GU J Sci, Part C, 9(4): 811-827 (2021)

Gazi University

Journal of Science

PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

High Performance Classification of Cancer Types with Gene Microarray Datasets: Hybrid Approach

Yılmaz ATAY[1,*] Muhterem Oğuzhan YILDIRIM[1] Cuma Umur DOĞAN[1]

[1]*Gazi Üniversitesi Mühendislik Fakültesi Eti Mh. Yükseliş Sk. No: 5, 06570 Maltepe/ANKARA*

Article Info:

Research article
Received: 15.12.2021
Revision: 14.12.2021
Accepted: 26.09.2021

Highlights

- *JSON.*
- *MongoDB.*
- *Location Services.*

Keywords

Ensemble Method
Genetic Algorithm
Cancer
Microarray
Naive Bayes
Classification

Graphical/Tabular Abstract

In this study, the classification of human cancer diseases is discussed by using different gene microarray datasets. In addition, a new methodology is presented for the efficient classification of cancer diseases over gene microarray datasets.

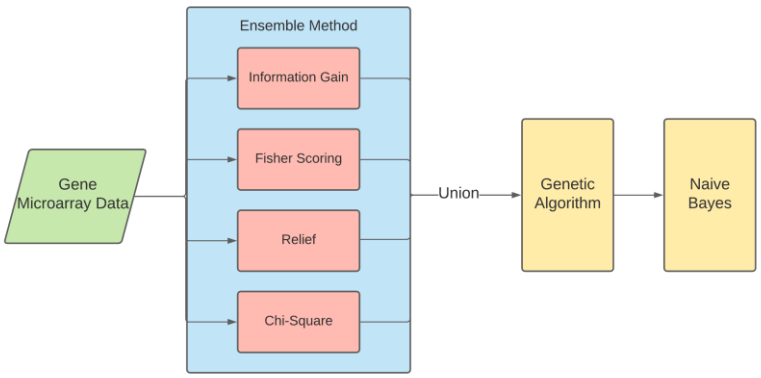


Figure A. Flowchart of proposed hybrid approach

Purpose: The main purpose of this study is to develop a machine learning model that can predict whether samples from gene microarray data are cancerous or not. The problem of classification of gene microarray data as a field of study has many difficulties due to its unique features, another aim of this study is to minimize these difficulties with effective methods, thus increasing the classification success of the proposed model.

Theory and Methods: In the proposed hybrid method, different filtering, wrapping and classification methods should be used together. For this reason, the materials and methods used in the development of the proposed system are explained in different headings. Fisher scoring, Chi-square, relief, information gain, customized genetic algorithm and Naive Bayes methods explained in detail according to the proposed hybrid method.

Results: The proposed model can be defined in three different parts. By applying the ensemble feature selection method instead of a single feature selection method in the first part, the probability of not selecting the features that can be decisive during classification is reduced. In the second part, the most successful feature combination is selected by performing a stochastic search with a genetic algorithm on the sub-dataset that result of the ensemble feature selection method. And in the third part, the classifier is trained with the sub dataset, which is the result of the previous part. The proposed model in this study was trained separately with Leukemia, Central Nervous System, and Colon Tumor datasets, and accuracy values were obtained as 97.06%, 85.48%, and 86.67% respectively.

Conclusion: In all Leukemia, Central Nervous System and Colon Tumor datasets considered in the study, the accuracy rate of the proposed hybrid method was observed to be better than the accuracy rates of other studies. The hybrid approach proposed within the scope of the study is both designed to directly affect the classification performance and presented in a structure with developable flexibility. Based on the test results obtained, it can be said that the proposed hybrid approach generally gives successful results in the classification problems of cancer diseases and has a high potential for maximizing the performance in datasets with different characteristics

*Corresponding author, e-mail: yilmazatay@gazi.edu.tr
DOI: 10.29109/gujsc.1000926

GU J Sci, Part C, 9(4): 811-827 (2021)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

http://dergipark.gov.tr/gujsc

High Performance Classification of Cancer Types with Gene Microarray Datasets: Hybrid Approach

Yılmaz ATAY[1,*] Muhterem Oğuzhan YILDIRIM[1] Cuma Umur DOĞAN[1]

[1]*Gazi Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, 06500, Yenimahalle/ANKARA*

Makale Bilgisi

Araştırma makalesi
Başvuru: 15.12.2021
Düzeltme: 14.12.2021
Kabul: 26.09.2021

Keywords

Ensemble Method
Genetic Algorithm
Cancer
Microarray
Naive Bayes
Classification

Anahtar Kelimeler

Ensemble Metot
Genetik Algoritma
Kanser
Mikrodizi
Naive Bayes
Sınıflandırma

Abstract

Currently the approach of biological meaningfulness detection from gene microarray datasets obtained with microarray technology is used effectively in many areas such as disease diagnosis and differentiation of cancer types. However, since datasets obtained with this technology measure gene expression profiles collectively, the number of features in the dataset can be quite high. The small number of samples in gene microarray datasets, the high number of features and where the data is noisy significantly complicates the preparation process of these datasets. In order for machine learning models to successfully classify, the number of features that represent the size of the dataset should be reduced. In the proposed method, gene microarray data is taken as input and Information Gain, Fisher Correlation Scoring, ReliefF and, Chi-Square methods are applied separately for feature selection. After this stage, a sub-dataset containing the new genes is obtained and a pool of genes for Genetic Algorithm is created according to this dataset. Bayes classifier is trained using the sub-dataset created with the genes of the most successful chromosome. Thus, the classification process of cancer data is successfully completed. The model proposed in this study was applied to datasets that are frequently used in the literature and high success rates were obtained in classification. As a result; acceptable feature selection methods and the hybrid method based on Genetic Algorithm generally provided the most appropriate results on all test data.

Gen Mikrodizi Veri Setleriyle Kanser Türlerinin Yüksek Başarımlı Sınıflandırılması: Hibrit Yaklaşım

Öz

Günümüzde mikrodizi teknolojisi ile elde edilen gen mikrodizi veri setlerinden biyolojik anlamlılık tespiti yaklaşımı, hastalık tanısı ve kanser türlerinin ayırt edilmesi gibi pek çok alanda etkin bir şekilde kullanılmaktadır. Fakat bu teknoloji ile elde edilen veri kümeleri, gen ifade profillerini toplu olarak ölçtüğü için veri kümesindeki özellik sayısı oldukça fazla olabilmektedir. Gen mikrodizi veri kümelerindeki örnek sayılarının az olması, özellik sayısının fazla olması ve verilerin gürültülü olması bu veri kümelerinin ön hazırlık işlemlerini oldukça karmaşık hale getirmektedir. Makine öğrenmesi modellerinin sınıflandırmayı başarıyla yapabilmesi için özellik sayısının, yani veri kümesinin boyutunun azaltılması gerekmektedir. Önerilen yöntemde, gen mikrodizi verileri girdi olarak alınır ve öznitelik seçimi amacıyla Bilgi Kazancı, Fisher Korelasyon Skorlama, ReliefF ve Ki-Kare yöntemleri ayrı ayrı uygulanır. Bu aşamadan sonra yeni gen alt veri kümesi elde edilir ve Genetik Algoritmanın gen havuzu oluşturulur. Bu algoritmanın uygun adımlarda tekrar çalıştırılması sonrasında seçilen en başarılı kromozomun genleri ile oluşturulan alt veri kümesi kullanılarak Naive Bayes sınıflandırıcısı eğitilir. Böylece kanser verilerinin sınıflandırılması işlemi tamamlanır. Bu çalışmada önerilen model, literatürde sıklıkla kullanılan veri kümelerine uygulanmış ve sınıflandırmada yüksek başarı oranları elde edilmiştir. Sonuç olarak; uygun öznitelik seçim yöntemleri ve Genetik Algoritma temelli hibrit yöntem genel anlamda tüm test verileri üzerinde en uygun sonuçlara ulaşılmasını sağlamıştır.

