	0.913	0.913
6	0.886	0.926	0.907	0.906
7	0.887	0.93	0.91	0.908
8	0.895	0.941	0.919	0.918
9	0.857	0.925	0.894	0.891
10	0.907	0.96	0.934	0.933
11	0.863	0.926	0.897	0.894
12	0.898	0.955	0.928	0.926
13	0.881	0.94	0.913	0.911
14	0.893	0.945	0.92	0.919
15	0.861	0.926	0.897	0.894
16	0.881	0.94	0.913	0.911
17	0.884	0.946	0.917	0.915
18	0.875	0.931	0.905	0.903
19	0.891	0.95	0.922	0.921
20	0.867	0.936	0.904	0.901
21	0.867	0.945	0.909	0.906
22	0.887	0.95	0.921	0.919
23	0.881	0.94	0.913	0.911
24	0.925	0.965	0.946	0.945

25	0.929	0.975	0.953	0.952
26	0.949	0.98	0.964	0.964
27	0.929	0.975	0.953	0.952
28	0.944	0.975	0.959	0.959
29	0,951	0,961	0,955	0,956

Tablo 13 – Akciğer kanser veri seti bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.924	0.843	0.880	0.883
2	0.838	0.889	0.876	0.864
3	0.564	0.801	0.718	0.682
4	0.901	0.911	0.910	0.906
5	0.935	0.889	0.912	0.912
6	0.309	0.691	0.576	0.500
7	0.915	0.861	0.887	0.888
8	0.309	0.691	0.576	0.500
9	0.844	0.916	0.887	0.880
10	0.917	0.959	0.942	0.938
11	0.931	0.972	0.956	0.951
12	0.875	0.936	0.912	0.905
13	0.875	0.936	0.912	0.905
14	0.762	0.808	0.794	0.785
15	0.651	0.675	0.673	0.663
16	0.834	0.854	0.847	0.844
17	0.678	0.786	0.750	0.732
18	0.889	0.920	0.909	0.905
19	0.878	0.866	0.874	0.872
20	0.309	0.691	0.576	0.500
21	0.879	0.840	0.860	0.860
22	0.570	0.745	0.686	0.657
23	0.924	0.963	0.948	0.943
24	0.893	0.912	0.906	0.903
25	0.933	0.946	0.943	0.939
26	0.917	0.959	0.942	0.938
27	0.843	0.911	0.883	0.877
28	0.837	0.847	0.844	0.842
29	0.779	0.847	0.826	0.813

Tablo 14 – Akciğer kanser veri seti en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	0.951	0.980	0.964	0.964
Sınıflandırıcı	0.935	0.972	0.956	0.951

Akciğer kanseri veri seti ile yapılan deney sonuçlarına bakıldığında, en iyi f-skorlama değerini veren sınıflandırıcı sayısının 26 olduğu görülmektedir (Tablo 12). Bunun anlamı, havuzdan seçilen 26 sınıflandırıcı havuzdaki en iyi kombinasyonu sağlayan sınıflandırıcılardır. Tablo 13'de bu veri seti ile ilgili bireysel sınıflandırıcıların sonuçları görülmektedir. Bu sonuçlara bakıldığında ise en iyi sonucu veren sınıflandırıcı 11 numaralı sınıflandırıcıdır. Diğerlerine göre daha iyi bir ortalama sağlamıştır. En iyi sonuçların karşılaştırmaları Tablo 14 'de gösterilmektedir. Bu tabloya bakıldığında Göğüs kanser veri seti için özgüllük, duyarlılık, doğruluk ve f-skorlama metotlarının hepsinde önerilen yöntem bireysel sınıflandırıcılara göre daha iyi sonuç vermiştir.

MLL Kanser

MLL kanser veri seti ile yapılan deney sonuçlarıTablo 15, Tablo 16 ve Tablo 17'de gösterilmiştir.

Tablo 15 – MLL kanser veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.711	0.648	0.669	0.679
2	0.898	0.791	0.835	0.844
3	0.884	0.801	0.835	0.843
4	0.949	0.875	0.908	0.912
5	0.98	0.942	0.96	0.961
6	0.975	0.946	0.959	0.96
7	0.975	0.942	0.957	0.959
8	0.975	0.917	0.944	0.946
9	0.991	0.971	0.98	0.981
10	1.0	1.0	1.0	1.0
11	0.994	0.985	0.989	0.99
12	1.0	1.0	1.0	1.0

13	0.989	0.985	0.986	0.987
14	0.964	0.957	0.96	0.96
15	1.0	1.0	1.0	1.0
16	1.0	1.0	1.0	1.0
17	0.991	0.971	0.98	0.981
18	0.985	0.932	0.957	0.959
19	0.991	0.971	0.98	0.981
20	0.997	0.985	0.99	0.991
21	0.972	0.957	0.964	0.965
22	0.994	0.985	0.989	0.99
23	0.984	0.958	0.97	0.971
24	0.986	0.957	0.97	0.971
25	0.991	0.971	0.98	0.981
26	0.988	0.971	0.979	0.98
27	0.994	0.985	0.989	0.99
28	0.994	0.985	0.989	0.99
29	0,958	0,887	0,919	0,923

Tablo 16 – MLL kanser veri seti bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.974	0.904	0.933	0.939
2	0.965	0.889	0.922	0.927
3	0.901	0.718	0.783	0.810
4	0.933	0.861	0.887	0.897
5	0.979	0.917	0.941	0.948
6	0.563	0.437	0.452	0.500
7	0.995	0.983	0.988	0.989
8	0.563	0.437	0.452	0.500
9	0.975	0.940	0.955	0.957
10	0.990	0.969	0.978	0.979
11	0.993	0.960	0.974	0.976
12	0.967	0.890	0.922	0.928
13	0.967	0.890	0.922	0.928
14	0.911	0.787	0.827	0.849
15	0.854	0.575	0.650	0.714
16	0.943	0.870	0.899	0.907
17	0.953	0.924	0.935	0.938
18	0.981	0.953	0.965	0.967
19	0.943	0.873	0.898	0.908
20	0.563	0.437	0.452	0.500
21	0.998	0.994	0.996	0.996