1. GİRİŞ (INTRODUCTION)

İsimlerini genellikle ortaya çıktıkları doku ve organlardan alan ve sayısı yüzü geçen kanser çeşidi bulunmaktadır. Kanserler doku hücrelerinin tipine göre Karsinom, Sarkom, Miyelom, Lösemi, Lenfoma ve karışık tipler olmak üzere altı ana kategoride incelenebilir [1]. Kanserli hücrelerin tıbbi olarak teşhisi ve tümör tiplerinin belirlenmesi büyük öneme sahiptir. Tümörlerin sınıflandırılması süreci, hastaların daha iyi tedavi görebilmesine olanak sağlarken; maruz kalacakları toksin ve yan etkilerin en aza indirgenmesine olanak sağlayabilir. Tümörlerin geleneksel yöntemler kullanılarak teşhis edilmesi ve sınıflandırması oldukça zor ve maliyetli bir süreçtir. Ayrıca bu işlemler insan hatalarına ve gözlemciler arası değişkenliğe karşı da oldukça hassastır. Bu sebeple tanı sürecinde yeni yöntemlerin geliştirilmesi hayati öneme sahiptir [2]. Kanser tanısında gen ekspresyon profili, hücresel özelliklerin tanımlanmasında etkin olabilir. Burada gen ekspresyon profili hücrenin fenotipini, işlevini ve uyaranlara karşı tepkisini belirler. Kanserli hücrelerin gen ekspresyon profilleri, normal hücrelerin gen ekspresyon profilleri ile karşılaştırıldığında edinilen bilgi ile kanser tanı süreci iyileştirilebilir. Diferansiyel gösterim, gen ekspresyonu seri analizi ve mikrodizi yöntemleri gibi gen ekspresyonu profilleme yöntemleri, kanser araştırmalarında başarıyla uygulanmaktadır. Kanserli hücrelerin gen ekspresyon verilerini çıkarmada en çok kullanılan teknoloji mikrodizi teknolojisidir. Bunun sebebi kullanımlarının kolay olması, büyük ölçekli DNA dizilemesi gerektirmemesi ve çoklu örneklerden sayısız genin paralel olarak sayısallaştırılmasına olanak sağlamasıdır [3]. Mikrodizi teknolojisi ile elde edilen DNA mikrodizi veri kümeleri biyoinformatik ve makine öğrenmesinde kullanılmaktadır. Bu veri kümeleri hastalık tanısı, kanser tiplerinin ayırt edilmesi gibi pek çok faydalı alanda kullanılmaktadır. Fakat bu teknoloji ile elde edilen veri kümeleri, gen ifadesini toplu olarak ölçtüğü için veri kümesindeki özellik sayısı oldukça fazladır. DNA mikrodizi veri kümelerindeki örnek sayılarının az olması, özellik sayısının fazla olması [4] ve verilerin gürültülü olması bu veri kümelerinin ön hazırlık işlemlerini oldukça zor bir hale getirmektedir. Makine öğrenmesi modellerinin sınıflandırmayı başarıyla tamamlayabilmesi için öznitelik sayısının indirgenmesi ve sonuç olarak veri kümesinin boyutunun azaltılması önemlidir.

Literatürdeki çalışmalar incelendiğinde; Yu ve diğerlerinin önerdiği çalışmada, temelde gen mikrodizi verilerinin dengesiz sınıflara sahip olması problemine odaklanılmıştır [5]. İncelenen çalışmanın kısıtları temelde uygulanan yöntemin oldukça zaman alıcı olması ve önerilen yöntemin sadece iki sınıflı veri kümeleri üzerinde çalışabilmesi ile ilgilidir. Önerilen modelde öznitelik seçme yöntemi olarak Weighted Metric, sınıflandırma yöntemi olarak ise Decision Rule önerilmiştir. Bu çalışmada önerilen model Leukemia ve Colon Tumor veri kümeleri ile eğitilmiş ve sırasıyla %95.55 ve %85.49 doğruluk oranları elde edilmiştir. Gunavathi ve Premalatha tarafından gerçekleştirilen çalışmada [6], önerilen modelde öznitelik seçme yöntemleri olarak T-istatistik, Signal-to Noise Ratio ve F-test yöntemleri; sınıflandırma yöntemleri olarak ise k-en yakın komşu ve destek vektör makinesi yöntemleri önerilmiştir. Çalışmada önerilen model Central Nervous System veri kümesi ile eğitilmiş ve Doğruluk Oranı olarak %81.25 değeri elde edilmiştir. Bu çalışmadaki en önemli kısıtlar diğer çalışmalarda da olduğu gibi üzerinde çalışılan gen mikrodizi verilerinin çok yüksek sayıda özniteliğe sahip olması ve az sayıda örnek içermesidir. [7] referas numaralı çalışmada önerilen yaklaşım, öznitelikleri seçmek için genetik algoritmayı kullanırken; sınıflandırma süreci için de destek vektör makinesini tercih etmiştir. İlgili çalışmada önerilen model ayrı ayrı Leukemia ve Colon Tumor veri kümeleri ile eğitilmiş ve sırasıyla %91.5 ve %84.6 başarı değerleri elde edilmiştir. Bu çalışmada da yüksek öznitelik sayısı ve düşük örnek sayısı temel zorluğu oluşturmaktadır ayrıcı bu çalışmanın zaman karmaşıklığının yüksek olması ve çok sınıflı veriler üzerinde uygulanabilir olmaması temel kısıtlarını oluşturnaktadır. Salem ve diğerlerini bilimsel çalışmasındaki [8] yöntemde, öznitelik seçme yöntemi olarak bilgi kazancı yaklaşımı tercih edilirken; sınıflandırma yöntemi olarak Small for Gestational Age yaklaşımı kullanılmıştır. Nguyen ve diğerlerinin sunduğu çalışmada [9], Leukemia veri kümesi üzerinde Analytic Hierarchy Process ile öznitelik seçimi yapılmış olup; Hidden Markov Models ile sınıflandırma yapılmıştır. Çalışmada önerilen modelin Leukemia veri seti üzerindeki doğruluk oranı %96.48 olarak belirlenmiştir. Hengpraprohm'un çalışmasında [10] Leukemia veri kümesi üzerinde Signal-to Noise Ratio yöntemi ile öznitelik seçimi yapılmış ve Genetik Algoritma uygulanmıştır. Çalışmada önerilen yöntemin Leukemia veri kümesi üzerindeki Doğruluk Oranı %91.9'dur. Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression [11] isimli çalışmada gen mikrodizi verilerinin yüksek boyutlu olduğu ve özellikle kanserli hastaların teşhisi için olan gen mikrodizi veri kümelerinin az örneğe sahip olduğundan bir diğer deyişle Curse of Dimentionality

probleminden bahsedilmiştir. Çalışma prostat kanserinin gen mikrodizi verilerinden teşhisi üzerinedir. Teşhis işlemi iki adımda tanımlanmıştır. Birinci adımda Korelasyon Öznitelik (Correlation Feature Selection) Seçimi yöntemi ile öznitelik seçimi yapılırken, ikinci adımda Random Committee Ensemble sınıflandırıcısı kullanılmıştır. Önerilen modelin teşhis doğruluk oranının tespiti için ise 10 katlı çapraz doğrulama yöntemi kullanılmış, %95.098 doğruluk oranı elde edilmiştir ve bu oranın aynı veri kümesi üzerinden teşhis yapmayı amaçlayan diğer çalışmaların sonuçlarından daha yüksek olduğu belirtilmiştir. Genetik Algoritma ve Sınıflandırıcı Yöntemler ile Kanser Tahmini [12] isimli çalışmada akciğer ve beyin kanseri ile ilgili gen mikrodizi verileri kullanılmış ve genetik algoritma ile öznitelik seçimi yapılmıştır. Naive Bayes, Bayes Net, k-En Yakın Komşu, Rastgele Orman ve Destek Vektör Makineleri sınıflandırma yöntemleri eğitilerek başarı oranları karşılaştırılmıştır ve sonucunda genetik algoritma ile öznitelik seçimi işleminin makine öğrenmesi ile yapılacak olan kanser teşhisi çalışmalarında başarım arttırıcı özelliği olduğu gösterilmiştir. Bir diğer çalışma olan Çok Amaçlı Genetik Algoritma Kullanarak DNA Mikrodizi Verilerinin Kümelenmesi [13] çalışmasında ise DNA mikrodizi verilerini örnek tabanlı kümelemek için küme sayısı önceden belirlenmeden çok amaçlı genetik algoritmalara göre yeni bir yöntem geliştirmektir. Önerilen yöntem daha önce geliştirilmiş olan hızlı genetik k-means algoritmasını çok amaçlı genetik algoritma süreci ile birleştirmiştir. Bu sayede daha etkin ve doğruluk oranı daha yüksek bir sınıflandırma yöntemi ortaya çıkmıştır. Diğer bir bilimsel çalışma ise Makine Öğrenmesi Yöntemleri Kullanarak Kanser Teşhisi [14] isimli araştırma çalışmasıdır. Göğüs kanseri veri seti kullanılarak kanser verilerinin sınıflandırması üzerine nitelik indirgeme metotlarının etkisinin incelenmesini amaçlayan bu çalışmanın sonucunda öznitelik eleme yöntemlerinin eğitim başarısı üzerindeki olumlu etkisi analiz edilmiş ve sonuçları gösterilmiştir. A Cancer Gene Selection Algorithm Based on the K-S Test and CFS isimli çalışmada [15] K–S (Kolmogorov–Smirnov) Test, Wilcoxon Test ve T-Test isimli yöntemler öznitelik seçme yöntemleri olarak tercih edilmiş ve sonuçlar karşılaştırılmıştır. Burada en iyi sonuçların K-S Test ile elde edildiği görülmüştür.

Bu çalışmada gen mikrodizi verileri kullanılarak insanlara ait kanser hastalıklarının sınıflandırılma konusu ele alınmış ve gen mikrodizi veri kümeleri üzerinden kanser hastalıklarının etkin bir şekilde sınıflandırılması için yeni bir metodoloji sunulmuştur. Bu metodolojiye göre öncelikle bilgi kazancı, fisher korelasyon skorlama, relief ve ki-kare yöntemleri kullanılarak elde edilen alt kümelerin birleşimi ensemble metot ile sağlanır. Bu işlemle öznitelik seçimi yapılır ve daha sonra genetik algoritma kullanılarak öznitelik azaltma süreci uygulanır. Son olarak, naive bayes sınıflandırıcısı kullanılarak sınıflandırma işlemi yapılır. Önerilen bu hibrit yaklaşım literatürde yaygın olarak kullanılan üç kanser veri seti üzerinde test edilmiş ve elde edilen sonuçlar literatürdeki güncel çalışmaların sonuçlarıyla karşılaştırılmıştır.

Bu çalışmanın sonraki kısımları şu şekilde düzenlenmiştir: bölüm 2'de temel problemler tanımlanmıştır, bölüm 3'te kullanılan veri kümeleri, kullanılan öznitelik indirgeme ve sınıflandırma yöntemleri ve önerilen model detaylı bir şekilde açıklanmıştır, bölüm 4'te önerilen modelin uygulanması sonucunda elde edilen bulgular açıklanmış ve tartışılmıştır, son olarak bölüm 5'te ise sonuçlar sunulmuştur

.2. PROBLEM TANIMI (PROBLEM DEFINATION)

Bu çalışmanın temel amacı, gen mikrodizi verilerinden eldeki örneklerin kanserli olup olmadığını tahmin edebilecek bir makine öğrenmesi modeli geliştirmektir. Çalışma alanı olarak gen mikrodizi verilerinin sınıflandırılması problemi, kendine has özelliklerinden kaynaklı birçok zorluğa sahiptir. Bu zorluklardan ilki gen mikrodizi verilerinin yapısından kaynaklı bir problemdir. Bu problem, veri kümelerinin genellikle binlerce gene yani özniteliğe sahip olmasından ve bunun aksine örnek sayısının birkaç yüzü geçmemesinden kaynaklanmaktadır [16]. İkinci zorluk, veri setlerindeki genlerin sadece birkaçının üstünde çalışılan kanser türüyle ilişkili olmasıdır. On binlerce gen arasından en alakalı genlerin ortaya çıkartılması, en basit haliyle genlerin mümkün olan tüm alt kümelerinin tespiti ve denenmesiyle bulunabilir. Teorik olarak bu uygulanabilir görünse de uygulamada birçok problemle karşılaşılmaktadır [16]. Çünkü gen mikrodizi verilerinde gen sayılarının çok fazla olması sebebiyle tüm alt kümelerin test edilmesi kaynak, teknik ve zaman kısıtları sebebiyle mümkün değildir. Bu sebepten dolayı, genlerin tüm alt kümelerinin denenmesi ile etkin genlerin bulunması problemi genellikle NP-hard problem sınıfına dahil edilir. NP-hard problemler, doğrusal zamanda çözülmesi ve çözümün doğrulanması mümkün olmayan problem sınıfındadır [17]. Üçüncü zorluk, gen mikrodizi verilerinin hem elde edilmesindeki sürecin maliyetli

oluşundan hem de biyolojik sebeplerden dolayı elde edilen veri setlerinin gürültü olmasından kaynaklanmaktadır. Bu verilerin gürültülü olması etkin genlerin tespit edilmesinde yanılmalara sebep olabilmektedir [18]. Dördüncü zorluk ise uygulama alanından ve çalışılan problemin tipinden kaynaklanan zorluktur. Kanser sınıflandırmasında etkin olacak genlerin tespitinden sonra eğitilecek modelin başarı (accuracy) skoru, modeli değerlendirmek için yeterli değildir. Bunun sebebi, her bir sınıfa ait örnek sayısının birbirinden sayısal olarak aralarında büyük farklar olabilme ihtimalidir. Böyle bir durumda accuracy metriği ile değerlendirme yapıldığında hatalı sonuçlar elde edilebilir [19]. Dolayısıyla, etkin genlerle eğitilen modelin sadece accuracy ölçütü ile değerlendirilmesi yeterli değildir. Bu yüzden farklı değerlendirme ve doğrulama metriklerine ihtiyaç duyulmaktadır.

3. MATERYAL VE METOD (MATERIAL AND METHOD)

3.1. Öznitelik Seçimi (Feature Selection)

Gen mikrodizi verileri genellikle çok sayıda özniteliğe sahiptir. Bu durum boyutsallığın lanetine (curse of dimensionality) sebep olmaktadır. Bu tip veri kümelerinde genlerin çoğu alakasız veya gereksizdir. Gen ifade verileri kanser oluşumunu gösterebilir. Fakat alakasız gen ifade verileri makine öğrenmesi modelinin düzgün şekilde eğitilmesini engeller. Böyle durumlarla baş edebilmek için verinin yapısını bozmadan veriyi daha az değişkenle temsil edebilmek mümkündür. Bunun için özellik seçimi (*feature selection*) ve özellik çıkarma (*feature extraction*) şeklinde iki temel yaklaşım tercih edilmektedir. Bunlardan öznitelik seçimi yapılmasıyla alakasız gen verileri veri kümesinden rahatlıkla çıkarılabilir. Bu sayede işlem ve zaman karmaşıklığı azaltılmış olur.

3.1.1. Bilgi Kazancı (Information Gain)

Bu yöntem, makine öğrenmesi alanında oldukça yaygın kullanılan bir değerlendirme yaklaşımıdır. Bu yöntem aslında entropide beklenen azalmayı ölçer [20]. X özelliğine bağlı şekilde Y özelliğinin entropisindeki azalma aşağıdaki gibi hesaplanır [21].

$$\mathrm{H}(Y) = -\sum_{y \epsilon \mathrm{Y}} p(y) log_2\big(p(y)\big) \tag{1}$$

$$H(Y \setminus X) = -\sum_{x \epsilon X} p(x) \sum_{y \epsilon Y} p(y \setminus x) \log_2\big(p(y \setminus x)\big) \tag{2}$$

$$BilgiKazancı = H(Y) - H(Y \setminus X) \tag{3}$$

Bilgi Kazancı, öznitelik seçimi için kullanılabilir. Bir veri kümesindeki tüm öznitelikler sınıflandırma için eşit etkiye sahip değildir. Kimi öznitelik sınıfları birbirinden daha iyi ayırt edilebilirken; kimi öznitelik sınıfları bu tür bir işlem için elverişli değildir. Bilgi kazancı skoru ile diğer özniteliklere göre ayırt ediciliği daha yüksek öznitelikler belirlenebilir. Bu öznitelikler, sınıflara karar verme konusunda diğer özniteliklere göre daha çok katkıda bulunabilir [22].

Bilgi Kazancı yöntemi ile skorlanan özniteliklerden en yüksek skora sahip olanı, sınıflandırma için kullanıldığında en yüksek ayırt edici öznitelik olacağı kabul edilir. Karar Ağaçlarında da kullanılan bu yöntem, bu çalışmada öznitelik seçimi için kullanılmıştır çünkü gen mikrodizi verilerinin içerisinde tespit edilmek istenen kanserle ilişkili genlerin yanında alakasız genler de bulunmaktadır ve bu genler model eğitimi sırasında öznitelik olarak kullanılırsa, modelin sınıflandırma başarısı olumsuz etkilenir. Bilgi Kazancı yönteminin bu çalışma da kullanılmasının sebebi çalışılan veri kümesinin gen mikrodizi veri kümesi olmasından dolayı tespit edilmeye çalışılan kanser türüyle alakasız veya az alakalı genlerin yani özniteliklerin elenmesi, dolayısıyla sınıflandırma modelinin kanser türüyle daha alakalı genler ile eğitilmesini sağlamaktır. Bu yöntem ile öznitelik seçim işlemi, özniteliklerin Bilgi Kazancı skorlarına göre büyükten küçüğe sıralanması ve sıralanmış özniteliklerin belirlenen sayıda seçilmesi ile yapılmıştır.

3.1.2. Fisher Skorlama (Fisher Scoring)

Bu skorlama türü, özellikler için ayrı ayrı ilişki skoru hesaplar. Bu yöntem ilişki skoru hesabında her sınıf için özelliklerin standart sapmasını ve ortalamasını kullanılır. Bu yöntemin formülü Denklem 4'te gösterilmiştir [23].

$$F(x_i) = \frac{|\mu_i^+ - \mu_i^-|}{\sigma^+ - \sigma^-} \qquad (4)$$

Bu formülde sınıflar (+) ve (-) işaretleri ile gösterilmektedir. Standart sapmalar $\boldsymbol{\sigma}^+$ ve $\boldsymbol{\sigma}^-$ sembolleriyle, ortalamalar ise $\boldsymbol{\mu}^+$ ve $\boldsymbol{\mu}^-$ sembolleriyle gösterilmektedir. Bu yöntem ile elde edilen en yüksek skorlu özellik iki sınıfı birbirinden ayırabilmede uygun bir seçenek olarak tercih edilebilir. Fisher skorlamanın özellik seçimi için kullanımı sırasında öncelikle bu yöntemle skorlanan özellikler büyükten küçüğe doğru sıralanır. Daha sonra en yüksek skorlu özellikten başlanarak istenilen sayıda özellik seçilir ve işlem tamamlanır.