22	0.888	0.699	0.764	0.794
23	0.995	0.983	0.989	0.989
24	0.981	0.953	0.965	0.967
25	0.980	0.970	0.975	0.975
26	0.990	0.969	0.978	0.979
27	0.920	0.837	0.868	0.879
28	0.913	0.802	0.838	0.858
29	0.965	0.933	0.941	0.949

Tablo 17 – MLL kanser veri seti en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	1.000	1.000	1.000	1.000
Sınıflandırıcı	0.998	0.994	0.996	0.996

MLL kanseri veri seti ile yapılan deney sonuçlarına bakıldığında, en iyi f-skorlama değerini veren sınıflandırıcı sayılarının 10, 12, 15 ve 16 olduğu görülmektedir (Tablo 15). Bunun anlamı, havuzdan seçilen en iyi kombinasyonlar MLL veri seti için birden fazladır. Tablo 16'da bu veri seti ile ilgili bireysel sınıflandırıcıların sonuçları görülmektedir. Bu sonuçlara bakıldığında ise en iyi sonucu veren sınıflandırıcı 21 numaralı sınıflandırıcıdır. Diğerlerine göre daha iyi bir ortalama sağlamıştır. En iyi sonuçların karşılaştırmaları Tablo 17'de gösterilmektedir. Bu tabloya bakıldığında MLL kanser veri seti için özgüllük, duyarlılık, doğruluk ve f-skorlama metotlarının hepsinde önerilen yöntemin sonuçlarının 1.0 olduğu görülmektedir. Bu sonuç bütün örneklerin doğru sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcıların bireysel olarak en iyi sonuçlarına bakıldığında ise, sonuçların 1'e çok yakın olduğu görülmektedir.

Yumurtalık Kanseri

Yumurtalık kanseri veri seti ile yapılan deney sonuçları Tablo 18, Tablo 19 ve Tablo 20'de gösterilmiştir.

Tablo 18 – Yumurtalık veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.468	0.652	0.597	0.56
2	0.934	0.928	0.93	0.931
3	0.959	0.96	0.959	0.959
4	0.978	0.968	0.972	0.973
5	0.9	0.944	0.923	0.922
6	0.986	0.98	0.982	0.983
7	0.965	0.984	0.974	0.974
8	0.984	0.984	0.984	0.984
9	0.97	0.976	0.972	0.973
10	0.977	0.988	0.982	0.982
11	0.994	0.992	0.992	0.993
12	1.0	1.0	1.0	1.0
13	0.973	0.976	0.974	0.975
14	0.981	0.988	0.984	0.984
15	0.981	0.988	0.984	0.984
16	0.987	0.996	0.991	0.991
17	0.977	0.988	0.982	0.982
18	0.985	0.992	0.988	0.988
19	0.994	0.992	0.992	0.993
20	0.977	0.988	0.982	0.982
21	0.974	0.984	0.978	0.979
22	0.99	0.992	0.99	0.991
23	0.987	0.996	0.991	0.991
24	0.982	0.992	0.987	0.987
25	0.978	0.988	0.982	0.983
26	0.974	0.984	0.978	0.979
27	0.987	0.996	0.991	0.991
28	0,982	0,988	0,984	0,985
29	0,984	0,992	0,987	0,988

Tablo 19 – Yumurtalık veri seti bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.912	0.913	0.913	0.913
2	0.890	0.883	0.886	0.886
3	0.471	0.647	0.593	0.559
4	0.951	0.956	0.954	0.954
5	0.890	0.883	0.886	0.886

6	0.389	0.586	0.528	0.488
7	0.924	0.898	0.910	0.911
8	0.822	0.871	0.850	0.846
9	0.834	0.821	0.827	0.827
10	0.971	0.971	0.971	0.971
11	0.963	0.964	0.963	0.963
12	0.904	0.918	0.912	0.911
13	0.904	0.918	0.912	0.911
14	0.920	0.937	0.929	0.929
15	0.929	0.945	0.937	0.937
16	0.961	0.964	0.962	0.962
17	0.932	0.941	0.937	0.937
18	0.942	0.949	0.945	0.945
19	0.950	0.960	0.955	0.955
20	0.388	0.583	0.526	0.486
21	0.953	0.953	0.953	0.953
22	0.921	0.933	0.928	0.927
23	0.971	0.971	0.971	0.971
24	0.927	0.933	0.931	0.930
25	0.967	0.968	0.967	0.967
26	0.971	0.971	0.971	0.971
27	0.901	0.933	0.920	0.917
28	0.871	0.864	0.868	0.868
29	0.916	0.926	0.921	0.921

Tablo 20 – Yumurtalık veri seti en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	1.000	1.000	1.000	1.000
Sınıflandırıcı	0.971	0.971	0.971	0.971

Yumurtalık kanseri veri seti ile yapılan deney sonuçlarına bakıldığında, en iyi f-skorlama değerini veren sınıflandırıcı sayılarının 12 olduğu görülmektedir (Tablo 18). Bunun anlamı, havuzdan seçilen 12 sınıflandırıcı havuzdaki en iyi kombinasyonu sağlayan sınıflandırıcılardır. Tablo 19'da bu veri seti ile ilgili bireysel sınıflandırıcıların sonuçları görülmektedir. Bu sonuçlara bakıldığında ise en iyi sonucu veren sınıflandırıcı 10, 23 ve 26 numaralı sınıflandırıcılardır. Diğerlerine göre daha iyi bir ortalama sağlamışlardır. En iyi sonuçların karşılaştırmaları Tablo

20'de gösterilmektedir. Bu tabloya bakıldığında Yumurtalık kanseri veri seti için özgüllük, duyarlılık, doğruluk ve f-skorlama metotlarında önerilen yöntemin sonuçlarının hepsinin 1.0 olduğu görülmektedir. Bu sonuç bütün örneklerin doğru sınıflandırıldığı anlamına gelmektedir. Sınıflandırıcıların bireysel olarak en iyi sonuçlarına bakıldığında ise, sonuçların 1'e yakın ve 0.971 olduğu görülmektedir.