Bilgi Kazancı yöntemi entropi metriğini kullanarak öznitelikleri skorlarken, Fisher Skorlama standart sapma ve ortalama metriklerini kullanır. Bu çalışmada ki hibrit öznitelik seçme modeli oluşturulurken kullanılan yöntemlerden bir tanesi de Fisher Skorlama'dır çünkü bu yöntem özniteliklerin sınıflandırma ile ilişkisini skorlarken diğer yöntemlerden farklı metrikler kullanır.

3.1.3. Relief-F

Kira ve Rendell [24] tarafından önerilen bu yöntem, özellikler arasındaki bağımlılıkları ortaya çıkartarak bu özelliklerin seçilebilirliği için anlamsal değerleri bulmayı hedefler ve ikili sınıflandırma problemlerinde etkin bir şekilde kullanılır. ReliefF yaklaşımı, komşuluk algoritmalarının çalışma mantığına yakın bir işleyişe sahiptir. Ele alınan özelliğin sınıflarda bulunup bulunmadığına göre sınıflardaki yakınlıklar göz önünde bulundurularak ağırlıklandırmalar yapılır. Relief algoritması üç temel adımdan oluşmaktadır. Bu adımlardan ilkinde, ele alınan örnekle aynı ve farklı sınıflarda bulunan en yakın örneklerin ilgili özellik skorlarının ayrı ayrı belirlenmesi sağlanır. İkincisinde ele alınan özelliklerin ağırlıkları hesaplanır. Üçüncüsünde ise ağırlıklandırılan özelliklerin sıralanması ve belirlenen bir eşik değerin üstünde kalan k adet özelliğin belirlenmesi işlemleri tamamlanır. Denklem 5'te ağırlıkların güncellenmesi ile ilgili formül verilmiştir. Burada n örnek sayısını, W_i belirlenen özelliğin ağırlığını, $nearHit_i$ aynı sınıftaki en yakın örnekle ilişkili özellik değerini, $nearMiss_i$ farklı sınıftaki yakın örnekle ilgili özellik değerini, x ise rastgele seçilen örneği gösterir. Böylece algoritmanın ikinci adımındaki ağırlıklar bu formülün n kez tekrarlanmasıyla hesaplanır [25].

$$W_i = W_{i-1} - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \qquad (5)$$

Relief-F öznitelik seçme yöntemi de bu çalışmada tanımlanan hibrit öznitelik seçme modelinde kullanılan yöntemlerden bir tanesidir. Relief-F'in hibrit modele dahil edilmesinin sebebi yine öznitelikleri skorlarken diğer yöntemlerden farklı metrikler kullanmasıdır.

3.1.4. Ki-Kare (Chi-Square)

Bu yaklaşım veri bilimindeki popüler öznitelik seçme yöntemlerinden biridir. Ki-kare testinin çalışma mantığı beklenen ve gözlemlenen frekanslar arasındaki farklılığın anlamlı olup olmadığının belirlenmesine dayanmaktadır. Bu yöntemde özellikler (X) ile sınıflar (Y) arasındaki ilişkinin varlığını incelemektedir. Bu inceleme sonucunda Y ile ilişkisi bulunmayan özellikler veri kümesinden çıkarılır. Ki-kare test skoru, Denklem 6, 7 ve 8'de verilen formüller kullanılarak hesaplanmaktadır.

$$X^2 = \sum_{i=1}^{i}\sum_{j=1}^{j}\left(N_{ij} - \widehat{N_{ij}}\right)/\widehat{N_{ij}} \qquad (6)$$

$$\widehat{N}_{ij} = \frac{N_i N_j}{N} \quad (7)$$

$$d = (I-1)(J-1) \quad (8)$$

Yukarıda verilen denklemlerdeki $\boldsymbol{N_{ij}}$, Y'nin i. ve X'in j. düzeyinde bulunan birim sayısını, $\boldsymbol{\widehat{N}_{ij}}$ iki özelik birbirinden bağımsızken $\boldsymbol{Y}$'nin $\boldsymbol{i}$. ve $\boldsymbol{X}$'in $\boldsymbol{j}$. düzeydeki beklenen birim sayısını ve d test istatistiği hesaplaması yapılacak olan Ki-kare dağılımının serbestlik derecesini ifade eder. $\boldsymbol{I}$ ve $\boldsymbol{J}$ ise satırlar ve sütunlardır. Bu yöntem kullanılarak öznitelik seçme işlemi yapılacağında, özellikler hesaplanan skorlara göre büyükten küçüğe doğru sıralanır ve en yüksekten başlayarak istenilen sayıda özellik elde edilmiş olur.

Ki-Kare öznitelik seçme yöntemi de bu çalışmada tanımlanan hibrit öznitelik seçme modelinde kullanılan yöntemlerden bir diğeridir. Bu yöntemin hibrit modele dahil edilmesinin sebebi yine öznitelikleri skorlarken diğerlerinden daha farklı bir yaklaşımda bulunmasıdır.

3.2. Sınıflandırma (Classification)

3.2.1. Naive Bayes

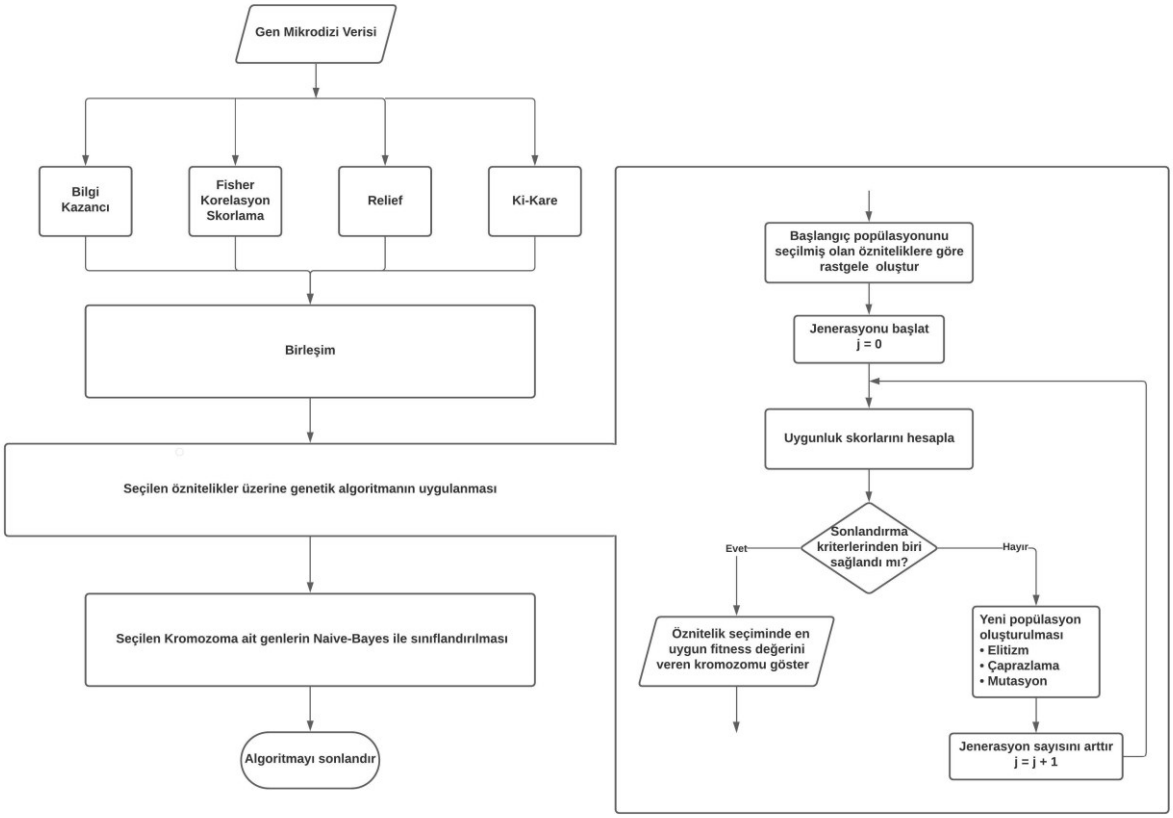
Bu sınıflandırıcı hem ayrık hem de sürekli özniteliklerle gerçek hayat problemleri üzerinde yüksek performansla çalışabilen hızlı ve artırımlı bir algoritmadır. Bu algoritmada etkin sınıflandırıcı her durumun olasılıklarını ayrı ayrı hesaplar ve olasılık değeri en yüksek olan sınıfa atanır. Naive Bayes sınıflandırıcısının kararları bilgi kazanımlarının toplamı olarak ifade edilebilir. Bu sınıflandırıcı, özniteliklerin verilen sınıftan bağımsız olduğunu varsayar. Bu durum öğrenmeyi oldukça basit bir hale getirir. Burada bağımsızlık genellikle zayıf bir varsayımdır fakat pratikte Naive Bayes'e göre daha karmaşık sınıflandırıcılarla başarı bağlamında oldukça yüksek başarımla rekabet eder. Buna karşın öznitelikler arası güçlü bağımlılıkların olduğu alanlarda yüksek performans görülmeyebilir. Naive Bayes yaklaşımının uygulama formülleri Denklem 9 ve 10'da gösterilmiştir [25].