Lenfoma

Lenfoma kanser veri seti ile ilgili yapılan deney sonuçları Tablo 21, Tablo 22 ve Tablo 23'de gösterilmiştir.

Tablo 21 – Lenfoma veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.707	0.809	0.769	0.758
2	0.921	0.914	0.916	0.917
3	0.897	0.952	0.926	0.924
4	0.972	0.957	0.964	0.965
5	0.942	0.9	0.919	0.921
6	0.957	0.942	0.948	0.95
7	0.967	0.957	0.961	0.962
8	0.989	0.985	0.986	0.987
9	0.917	0.957	0.938	0.937
10	0.989	0.985	0.986	0.987
11	0.978	0.971	0.974	0.975
12	0.989	0.985	0.986	0.987
13	0.989	0.985	0.986	0.987
14	0.989	0.985	0.986	0.987
15	1.0	1.0	1.0	1.0
16	0.909	0.94	0.925	0.925
17	0.989	0.985	0.986	0.987
18	1.0	1.0	1.0	1.0
19	0.989	0.985	0.986	0.987
20	0.978	0.971	0.974	0.975
21	0.989	0.985	0.986	0.987
22	0.989	0.985	0.986	0.987
23	0.978	0.971	0.974	0.975
24	0.989	0.985	0.986	0.987
25	0.989	0.985	0.986	0.987
26	0.989	0.985	0.986	0.987
27	0.989	0.985	0.986	0.987

28	0.989	0.985	0.986	0.987
29	0,845	0,935	0,901	0,892

Tablo 22 – Lenfoma bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.989	0.986	0.987	0.988
2	0.937	0.878	0.900	0.908
3	0.778	0.852	0.828	0.815
4	0.899	0.914	0.909	0.907
5	1.000	1.000	1.000	1.000
6	0.374	0.626	0.549	0.500
7	0.979	0.971	0.975	0.975
8	0.627	0.748	0.704	0.688
9	0.910	0.940	0.929	0.925
10	0.989	0.986	0.987	0.988
11	1.000	1.000	1.000	1.000
12	1.000	1.000	1.000	1.000
13	1.000	1.000	1.000	1.000
14	0.951	0.928	0.939	0.940
15	0.935	0.778	0.839	0.857
16	0.984	0.957	0.969	0.970
17	0.989	0.971	0.980	0.980
18	0.959	0.955	0.958	0.957
19	0.904	0.805	0.843	0.855
20	0.374	0.626	0.549	0.500
21	0.979	0.928	0.947	0.954
22	0.935	0.807	0.856	0.871
23	0.989	0.986	0.987	0.988
24	0.959	0.955	0.958	0.957
25	0.989	0.986	0.987	0.988
26	0.989	0.986	0.987	0.988
27	0.979	0.971	0.975	0.975
28	0.937	0.893	0.912	0.915
29	0.905	0.836	0.862	0.870

Tablo 23 – Lenfoma veri seti en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	1.000	1.000	1.000	1.000
Sınıflandırıcı	1.000	1.000	1.000	1.000

Lenfoma veri seti ile yapılan deney sonuçlarına bakıldığında en iyi f-skorlama değerini veren sınıflandırıcıların 15 ve 18 olduğunu görülmektedir. En iyi kombinasyonlar 15 ve 18 tane sınıflandırıcı seçimi ile gerçekleşmiştir (Tablo 21). Sınıflandırıcıların bireysel olarak performanslarına bakıldığında ise (Tablo 22), en iyi sınıflandırıcıların 11,12 ve 13 numaralı sınıflandırıcılar olduğu görülmektedir.Tablo 23'e bakıldığında Tablo 21 ve Tablo 22 'deki maksimum sonuçlar görülmektedir. Burada hem önerilen yöntem, hem de sınıflandırıcıların bireysel performansları maksimum seviyededir.

Mavi Hücre Tümör

Mavi hücre tümör veri seti ile yapılan deney sonuçları Tablo 24 , Tablo 25 ve Tablo 26'da gösterilmiştir.

Tablo 24 – Mavi hücre tümör veri seti ile yapılan deney sonuçları

Sınıflandırıcı Sayısı	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	0.825	0.6	0.675	0.712
2	0.921	0.702	0.788	0.812
3	0.959	0.905	0.93	0.932
4	0.924	0.904	0.912	0.914
5	0.98	0.94	0.959	0.96
6	0.967	0.927	0.946	0.947
7	0.988	0.963	0.975	0.976
8	0.979	0.951	0.964	0.965
9	0.99	0.963	0.976	0.977
10	0.988	0.963	0.975	0.976
11	0.98	0.963	0.97	0.971
12	0.929	0.894	0.909	0.911
13	0.973	0.941	0.956	0.957
14	0.994	0.988	0.99	0.991
15	1.0	0.988	0.993	0.994
16	0.983	0.963	0.972	0.973
17	1.0	0.989	0.994	0.994
18	1.0	0.988	0.993	0.994
19	0.988	0.963	0.975	0.976
20	0.987	0.976	0.981	0.981
21	0.987	0.976	0.981	0.981

22	0.98	0.951	0.965	0.966
23	0.983	0.963	0.972	0.973
24	0.987	0.976	0.981	0.981
25	0.992	0.976	0.983	0.984
26	0.977	0.952	0.964	0.965
27	0.992	0.987	0.989	0.99
28	1.0	0.988	0.993	0.994
29	0,99	0,963	0,976	0,977

Tablo 25 – Mavi hücre tümör veri bireysel sınıflandırıcı sonuçları

Sınıflandırıcı Kimlik	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
1	1.0	0.988	0.993	0.994
2	0.956	0.865	0.904	0.911
3	0.779	0.588	0.645	0.683
4	0.958	0.917	0.934	0.937
5	0.967	0.865	0.908	0.916
6	0.599	0.401	0.420	0.500
7	0.992	0.975	0.982	0.983
8	0.599	0.401	0.420	0.500
9	0.971	0.939	0.953	0.955
10	0.998	0.988	0.993	0.993
11	1.0	0.988	0.993	0.994
12	0.977	0.926	0.948	0.952
13	0.977	0.926	0.948	0.952
14	0.917	0.820	0.860	0.869
15	0.803	0.485	0.569	0.644
16	0.876	0.760	0.805	0.818
17	0.874	0.593	0.680	0.733
18	0.967	0.917	0.939	0.942
19	0.927	0.856	0.885	0.891
20	0.599	0.401	0.420	0.500
21	0.932	0.770	0.833	0.851
22	0.807	0.571	0.638	0.689
23	0.998	0.976	0.987	0.987
24	0.964	0.903	0.930	0.934
25	0.969	0.939	0.953	0.954
26	0.998	0.988	0.993	0.993
27	0.945	0.890	0.915	0.918
28	0.935	0.796	0.849	0.865
29	0.909	0.867	0.886	0.888

Tablo 26 – Mavi hücre tümör veri seti en iyi sonuçlar

Maksimum	Özgüllük	Duyarlılık	F-skorlama	Doğruluk
Önerilen Yöntem	1.000	0.989	0.994	0.994
Sınıflandırıcı	1.000	0.988	0.993	0.994

Mavi hücre tümör veri seti için yapılan deney sonuçlarına bakıldığında önerilen metot üzerinden en iyi kombinasyonu sağlayan sınıflandırıcı kümesi 17 olarak görülmektedir (Tablo 24). Sınıflandırıcıların bireysel sonuçları karşılaştırıldığında ise (Tablo 25), en iyi f-skorlama sonucunu veren sınıflandırıcının 1, 10, 11 ve 26 olduğu görülmektedir. Her ne kadar maksimum sonuçlar birbirine çok yakınsa da (Tablo 26), iki seviyeli genetik algoritma ile yapılan çalışmanın f-skorlama değeri, en iyi sınıflandırıcının bireysel sonucundan daha fazladır.

En iyi sonuçların kombinasyonları

Yapılan deneylerde ortaya çıkan sonuç, 10 farklı kombinasyonun ortalamasıdır. Bu sonuç veri setinin 10 farklı parçaya bölünmesinden oluştuğu için, her bir parçada farklı bir kombinasyon oluşmakta ve her bir parçaya farklı ağırlıklar atanmaktadır. Tablo 27, Tablo 28, Tablo 29, Tablo 30, Tablo 31, Tablo 32 ve Tablo 33'de 10 farklı parçanın her biri için, soldan sağa veri setinin test yapıldığı parçalar belirtilirken, yukarıdan aşağı ise bu parçaların testinden oluşan kombinasyonların sınıflandırıcı numaraları ve ağırlıkları belirtilmiştir.

Tablo 27 – Göğüs kanser veri seti en iyi kombinasyon ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	-	-	-	0.050	0.073	0.134	0.012	-	0.005	0.048
2	-	-	0.010	0.022	-	-	-	0.121	0.068	0.010
3	-	-	-	-	0.034	0.009	-	-	-	0.014
4	0.046	0.104	0.036	-	0.073	0.117	0.030	-	0.041	-
5	0.129	-	0.041	-	0.034	0.039	0.077	-	0.077	0.130
6	0.012	0.114	0.005	0.072	0.034	0.065	0.048	0.101	-	0.005
7	0.087	0.014	0.046	0.194	-	0.048	-	-	-	-
8	-	-	0.157	0.043	0.044	0.004	-	0.061	-	0.116
9	-	0.114	-	0.050	-	0.048	-	0.015	-	0.005
10	0.091	0.043	0.015	0.022	-	0.061	0.065	0.030	-	0.019

11	-	-	0.122	-	-	0.121	0.167	0.020	0.113	-
12	0.087	0.014	-	0.151	0.015	-	0.018	-	0.122	0.048
13	0.083	-	0.157	0.007	0.063	0.004	0.030	0.096	0.005	-
14	0.071	0.047	0.010	0.007	0.053	-	0.095	0.040	0.108	-
15	0.012	0.019	0.066	-	0.049	0.061	0.012	-	-	0.126
16	-	0.033	-	0.022	0.068	0.082	-	0.101	0.041	-
17	-	0.062	-	-	-	-	0.030	0.131	0.005	-
18	0.108	0.019	0.015	0.007	0.049	-	-	0.010	0.014	0.048
19	0.008	0.057	0.020	-	0.150	0.004	0.089	0.056	0.014	0.092
20	0.025	-	0.015	0.007	-	-	0.054	-	0.050	-
21	0.029	-	0.096	-	-	-	-	0.056	0.036	-
22	-	0.043	-	0.036	-	-	0.036	0.010	-	0.019
23	0.108	-	0.102	-	0.005	0.048	-	-	0.104	0.150
24	0.004	0.066	0.005	0.043	0.083	-	0.054	-	-	0.024
25	0.004	-	-	0.194	-	-	0.012	0.030	0.095	0.014
26	-	0.028	-	0.036	0.117	0.030	-	0.015	-	0.106
27	0.050	0.095	0.081	0.036	0.034	0.095	0.155	-	0.099	0.024
28	-	0.123	-	-	0.024	-	-	0.056	-	-
29	0.046	0.005	-	-	-	0.030	0.018	0.051	0.009	-