$$P(a_1, a_2, \ldots a_n | v_j) = argmax_{v_j \epsilon V} \prod_i P(a_i | v_j) \quad (9)$$

$$v_{NB} = argmax_{v_i \epsilon V}(v_j) \prod_i P(a_i | v_j) \quad (10)$$

3.3. Önerilen Hibrit Yaklaşım (The Proposed Hybrid Approach)

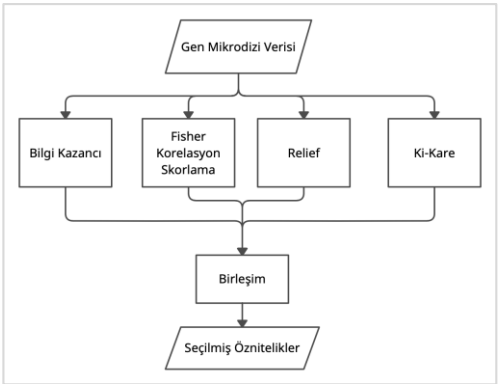
Önerilen hibrit yöntemde, gen mikrodizi verisi girdi olarak alınır ve Bilgi Kazancı, Fisher Korelasyon Skorlama, ReliefF, Ki-Kare isimli yöntemler bu veriye ayrı ayrı uygulanır ve elde edilen sonuçların küme bileşimi alınırak yeni gen alt veri kümesi elde edilir. Bu alt veri kümesi Genetik Algoritma'nın girdisini oluşturur ve başlangıç popülasyonu bu verilere göre oluşturulur. Daha sonra bu alt veri kümesine Genetik Algoritma uygulanarak, sınıflandırmada en etkin olan öznitelikler bulunur. Bu işlemler önerilen yaklaşımın sınıflandırmada etkin olan özniteliklerin seçiminde çok etkilidir. Son olarak, Genetik Algoritma'nın işlemleri sonucunda geri döndürülen öznitelik alt kümesi kullanılarak Naive Bayes sınıflandırıcısı eğitilir ve kanser sınıflandırılması yapılır. Şekil 1'deki akış şemasında önerilen modelin bütüncül yapısı sunulmuştur.



Şekil 1. *Önerilen hibrit yöntem çalışma adımları*

3.3.1. Öznitelik Seçim Yaklaşımı (Feature Selection Approach)

Önerilen yöntemde öznitelik seçimi, gen mikrodizi verisi üzerine Bilgi Kazancı, Fisher Korelasyon Skorlama, ReliefF ve Ki-Kare yöntemleri kullanılarak oluşturulan Ensemble yöntemle gerçekleştirilir. Ensemble Öznitelik Seçim Algoritması'nın ilk adımında veri kümesi öznitelikleri ayrı ayrı Bilgi Kazancı, Fisher Korelasyon Skorlama, ReliefF ve Ki-kare yöntemleri ile skorlanır. Her bir yöntem ile ölçeklenen bu özniteliklerin ilk N tanesi kaydedilir. Daha sonra elde edilen gen alt kümelerinin birleşimi sağlanır ve bu aşamalardan sonra uygun öznitelikler belirlenir. İlgili işlem adımları Şekil 2'de gösterilmiştir. Birleşim yöntemi sayesinde, farklı öznitelik seçme yöntemlerinde birden fazla sayıda yüksek skor almış ve seçilmiş özniteliklerin tekrar etmesi engellenir. Ayrıca bu Ensemble yöntem sayesinde sınıflandırmada etkin olabilecek özniteliklerin elenmesinin büyük oranda önüne geçilmiş olunur.



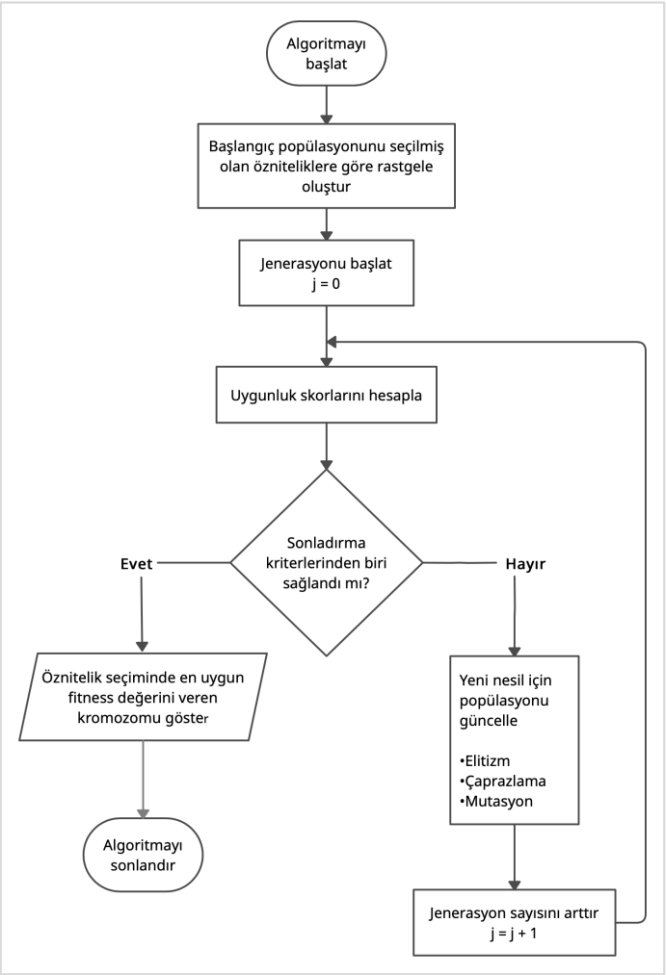
Şekil 2. *Önerilen ensemble seçim yaklaşımı*

3.3.2. Genetik Algoritma (Genetic Algorithm)

Genetik Algoritma doğadan esinlenerek oluşturulmuş evrimsel bir optimizasyon algoritmasıdır [12]. Bu algoritma, çevreye adaptasyon konusunda en başarılı olan bireyin hayatta kalması prensibine dayanan stokastik bir algoritmadır. Stokastik bir algoritma olduğu için büyük veri uzayındaki aramayı daha hızlı yapabilmekte ve optimal sonuçlara yakın çözümlerin bulunmasında önemli avantajlara sahiptir.

Genetik Algoritmadaki genler, gen mikrodizi verilerindeki özniteliklere karşılık gelmektedir. Belirli sayıda gen bir araya gelerek kromozomlar oluşturur. Bu sebepten dolayı kromozomlar aslında gen mikrodizi veri kümesinin özniteliklerinin farklı alt kümeleridir. Kromozomlar ise bir araya gelerek popülasyonu oluşturmaktadır.

Ele alınan problemin çözümü için belirlenen algoritmanın başlangıç aşamasında bir önceki Ensemble Öznitelik Seçimi aşamasında seçilmiş gen alt kümesi kullanılarak rastgele bir popülasyon oluşturulur ve popülasyondaki tüm kromozomlar için uygunluk (fitness) skoru hesaplanır. Fitness skoru hesaplanan kromozomlar, fitness skorlarına göre büyükten küçüğe sıralanır ve elitizm yüzdeliği içerisinde olan kromozomlar doğrudan sonraki nesle (jenerasyon) aktarılır. Sonrasında elitizm ile seçilen kromozomlar kendi aralarında çaprazlanır. Çaprazlama işlemi sonrası mutasyonlar ile gen değişimi gerçekleşebilir. Bu şekilde yeni jenerasyon oluşturulur. Yeni jenerasyon için tekrar fitness skorları hesaplanır ve durdurma koşulu kontrol edilir. Eğer koşul sağlanmazsa elitizm aşamasına geri dönülür ve kalan adımlar tekrarlanır. Eğer koşul sağlanırsa seçilen kromozom, problemdeki biyobelirteç genler olarak belirlenir. Bu algoritmanın bu temel işlem adımları aşağıda kısaca açıklanmış olup; çalışma adımları Şekil 3'de sunulmuştur.



Şekil 3. *Genetik algoritma çalışma adımları*

3.3.2.1. Gen Havuzunun Belirlenmesi (Determination of Gene Pool)

Bu çalışmada genetik algoritmanın gen havuzu, ensemble öznitelik seçimi adımında elde edilen öznitelik altkümesi kullanılarak oluşturulmaktadır.