Tablo 28 – ALL-MLL-4 veri seti en iyi kombinasyon ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	-	0.155	-	0.085	-	-	-	0.211	-	-
2	-	-	-	-	-	0.102	0.133	-	-	-
3	-	-	-	-	-	-	-	0.114	-	0.099
4	0.012	-	-	-	0.190	-	-	-	0.095	-
5	-	0.109	0.101	-	-	0.024	-	0.061	-	-
6	-	-	0.081	-	-	-	0.240	-	0.168	-
7	-	0.217	-	-	-	-	-	-	-	0.257
8	-	-	-	-	-	-	0.240	-	-	0.020
9	-	-	0.061	0.037	-	0.087	-	-	-	-
10	-	0.070	0.030	0.024	0.107	-	0.027	0.211	-	-
11	-	-	-	-	-	0.055	0.013	-	0.022	-
12	-	0.163	-	0.110	-	0.087	0.093	-	0.007	0.099
13	-	0.124	0.040	0.232	-	-	0.040	-	-	-
14	0.212	0.047	-	0.049	-	-	-	0.167	-	0.059
15	-	-	0.121	0.256	0.036	0.102	0.187	-	0.153	0.178
16	-	0.093	-	-	-	-	-	-	0.044	-
17	-	-	0.051	-	0.060	-	-	0.123	-	-
18	-	-	-	-	0.119	0.228	-	-	-	-
19	0.012	-	-	-	0.036	-	0.027	-	-	-
20	-	-	-	-	-	-	-	-	-	0.129
21	-	-	0.293	-	0.274	0.197	-	-	-	-
22	0.012	-	0.222	0.012	-	-	-	-	0.146	-
23	-	0.023	-	-	-	-	-	-	-	0.059
24	-	-	-	-	0.167	-	-	-	0.168	0.099
25	0.012	-	-	-	-	0.118	-	0.053	-	-
26	0.200	-	-	0.195	-	-	-	-	-	-

27	0.012	-	-	-	-	-	-	0.026	-	-
28	0.338	-	-	-	-	-	-	0.035	0.197	-
29	0.188	-	-	-	0.012	-	-	-	-	-

Tablo 29 – Akciğer kanseri veri seti en iyi kombinasyon ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	0.064	0.043	0.011	0.087	0.057	0.004	0.036	-	0.101	0.004
2	0.011	0.036	0.008	0.034	-	0.027	0.003	0.004	0.076	-
3	0.032	0.096	0.011	0.034	0.017	0.027	0.062	0.066	-	0.016
4	-	0.033	0.038	0.090	0.004	0.027	0.068	0.004	0.080	0.070
5	0.004	0.003	0.011	0.062	0.031	0.004	0.032	0.037	0.040	0.012
6	0.093	0.023	0.023	0.050	0.013	0.030	0.023	0.066	0.022	-
7	0.014	0.007	0.008	0.015	0.004	0.019	0.016	0.033	0.018	0.066
8	-	0.023	0.094	0.031	0.013	0.011	0.068	0.044	0.018	0.008
9	0.025	0.073	0.008	0.031	0.127	-	0.065	0.052	-	0.066
10	0.039	0.070	0.041	0.022	0.017	0.004	0.088	0.026	0.109	0.019
11	0.061	0.060	0.056	0.065	0.061	0.053	0.026	-	0.007	0.074
12	0.032	0.013	0.083	0.053	0.061	0.015	0.019	0.100	0.007	0.081
13	0.043	0.003	0.079	0.022	0.048	0.057	0.032	0.004	0.040	0.058
14	-	0.020	-	0.015	0.004	0.004	0.029	0.004	0.011	0.004
15	0.007	0.023	0.038	0.003	0.004	0.095	0.003	0.037	0.018	0.019
16	0.107	0.030	0.068	0.022	0.031	0.011	-	0.063	0.007	0.012
17	0.004	0.033	0.053	0.084	0.031	0.004	0.016	0.074	0.033	0.023
18	0.029	0.050	0.075	0.056	0.100	0.068	0.029	0.074	0.004	0.047
19	0.043	-	0.004	0.003	0.057	0.023	0.097	0.026	0.014	0.023
20	0.082	0.056	0.064	0.071	0.066	0.080	0.088	0.044	-	0.004
21	0.043	0.010	0.019	-	0.039	-	0.026	-	0.018	0.031
22	0.004	0.103	0.026	0.019	0.066	0.019	0.023	0.022	0.004	0.027
23	0.111	0.020	0.026	0.046	-	0.084	0.026	0.022	0.054	0.058
24	0.029	0.017	0.068	-	0.004	0.110	-	0.103	0.083	0.074
25	0.004	0.070	-	0.003	0.004	0.011	0.036	0.004	0.072	-
26	0.054	-	-	0.034	-	-	0.032	0.063	0.062	0.047
27	0.004	0.050	0.019	-	0.031	0.027	0.052	0.022	0.004	0.078
28	0.039	-	0.064	0.003	0.105	0.068	-	0.004	0.004	0.008
29	0.025	0.036	0.008	0.046	0.004	0.118	0.003	0.004	0.094	0.074

Tablo 30 – MLL veri seti 10 sınıflandırıcı için kromozom ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	-	-	-	-	-	-	-	0.172	0.151	-
2	-	0.036	-	0.010	0.184	-	0.040	-	-	-
3	-	0.024	-	-	0.014	0.167	0.010	-	-	0.008
4	0.175	-	0.101	-	0.199	-	-	0.090	0.019	0.226
5	-	-	-	-	0.177	-	-	0.041	-	-
6	0.194	0.060	0.209	-	0.071	-	-	0.197	0.274	-
7	-	-	0.054	-	-	-	-	-	-	0.143
8	0.081	-	0.014	0.030	-	0.153	0.051	-	-	-
9	-	-	-	0.290	-	-	0.040	-	-	-

10	-	0.060	-	-	-	-	0.131	-	0.047	0.060
11	-	-	0.081	-	-	0.167	-	0.189	0.019	-
12	0.069	-	0.088	0.100	-	0.042	-	-	-	-
13	0.038	-	-	0.050	0.135	-	-	0.049	-	-
14	0.025	0.167	-	0.020	0.021	-	-	-	-	-
15	-	-	-	-	-	-	-	-	-	-
16	-	-	0.182	-	-	-	-	-	0.094	-
17	-	0.131	0.081	-	-	-	0.101	-	-	-
18	0.038	-	-	-	0.184	-	0.212	-	-	-
19	0.038	-	0.182	0.140	-	0.083	-	-	0.094	0.053
20	-	-	-	-	-	-	0.101	0.074	0.075	0.135
21	0.162	0.012	-	0.020	-	0.083	-	-	0.009	-
22	-	-	-	-	-	0.194	-	-	-	0.113
23	0.181	0.095	-	-	-	0.014	-	-	-	0.030
24	-	-	-	0.250	-	-	-	-	-	-
25	-	-	-	-	0.007	-	-	-	-	0.098
26	-	-	-	-	0.007	0.069	-	-	-	-
27	-	-	0.007	0.090	-	0.028	-	0.131	-	-
28	-	0.190	-	-	-	-	0.020	0.025	0.217	-
29	-	0.226	-	-	-	-	0.293	0.033	-	0.135

Tablo 31 – Yumurtalık kanseri veri seti en iyi kombinasyon ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	-	-	-	0.259	-	-	0.031	0.019	0.166	-
2	-	-	0.033	-	0.179	0.044	-	-	-	0.046
3	-	0.041	0.165	0.125	-	-	-	-	-	0.176
4	0.233	-	0.149	-	-	-	0.031	-	0.020	0.084
5	-	0.102	-	-	-	0.108	-	-	0.046	0.176
6	0.039	0.020	-	0.036	-	-	-	-	-	0.183
7	-	-	0.058	-	0.096	0.127	0.044	-	0.139	-
8	0.054	-	0.083	-	0.013	0.006	-	-	-	-
9	0.124	0.204	-	0.009	-	-	-	0.159	0.007	0.031
10	0.039	0.092	0.157	0.062	0.122	-	0.050	0.178	-	0.023
11	0.140	-	-	-	0.038	0.032	0.189	-	0.199	-
12	-	0.153	-	0.125	-	0.013	-	0.172	-	-
13	-	-	-	0.062	-	-	0.019	-	0.007	-
14	-	-	-	0.009	-	-	-	-	-	-
15	-	0.143	-	0.125	-	0.133	0.126	0.166	-	0.076
16	0.008	0.092	-	-	-	-	-	-	-	-
17	0.186	-	0.058	0.098	0.032	-	-	-	-	0.015
18	-	-	0.074	-	0.006	-	0.189	0.096	-	-
19	-	-	0.165	-	-	0.019	0.019	0.019	0.060	-
20	-	-	-	-	0.090	0.190	-	-	-	-
21	0.023	-	0.008	0.080	-	0.133	0.113	0.076	0.093	-
22	-	-	-	-	0.173	-	-	-	-	0.084
23	0.008	0.041	0.041	-	-	-	-	-	-	0.053
24	0.093	-	-	-	0.045	-	-	0.038	0.139	-
25	-	0.031	-	-	0.122	0.171	-	0.006	-	0.053

26	-	0.020	0.008	-	-	0.025	0.006	-	-	-
27	-	0.061	-	-	0.083	-	-	-	0.053	-
28	0.054	-	-	-	-	-	0.182	0.064	-	-
29	-	-	-	0.009	-	-	-	0.006	0.073	-

Tablo 32 – Lenfoma veri seti 15 sınıflandırıcı için ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	0.058	-	0.005	0.032	-	0.148	-	0.045	-	0.128
2	0.044	0.054	-	0.019	0.041	0.037	0.008	0.005	-	-
3	0.073	-	0.086	-	0.057	0.048	0.047	-	0.056	0.007
4	0.058	0.094	0.057	-	0.033	0.032	0.054	-	-	0.020
5	0.095	0.069	0.100	0.006	0.008	-	-	-	-	0.020
6	-	-	-	-	-	-	0.085	-	-	-
7	0.051	-	-	-	-	0.026	0.070	0.086	0.084	0.128
8	0.153	0.109	-	-	0.008	-	-	0.032	0.017	0.047
9	-	0.040	-	0.006	-	-	0.070	-	-	-
10	-	-	-	-	0.057	0.095	-	-	0.124	-
11	-	0.094	0.124	-	-	0.011	0.054	0.068	-	0.081
12	-	-	-	0.181	-	0.079	0.085	0.108	0.112	0.135
13	-	0.104	0.090	0.045	0.065	-	-	0.072	-	0.088
14	0.022	0.094	-	0.116	0.081	0.079	-	0.041	0.028	-
15	-	-	0.071	0.194	0.138	-	-	0.050	0.011	-
16	-	-	0.005	0.097	0.179	0.048	-	-	-	0.020
17	0.051	-	0.105	0.026	-	-	0.124	0.045	-	0.041
18	0.073	-	-	-	-	0.016	-	-	0.022	0.014
19	0.036	-	0.048	0.103	-	0.153	0.016	0.135	0.045	-
20	-	-	-	-	0.098	-	-	-	0.107	-
21	-	0.005	-	-	-	-	-	-	-	0.115
22	0.139	0.005	0.014	-	-	-	0.124	0.059	0.084	-
23	-	-	0.071	0.077	0.008	-	0.023	-	-	-
24	0.124	0.094	0.062	0.032	0.016	-	0.132	-	-	-
25	0.007	-	-	0.045	0.122	-	-	0.140	0.135	-
26	-	0.045	0.133	-	0.089	0.095	-	0.014	-	-
27	-	0.020	-	0.019	-	0.058	0.093	-	0.017	0.074
28	0.015	0.025	-	-	-	-	0.016	-	0.062	-
29	-	0.149	0.029	-	-	0.074	-	0.104	0.096	0.081

Tablo 33 – Mavi hücre tümör seti en iyi kombinasyon ağırlık seti

	1	2	3	4	5	6	7	8	9	10
1	-	-	0.006	-	0.067	0.147	0.016	0.021	-	0.160
2	-	0.040	-	0.086	0.086	-	-	0.005	-	0.047
3	-	0.166	-	-	0.049	-	0.068	0.132	0.018	-
4	-	0.040	0.012	0.086	-	-	0.037	-	-	0.053
5	-	0.080	0.078	0.005	-	0.052	-	-	0.078	-
6	0.167	0.114	0.090	-	0.074	0.016	0.089	-	-	-
7	-	0.126	0.150	0.048	0.049	-	0.079	-	0.102	0.100
8	0.048	0.040	0.024	0.124	-	0.073	0.084	0.068	0.006	-
9	0.008	0.097	0.036	0.145	0.129	-	-	0.037	0.078	-