3.3.2.2. Popülasyon Oluşturma (Generation of Population)

Genetik algoritmada popülasyonu, kromozomlar oluşturmaktadır. Kromozomları ise genler oluşturmaktadır. Popülasyonu oluşturmak için kromozomların oluşturulması gerekmektedir. Kromozomların uzunluğu kaç tane gen taşınacağını belirler. Örneğin; bir kromozomun uzunluğu 10 ise ilgili kromozom içerisinde 10 farklı gen bulunur. Kromozomların oluşturulması süreci, kromozom uzunluğu kadar rastgele ve birbirinden farklı genlerin seçilmesiyle gerçekleştirilir. Bu şekilde popülasyon sayısı kadar rastgele kromozomlar oluşturulur.

3.3.2.3. Fitness Fonksiyonu (Fitness Function)

Genetik algoritmanın başarısını belirleyen en önemli fonksiyon fitness fonksiyonudur. Çünkü bir sonraki jenerasyona hangi kromozomların aktarılacağını bu fonksiyon belirler. Bu çalışmada fitness skoru, f1 skor olarak belirlenmiştir. Buradaki fitness fonksiyonu uygunluk skorunun hesaplanmasında kullanılır. Bu çalışmada accuracy yerine F1 Uygunluk değerinin kullanılmasının en temel sebebi sadece false-negative ya da false-positive değil; aynı zamanda tüm hata maliyetlerini de içerecek bir ölçme metriğine ihtiyaç duyulmasıdır [26].

Her bir jenerasyonda, ilgili nesilde bulunan tüm kromozomlar fitness fonksiyonu ile skorlanır. Fitness fonksiyonu her bir kromozomun içerdiği genlerden oluşan veri alt kümeleri ile makine öğrenmesi modelini eğitir ve eğitilen modelin F1 skorlarını hesaplar. Fitness fonksiyonunda makine öğrenmesi modeli olarak Gausiann Naive Bayes yaklaşımı kullanılmıştır.

3.3.2.4. Yeni Jenerasyonun Belirlenmesi (Determination of New Generation)

Genetik algoritmanın temel motivasyonlarından biri en iyi kromozomların sonraki nesillere sistematik olarak aktarılmasıdır. Bunun sağlanması için öncelikle popülasyondaki tüm kromozomlar, fitness fonksiyonu ile skorlanır. Sonrasında kromozomlar fitness skorlarına göre yüksek skordan düşük skora doğru sıralanır. Daha önceden belirlenen elitizm oranına karşılık gelen kromozom sayısı kadar birey en yüksek skorlu kromozomdan başlanarak seçilir. Burada seçilen kromozomlar, doğrudan sonraki nesile aktarılır. Fakat yeni nesile aktarılan kromozomlar yeni neslin tamamını oluşturmaz. Yeni nesilde eksik kalan kromozomlar, elitizm ile seçilen kromozomların kendi aralarında çaprazlanmasıyla oluşturulan yeni kromozomlarla tamamlanır.

3.3.2.5. Çaprazlama ve Mutasyon (Crossover and Mutation)

Elitizim ile seçilen kromozomlar kendi aralarında çaprazlanarak yeni jenerasyonun kalan kromozomlarını oluşturur. Çaprazlama (cross-over) işlemi sırasında her yeni kromozom elitizm ile seçilen kromozomlar arasından rastgele belirlenen iki ebeveyn kromozom tarafından oluşturulur. Yeni kromozomun genleri belirlenirken üç olasılık mevcuttur. İlk olasılıkta seçilen gen, birinci ebeveynden alınmış olabilir. İkinci olasılıkta gen ikinci ebeveynden alınmış olabilir ve son olasılıkta ilgili gen mutasyon sonucu elde edilmiş olabilir. Mutasyon işlemi sayesinde yeni kromozoma dahil edilecek genin ebeveyn kromozomlardan seçilen bir gen ile değil; gen havuzundan rastgele seçilmesiyle gerçekleştirilir. Mutasyonun gerçekleşme ihtimali, daha önceden belirlen mutasyon yüzdesi parametresi ile belirlenir.

3.3.2.6. Durdurma Kriteri (Stopping Condition)

Algoritmadaki durdurma kriteri iki farklı durumda sağlanabilir. Birinci durum, Genetik Algoritmanın başlangıçta belirlenen maksimum jenerasyon sayısına ulaşılmasıdır. Algoritma bu sayıya ulaştığında işlemler tamamlanır. İkinci durum ise algoritma çalışmadan önce belirlenen kontrol aralığı durumuna ve hata parametrelerine bağlıdır. Bu durum, kontrol aralığı parametresi ile belirtilen sayıdaki jenerasyonlar

boyunca tüm nesillerdeki en yüksek fitness skorlarının değişiminin belirlenen değişim oranından büyük olmadığında gerçekleşir.

Algoritma sonlandırıldığında, en yüksek uygunluk skoruna sahip kromozom belirlenir. Sonrasında belirlenen kromozom içerisindeki genlerden oluşan veri alt kümesi, sınıflandırıcının eğitilmesi için önerilen modelin sonraki aşamasına girdi olarak verilir.

3.3.3. Sınıflandırma (Classification)

Genetik Algoritma'nın işlem adımları sonrasında seçilen gen alt kümesi, sınıflandırmada en etkin genler olarak kabul edilir. Bu genlerden oluşan veri altkümesi ile sınıflandırma yapılması için bir makine öğrenmesi modeli eğitilir. Bu problem için sınıflandırma işlemlerinde başarılı olan ve istatistiksel alt yapısı problem çözümünde etkin olan Gaussian türü Naive Bayes Algoritması kullanılmıştır. Sınıflandırıcı ile ilgili formüller Denklem 11 ve 12'de verilmiştir.

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \tag{11}$$

$$V_{nb} = argmax_{v_j \epsilon V} P(V_j) \prod P(a_i|v_j) \tag{12}$$

Burada $\boldsymbol{n}$, $\boldsymbol{v = v_j}$ için eğitim örneklerinin sayısını gösterirken; $\boldsymbol{n_c}$, $\boldsymbol{v = v_j}$ ve $\boldsymbol{a = a_i}$ için belirlenen örneklerin sayısını ifade eder. Denklem 12'deki $\boldsymbol{V_{nb}}$ ise sınıf değişkenini gösterir.

3.4. Deneysel Çalışmalar (Experimental Studies)

Bu bölümde, hibrit seçim ve genetik algoritma temelli önerilen yaklaşımın performans değerlendirme sonuçları sunulmuş ve sonuçlar analiz edilmiştir.

3.4.1. Veri Kümeleri (Datasets)

Bu çalışmada literatürde yaygın olarak kullanılan üç kanser türü için üç ayrı gen mikrodizi veri kümesi üzerinde testler yapılmıştır. Bu veri setleri Tablo 1'de sunulmuştur. Ayrıca çalışmada kullanılan genetik algoritmaya ait parametrelerin belirlenmesi için daha güncel teknoloji ile elde edilmiş prostat kanseri gen mikrodizi veri seti de kullanılmıştır.

Tablo 1. *Veri Kümesi Özellikleri*

Veri Kümesi	*Sınıf Türü Sayısı*	*Gen Sayısı*	*Örnekler Sayısı*	*Sınıf Dağılımı*
Leukemia	2	7129	72	25 AML- 47 ALL
Central Nervous System	2	7129	60	39 Class0 ve 21 Class1
Colon Tumor	2	6500	62	22 Pozitif ve 40 Negatif

Çalışmada kullanılan ilk veri kümesinde, kolon kanseri hastalarından toplanan 62 örnek ele alınmıştır. Bunlar arasından 40 tanesi tümör biyopsisi sonucunda negatif olarak etiketlenmişken; 22 tanesi aynı insanların sağlıklı kısımlarından alınarak pozitif etiketlenmiştir. Veri kümesinde toplamda 2000 adet özellik bulunmaktadır [27].

İkinci veri kümesi, merkezi sinir sistemi ile ilişkili tümör hastalıkları için kişilerin hayatta kalıp kalamadığı bilgisini gösteren veri setidir. Bu veri setindeki birinci sınıf, tedaviden sonra hayatta kalan hastaların grubunu gösterirken; ikinci sınıf, tedavide başarısız olanların grubunu ifade eder. Veri seti 60 hasta örneği içerir. Bunlardan 21'i survivors olarak ifade edilir ve Sınıf1 (Class-1) olarak etiketlenir. Kalan 39'u ise failures olarak isimlendirilen Sınıf0 (Class-0) etiketini ifade eder. Ayrıca bu veri setinde 7129 gen bulunmaktadır [28].

Üçüncü veri kümesinde, lösemi hastalarından toplanan 72 örnek vardır. Bunların arasından 25 tanesi Akut Miyeloid Lösemi (AML) hastasıyken; 47 tanesi ise Akut Lenfoblastik Lösemi (ALL) hastasıdır. Bu veri kümesi toplamda 7129 gen verisine sahiptir [29].

Ensemble öznitelik seçim yaklaşımının ve genetik algoritmanın parametre değerleri belirlenirken, daha güncel teknoloji ile elde edilen ve CuMiDa [30] çalışmasıyla sunulan prostat kanseri veri kümesi kullanılmıştır. Prostate_GSE6919_U95C veri kümesi içerisinde toplamda 115 örnek ve 12648 özellik yani gen verisi bulunmaktadır. Veri kümesinin analiz zorluğu, az sayıda örneğe karşılık çok sayıda özelliğin bulunmasından kaynaklanmaktadır. Bu veri kümesinde veriler kanserli veya normal olarak etiketlenerek algoritmaların parametrik değerleri belirlenmiştir.