10	0.119	-	0.138	0.097	0.012	0.068	-	0.053	-	0.080
11	-	0.063	0.156	-	0.025	0.021	0.031	0.005	-	-
12	-	0.006	-	-	0.092	0.073	0.010	0.116	0.066	-
13	0.008	0.080	0.060	0.027	0.043	0.016	0.136	0.047	0.108	0.033
14	0.222	-	-	-	-	0.021	-	-	0.042	0.040
15	0.016	-	-	0.016	0.043	0.037	-	0.053	0.018	0.060
16	0.063	0.057	0.012	-	0.055	-	0.068	0.121	0.030	-
17	0.008	-	-	0.048	-	0.162	-	-	-	-
18	0.008	0.011	0.102	0.016	0.080	-	-	-	-	0.047
19	0.032	-	-	-	0.018	-	0.084	0.053	0.024	-
20	0.063	-	-	0.005	0.117	-	-	-	0.078	0.073
21	0.008	-	0.018	0.059	-	-	0.016	0.105	-	0.060
22	-	0.040	0.006	-	0.006	-	0.047	-	0.048	0.013
23	0.008	0.017	0.072	-	-	0.037	-	0.100	0.169	0.087
24	-	-	-	-	0.055	0.047	0.120	-	0.006	0.040
25	0.111	0.006	-	-	-	0.084	-	0.074	0.114	-
26	0.087	-	-	0.075	-	0.058	0.016	0.005	0.012	-
27	-	-	0.036	0.054	-	0.016	-	-	-	0.007
28	0.024	0.017	-	0.097	-	0.073	0.073	0.005	-	0.067
29	-	-	0.006	0.011	-	-	0.026	-	-	0.033

Değerlendirme

Uygulanan yöntemlerin hata oranlarına baktığımızda, önerilen yöntemin özellikle son yıllarda kullanılan diğer yöntemler kadar iyi ve bireysel sınıflandırıcılara göre ise yüksek performans gösterdiği gözlemlenmiştir. Ayrıca deney için belirlenen 29 sınıflandırıcı havuz istenildiği gibi değiştirilebilir ve arttırılabilir. Bu sayede istenildiği kadar çeşit elde edilebilir. Önerilen yöntem ile aynı zamanda düşük performans gösteren sınıflandırıcı seçme şansı da azaltılmıştır. Performans anlamında veri setine göre doğruluklar değişebilmektedir. Bazı veri setlerinde (Lenfoma) sonuçlar aynı çıkabilirken, bazı veri setlerinde çok az da olsa gelişme kaydedilmiştir. Mavi hücre tümör veri seti gibi. Ancak, algoritmanın çalışma zamanı, tek bir sınıflandırıcının çalışma zamanından daha uzun olduğu gözlemlenmiştir. Fakat sınıflandırıcı topluluğu içindeki sınıflandırıcıların ağırlıklandırılması ile oluşturulan model ile tek başına en iyi sonuçların elde edildiği sınıflandırıcı yönteminden daha iyi sonuçlar elde edilmiştir. Oluşturulan modelde bazı sınıflandırıcılar yer almamaktadır. Yer alanların ağırlıkları da farklı veri kümeleri için değişkenlik göstermektedir.

6 SONUÇ

Sınıflandırma yöntemi günümüzde birçok alanda kullanılmaktadır. Kullanılan sınıflandırıcılar ise farklı amaçlar için üretilmiş metotlardır. Bu metotlar tüm veriler için aynı anda çok iyi performansı verememektedirler. Bu yüzden bir veri seti için ideal sınıflandırıcıyı bulmak zordur. Çoğu zaman veri seti için yanlış sınıflandırıcı seçme olasılığı yüksektir. Bunun için ise sınıflandırıcı topluluğu kavramı geliştirilmiştir. Bu kavramda havuzda bulunan sınıflandırıcıların bireysel tahminleri birleştirilerek, yeni bir sınıflandırıcı model oluşturularak sınıflandırma doğruluğunun arttırılacağı düşünülmüştür. Bazı veri setlerinde havuzdaki sınıflandırıcılardan bir kısmının performansı, diğer sınıflandırıcılardan daha iyi olabilir. Performansın daha iyi olduğu sınıflandırıcıların tahminlerini ön plana çıkarmak için sınıflandırıcılara ağırlıklar atanmıştır. Bu sayede genel doğrulukta artış yapılması hedeflenmektedir.

Bu tez çalışmasında, sınıflandırıcı topluluk oluşturma ve topluluktaki sınıflandırıcılara ağırlık atamak için 2 seviyeli bir genetik algoritma geliştirilmiştir. Bu algoritmanın ilk seviyesinde sınıflandırıcılar seçilirken, ikinci seviyesinde seçilmiş olan sınıflandırıcıların ideal ağırlıkları bulunmaya çalışılmıştır. Önerilen yöntem, bireysel sınıflandırıcı sonuçları ile karşılaştırıldığında, önerilen yöntemin daha iyi sonuç verdiği gözlemlenmiştir. Sınıflandırıcı model üretmek için harcanan zaman, tek bir sınıflandırıcının harcadığı zamandan fazla olmasına rağmen, bilinmeyen örnekleri sınıflandırırken zaman farkı gözlemlenmemiştir.

Yapılan bu çalışmanın, ileride çalışma zamanı açısından geliştirilmesi hedeflenmektedir. Çalışma zamanının arttırılması için, sınıflandırıcıların paralel olarak çalıştırılması ve de CPU üzerinde çekirdek bölümünün arttırılması gerekmektedir. Bu sayede model üretim süresinin azaltılması hedeflenmektedir. Ayrıca bellek kullanımının azaltılması ve algoritmanın yerel maksimumda kalmamasını sağlayacak yöntemler üzerinde çalışılması planlanmaktadır. Son olarak önerilen algoritmanın çalışmasından önce, veri setleri üzerinde bir ön işlemden geçirilerek, özellik sayısının azaltılması için çalışma planlanmaktadır.

KAYNAKLAR

[1] HOLLAND, John H. "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence." *U Michigan Press*, 1975.

[2] XU, Lei; KRZYZAK, Adam; SUEN, Ching Y. "Methods of combining multiple classifiers and their applications to handwriting recognition." *Systems, Man and Cybernetics, IEEE Transactions on*, 1992, 22.3: 418-435.