3.4.2. Parametrelerin Belirlenmesi (Determination of Parameters)

Bu çalışmada önerilen hibrit yöntemin birçok parametresi bulunmaktadır. Ensemble öznitelik seçimi kısmında seçilecek ilk N adet gen parametresi; genetik algoritmada kontrol aralığı, değişim miktarı, maksimum jenerasyon sayısı, popülasyon sayısı, kromozom uzunluğu, elitizm yüzdesi ve mutasyon oranı parametreleri sınfılandırma sonuçlarını doğrudan etkilemektedir. Bu sebepten dolayı parametre değerlerinin uygun yöntemler ile belirlenmesi sağlanmıştır.

Tablo 2. *Genetik Algoritma ile ilgili parametre değerleri*

Seçilen N gen	***Popülasyon x Kromozom***	***Elitizm Oranı***	***Mutasyon Oranı***
550	600 genlik veri	%25	%12

Bu çalışmadaki temel parametre değerleri Tablo 2'de sunulmuştur. Sınıflandırma için ilk N adet gen 550 olarak belirlenmiştir. "Popülasyon x Kromozom" kısmı genleri temsil eder ve bu çarpımın belirlenen değeri 600'dür. Elitizm ve mutasyon oranları sırasıyla; yüzde 25 ve yüzde 12 olarak belirlenmiştir. Tüm bu değerler *Prostate_GSE6919_U95C* veri seti[30] üzerindeki testlerle belirlenmiştir. Tablo 2'deki parametre içeriklerinin belirlenmesi ile ilgili adımlar aşağıdaki başlıklarda özetlenmiştir.

3.4.2.1. İlk N Gen Parametresi (Determination of First N Gene)

N parametresi, ensemble öznitelik seçim yöntemi ile skorlanıp, yüksek skordan küçük skora doğru sıralanan genlerin, ilk kaç adedinin seçileceğini ifade eder. Bu değişken değerinin doğru belirlenebilmesi, çalışmanın sonucunu doğrudan etkilemektedir. N parametresinin içeriğinin uygun bir şekilde belirlenebilmesi için önerilen yöntem, parametrenin iteratif şekilde arttırılıp her bir arttırım sonucunda hesaplanan F1 skor değerlerinin karşılaştırılmasıdır. Deneysel çalışmalarda N parametresinin her bir arttırım işleminden sonra seçilen genleri Genetik Algoritma'ya girdi olarak verilmiştir. Genetik Algoritma ise bu girdilere göre 10 kez çalıştırılmış ve işlem sonlandırıldığında elde edilen F1 skorları karşılaştırılmıştır. Sonuçlar incelendiğinde, en yüksek başarı ilk 550'şer özniteliğin seçimi ve bileşimi ile elde edilmiştir. Bu testten sonra N sayısı 550 olarak belirlenmiştir.

3.4.2.2. Popülasyon Ve Kromozom Uzunluğu Parametreleri (Determination of Population and Chromosome Lengths Parameters)

Kromozom uzunluğu ve popülasyon sayısı parametrelerinin nasıl belirleneceği önemli bir konudur. Literatürdeki araştırmalar incelendiğinde, deneme ve yanılma yöntemi bu parametrelerin belirlenmesinde en çok kullanılan yaklaşımlardan biridir. Deneysel çalışmalarda algoritmanın çalışma süreleri göz önüne alındığında "kromozom uzunluğu x popülasyon" değişkeni içeriğinin 600'den fazla olması işlemsel/zamansal karmaşıklığı arttırmakta ve deneme yanılma yönteminin işlevselliğini kaybettirmektedir. Yapılan denemeler ve analizler sonucunda ilgili parametre değerinin 600 olmasına karar verilmiştir.

3.4.2.3. Elitizm Orani Parametresi (Determination of Elitizm Ratio Parameter)

Elitizm yüzdesi, genetik algoritmanın başarısına önemli ölçüde etki etmektedir. Bu parametrenin küçük seçilmesi daha kısır jenerasyonlara, sebep olurken; büyük seçilmesi ise yerel maksimum (optimum)

sonuçtan uzaklaşılmasına ve daha düşük yerel maksimum sonuçlar elde edilmesine sebep olabilir. Bu parametrenin optimal değerinin tespiti için önerilen yöntem deneme yanılma yöntemidir. %5 ile %30 arasındaki tüm değerler için Genetik Algoritma 10'ar kez çalıştırılmış ve ortalama sonuçlar karşılaştırılmıştır. Deney sonucunda Elitizm parametresinin değeri %25 olarak belirlenmiştir.

3.4.2.4. Mutasyon Oranı Parametresi (Determination of Mutation Ratio Parameter)

Mutasyon oranı parametresi de elitizm parametresi gibi Genetik Algoritma'nın başarısını doğrudan etkiler. Genetik algoritma doğadan esinlenen bir algoritma olduğu için ele alınan mutasyon parametresi de doğadaki mutasyon ihtimaline uygun olarak oldukça küçük belirlenmelidir. Fakat genetik algoritmanın uygulandığı probleme göre mutasyon oranının farklı değerler şeklinde belirlenmesi de mümkündür. Mutasyon oranının yüksek olması, yerel maksimuma ulaşılma ihtimalini düşürmektedir. Fakat bu durum mevcut yerel maksimumdan daha yüksek değere sahip farklı bir yerel maksimuma ulaşılma ihtimalini de arttırabilmektedir. Dolayısıyla bu parametrenin optimal değerinin tespiti, algoritmanın test sonuçlarına fazlaca etki etmektedir. Bu parametrenin optimal değerinin tespiti için önerilen yöntem yine deneme yanılma yöntemidir. Yapılan deneylerde mutasyon oranı %2 ile %20 arasındaki tüm değerler için genetik algoritma 10'ar kez tekrar çalıştırılmış ve sonuçların ortalaması karşılaştırılmıştır. Deney sonucunda mutasyon oranı %12 iken en yüksek başarı elde edilmiştir. Bu sebepten dolayı mutasyon oranı parametresi %12 olarak belirlenmiştir. Tüm bu belirlenen parametre değerlerine göre yapılan testlerin sonuçları Tablo 3 ve 4'te özetlenmiştir.

3.4.3. Performans Değerlendirme Metrikleri (Metrics of Performance Evaluation)

Makine öğrenme modellerinin başarısını ölçmek için bazı performans ölçme yöntemlerinden yararlanılır. Bu ölçümün gerçekleştirilmesinde gerçek ve tahmin edilen sınıflar için True Positive (TP) değişkeni pozitif tahmin edilen ve gerçekte de pozitif olanları ifade ederken; True Negative (TN) değişkeni negatif tahmin edilen ve gerçekte de negatif olanları gösterir. Yanlış tahminlemede ise False Negative (FN) negatif olarak tahmin edilen ama gerçekte pozitif olanları ifade ederken; False Positive (FP) pozitif olarak tahmin edilen fakat gerçekte negatif olanları gösteren değişkenleri gösterir. Bu paretreler kullanılarak başarıların test edilenmesinde bir karışıklık/hata matrisi (confusion matrix) tanımlanır [31].

Tablo 3. *Önerilen yöntemin performans değelendirme metrikleri (yüzde olarak)*

Veri Kümesi	***Doğruluk***	***Duyarlılık***	***Kesinlik***	***F1-Skor***
Leukemia	98.28	98.66	98.38	98.26
Central Nervous System	93.66	95.25	92.95	93.18
Colon Tumor	93.33	83.00	90.00	86.03

Önerilen sistemin performansı Denklem 13-16 arasında verilen formüller kullanılarak ölçülür. Bu denklemlerle doğruluk (*accuracy*), duyarlılık (*precision*), kesinlik (*recall*) ve F1 skor değerleri hesaplanır.

$$doğruluk = \frac{TP + TN}{TP + FN + TN + FP} \tag{13}$$

$$kesinlik = \frac{TP}{TP + FP} \tag{14}$$

$$duyarlılık = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2xkesinlikxduyarlılık}{kesinlik + duyarlılık} \tag{16}$$

3.4.4. K-Katlamalı Çapraz Doğrulama (K-Fold Cross Validation)

K-kez çapraz doğrulamada, veri kümesi rastgele k adet, birbirini dışlayan alt kümelere bölünür. (*1/k*) kadarlık örnek, test alt kümesi olarak belirlenirken; geriye kalan (*k-1/k*) kadarlık örnek eğitim verisi olarak belirlenir ve model bu verilerle hem eğitilir hem de test edilir [33]. Bu işlem, test verisinin öncekilerden farklı olması şartı ile *k* kez tekrar edilir. Burada modelin performans metrikleri hesaplanırken tüm iterasyondaki performans metriklerinin ortalaması alınır.