[3] Polikar, R., "Ensemble learning," *Scholarpedia*, vol. 4, no. 1, pp. 2776, 2009.

[4] Nagi, Sajid, and Dhruba Kr Bhattacharyya. "Classification of microarray cancer data using ensemble approach." *Network Modeling Analysis in Health Informatics and Bioinformatics*: 1-15.

[5] Zhao, Tian-Zhong, et al. "Classifier ensemble based-on bi-coded chromosome genetic algorithm for automatic image annotation." *Machine Learning and Cybernetics, 2008 International Conference on*. Vol. 1. IEEE, 2008.

[6] Fida, B., et al. "Heart disease classification ensemble optimization using Genetic algorithm." *Multitopic Conference (INMIC), 2011 IEEE 14th International*. IEEE, 2011.

[7] Salem, D.A. ; AbulSeoud, R.A.A.A. ; Ali, H.A. "Merging Genetic Algorithm with Different Classifiers for Cancer Classification using Microarrays" *Radio Science Conference* (NRSC), (2012): 659-666

[8] Holmes, Geoffrey, Andrew Donkin, and Ian H. Witten. "Weka: A machine learning workbench." *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. IEEE, 1994.

[9] Alizadeh, Ash A., et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." *Nature* 403.6769 (2000): 503-511.

[10] Gordon, Gavin J., et al. "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma." *Cancer research* 62.17 (2002): 4963-4967.

[11] Armstrong, Scott A., et al. "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia." *Nature genetics* 30.1 (2001): 41-47.

[12] van't Veer, Laura J., et al. "Gene expression profiling predicts clinical outcome of breast cancer." *nature* 415.6871 (2002): 530-536.

[13] Li, Jingyan, and Huiqing Liu. "Kent ridge bio-medical data set repository."*Institute for Infocomm Research. http://sdmc. lit. org. sg/GEDatasets/Datasets. html* (2002).

[14] Petricoin III, Emanuel F., et al. "Use of proteomic patterns in serum to identify ovarian cancer." *The lancet* 359.9306 (2002): 572-577.

[15] Khan, Javed, et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nature medicine* 7.6 (2001): 673-679.

[16] Moreau, Y., De Smet, F., Thijs, G., Marchal, K., & De Moor, B. (2002). "Functional bioinformatics of microarray data: from expression to regulation." *Proceedings of the IEEE*, 90(11), 1722-1743.

[17] Tan, Aik Choon, and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification." (2003).

[18] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286.5439 (1999): 531-537.

[19] Schadt, Eric E., et al. "Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data." *Journal of Cellular Biochemistry* 84.S37 (2001): 120-125.

[20] Schena, Mark, et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270.5235 (1995): 467-470.

[21] Lockhart, David J., et al. "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nature biotechnology* 14.13 (1996): 1675-1680.

[22] Rennie, Jason D., et al. "Tackling the poor assumptions of naive bayes text classifiers." *ICML*. Vol. 3. 2003.

[23] Su, Jiang, et al. "Discriminative parameter learning for Bayesian networks."*Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

[24] Cohen, William W. "Fast effective rule induction." *ICML*. Vol. 95. 1995.

[25] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1998.

[26] John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995.

[27] Sumner, Marc, Eibe Frank, and Mark Hall. "Speeding up logistic model tree induction." *Knowledge Discovery in Databases: PKDD 2005*. Springer Berlin Heidelberg, 2005. 675-683.

[28] Keerthi, S. Sathiya, et al. "Improvements to Platt's SMO algorithm for SVM classifier design." *Neural Computation* 13.3 (2001): 637-649.

[29] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." *Machine learning* 6.1 (1991): 37-66.

[30] Frank, Eibe, Mark Hall, and Bernhard Pfahringer. "Locally weighted naive bayes." *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2002.

[31] Kohavi, Ron. "The power of decision tables." *Machine Learning: ECML-95*. Springer Berlin Heidelberg, 1995. 174-189.

[32] Holte, Robert C. "Very simple classification rules perform well on most commonly used datasets." *Machine learning* 11.1 (1993): 63-90.

[33] Frank, Eibe, and Ian H. Witten. "Generating accurate rule sets without global optimization." (1998).

[34] Gaines, Brian R., and Paul Compton. "Induction of ripple-down rules applied to modeling large databases." *Journal of Intelligent Information Systems* 5.3 (1995): 211-228.

[35] Shi, Haijian. “Best-first decision tree learning.” *Diss. The University of Waikato*, 2007.

[36] Gama, João. "Functional trees." *Machine Learning* 55.3 (2004): 219-250.

[37] Quinlan, John Ross. “C4. 5: programs for machine learning.” *Vol. 1. Morgan kaufmann*, 1993.

[38] Holmes, Geoffrey, et al. "Multiclass alternating decision trees." *Machine Learning: ECML 2002*. Springer Berlin Heidelberg, 2002. 161-172.

[39] Landwehr, Niels, Mark Hall, and Eibe Frank. "Logistic model trees." *Machine Learning* 59.1-2 (2005): 161-205.

[40] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[41] El-Manzalawy, Yasser, and Vasant Honavar. "WLSVM: Integrating libsvm into WEKA environment." *Software available at http://www. cs. iastate. edu/yasser/wlsvm* (2005).

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, Adı	:	İYİDOĞAN, Taylan
Uyruğu	:	T.C.
Doğum tarihi ve yeri	:	27.01.1989 Ankara
Medeni hali	:	Bekâr
Telefon	:	0 (555) 563 76 40
E-posta	:	tiyidogan@etu.edu.tr

Eğitim

Derece	**Eğitim Yeri**	**Mezuniyet Tarihi**
Lisans	TOBB ETÜ Bilgisayar Mühendisliği	2011

İş Deneyimi		**Görev**
2011-	Aydın Yazılım	Yazılım Mühendisi
2010-2011	Aydın Yazılım	Stajyer
2009-2010	Ventura Yazılım	Stajyer
2008-2009	Portakal Teknoloji	Stajyer

Yabancı Dil

İngilizce

Yayınlar

ASSIST: An Integrated Measurement Tool (IWSM-Mensura 2013)