4. BULGULAR VE TARTIŞMA (RESULTS AND DISCUSSION)

Bu bölümde literatürdeki diğer çalışmaların sonuçlarıyla bu makalede yapılan çalışmanın sonuçları karşılaştırılmış olup; çıktılara göre bazı değerlendirmeler yapılmıştır. Tablo 4, literatürdeki güncel 6 çalışma ve önerilen yöntemin doğruluk değerlerini karşılaştırmalı olarak göstermektedir. Bu tabloda (-) ile doldurulmuş hücreler ismi geçen çalışmada ilgili veri setinin kullanılmadığı anlamına gelir. Karşılaştırılan yöntemlerin ilgili test verileri üzerindeki sonuçları aşağıda analiz edilmiştir.

Tablo 4. *Önerilen yöntemin ve Diğer Yöntemlerin Sınıflandırma Başarıları*

Çalışma	*Öznitelik Seçme Yöntemi*	*Sınıflandırma Yöntemi*	*Leukemia*	*Central Nervous System*	*Colon Tumor*
T. Nguyen	Analytic hierarchy process (AHP)	Hidden Markov models	96.48%	-	-
S. Hengpraprohm	Signal-to Noise Ratio (SNR)	GA	91.9%	-	-
H. Yu	A weighted metric	Decision rule	95.55%	-	85.49%
C. Gunavathi	T-Statistics SNR, F-Test values	KNN, SVM	-	81.25%	-
J. C. H. Hernandez	GA	SVM	91.5%	-	84.6%
H. Salem	IG	SGA	97.06%	85.48%	86.67%
Önerilen Yöntem	Ensemble	GA	98.28%	93.66%	93.33%

Çalışmada önerilen model ayrı ayrı Leukemia, Central Nervous System ve Colon Tumor veri kümeleri ile eğitilmiş ve sırasıyla %97.06, %85.48 ve %86.67 değerleri elde edilmiştir. Tablo 3'te, bu çalışmada önerilen metodun ele alınan Leukemia, Central Nervous System ve Colon Tumor veri kümelerinin hepsinde literatürdeki güncel çalışmalara kıyasla daha yüksek sınıflandırma doğruluğu elde ettiği görülmektedir. Bir diğer test sonucu Tablo 4'te gösterilmektedir. Burada algoritmanın ilgili veriler üzerindeki stabil çalışma performansı test edilmiştir. Önerilen model her bir veri seti üzerinde 10 kez çalıştırılmış ve sonuçlar dört farklı metriğe göre elde edilmiştir. Bu sonuçlar tüm çalıştırmalarda elde edilen genel sonuçların ortalamasını göstermektedir. Elde edilen tüm sonuçlar değerlendirildiğinde, önerilen yöntemin hem farklı veri setleri üzerinde başarıyla çalıştığı hem de literatürdeki güncel yaklaşımlardan daha iyi performansa sahip olduğu anlaşılmaktadır. Önerilen modelin başarılı olmasının temel sebebi, çalışılan verinin yani gen mikrodizi verisinin temel problemlerinin çoğuna etkili çözümler bulunmasıdır. Yüksek öznitelik sayısı, az örnek sayısı, verilerin gürültü olması, sınıflandırmada etkin genlerin sayısının az olması gibi problemler gen mikrodizi verisinin temel problemlerindendir [16]. Yüksek öznitelik sayısı önerilen ensemble öznitelik seçimi yöntemi ile azaltılmıştır. Fakat elde edilen yeni öznitelik alt kümesindeki öznitelik sayısı, sınıflandırmada etkin özniteliklerin alt kümesindeki öznitelik sayısından hala oldukça fazladır. Bu sebepten dolayı, indirgenen öznitelikler arasında optimum gen kombinasyonunu bulmak amacıyla stokastik arama yapma ihtiyacı doğmuştur. Bu arama işlemi, stokastik arama algoritması olan ve mevcut problemlere göre özelleştirilmiş Genetik Algoritma kullanılarak gerçekleştirilmiştir. Bu özelleştirme de genetik algoritmanın

gen havuzuna, önerdiğimiz ensemble öznitelik seçme yöntemiyle indirgenmiş öznitelikler tanımlanmıştır. Popülasyondaki kromozom sayısı, kromozomdaki gen sayısı, elitizm yüzdesi, mutasyon yüzdesi gibi parametreler birçok farklı gen mikrodizi veri kümeleri kullanılarak deneme ve yanılma yöntemiyle optimize edilmiştir. Genetik Algoritmanın başarısını doğrudan etkileyen fitness fonksiyonu: az örnek sayısı, verilerin gürültülü olması ve sınıflara ait örnek sayılarının dengesiz olması gibi problemleri indirgeyecek şekilde düzenlenmiştir. Fitness fonksiyonu sonraki nesile aktarılacak kromozomların skorlanması için kullanılır. Dolayısıyla yanlış bir skorlama yöntemi kullanılması seçilecek kromozomların sınıflandırmada en etkili genler olması olasılığını düşürmektedir. Çalışmada fitness fonksiyonu olarak Naive Bayes sınıflandırıcısı kullanılmıştır. Çünkü Naive Bayes sınıflandırıcısı, sınıflara ait örnek sayısında dengesizlik olması durumundan etkilenmemesi, örnek sayısı az olan eğitim verilerinde diğer sınıflandırma yöntemlerine göre daha başarılı olması, hızlı eğitilmesi gibi avantajlara sahiptir [32].

5. SONUÇ (CONCLUSION)

Gen mikrodizi verileri kullanılarak kanser hastalıklarının sınıflandırması problemi, veri madenciliği ve makine öğrenmesi alanlarında oldukça fazla dikkat çeken konulardan biridir. Bu çalışmada dünyadaki en önemli sağlık sorunlarından biri olan kanserin erken teşhisi konusuna odaklanılmıştır. Burada DNA mikrodizi verileri kullanılarak kanser hastalıklarını ayırt etmeyi amaçlayan yeni bir metodoloji önerilmiştir. Bu metodolojiye göre öncelikle bilgi kazancı, fisher korelasyon skorlama, relief ve ki-kare yöntemleri kullanılarak elde edilen alt kümelerin birleşimi sağlanır. Bu yaklaşımların tümü ensemble metot ile bir araya getirilerek öznitelik seçimleri yapılır. Daha sonrada genetik algoritma kullanılarak öznitelik azaltma işlemleri uygulanır. Son aşamada ise naive bayes sınıflandırıcısı kullanılarak kanser türlerinin sınıflandırması işlemi başarılı bir şekilde tamamlanır. Önerilen metodoloji literatürde yaygın olarak kullanılan üç farklı veri kümesiyle test edilmiş ve elde edilen deneysel sonuçlar literatürdeki güncel altı farklı çalışmanın sonuçlarıyla karşılaştırılmıştır. Çalışmada ele alınan Leukemia, Central Nervous System ve Colon Tumor veri kümelerinin tümünde sınıflandırma doğruluğu kriter sonuçlarının diğer çalışmaların sonuçlarına kıyasla daha iyi olduğu anlaşılmıştır. Bu çalışmadaki literatürel karşılaştırmalar sonucunda önerilen yöntemin diğer yöntemlerden daha başarılı olduğu gözlemlenmiştir. Çalışma kapsamında önerilen hibrit yaklaşım hem sınıflandırma performansına doğrudan etki edecek şekilde tasarlanmış hem de geliştirilebilir esnekliklere sahip bir yapıda sunulmuştur. Elde edilen test sonuçlarına dayanarak önerilen hibrit yaklaşımın kanser hastalıklarının sınıflandırılması problemlerinde genellikle başarılı sonuçlar verdiği ve farklı özelliklerdeki veri setlerinde performasın en üst seviyeye çıkarılması konusunda yüksek potansiyele sahip olduğu söylenebilir. Bu çıkarımlara göre önümüzdeki süreçte önerilen yöntemin farklı karakteristik özelliklerdeki veri setlerine uygulanması planlanmaktadır. Buna göre elde edilecek deneysel çıktıların biyolojik anlamlılık açısından analiz edilmesi sağlanacaktır.

KAYNAKLAR (REFERENCES)

[1] SEER Training Modules, Cancer Classification. U. S. National Institutes of Health, National Cancer Institute. (2019, May 21). <https://training.seer.cancer.gov/disease/categories/classification.html>.

[2] Al-shamasneh, A. R. M., Obaidellah, U. H. B. (2017). Artificial Intelligence Techniques for Cancer Detection and Classification: Review Study. European Scientific Journal, 13(3). https://doi.org/10.19044/esj.2016.v13n3p342.

[3] Russo, G., Zegar, C., Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer - Oncogene. Oncogene, 22, 6497–6507. doi: 10.1038/sj.onc.1206865.

[4] Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Inform. Sci., 282, 111–135. doi: 10.1016/j.ins.2014.05.042.

[5] Yu, H., Ni, J., Dan, Y., Xu, S. (2012). Mining and integrating reliable decision rules for imbalanced cancer gene expression data sets. Tsinghua Sci. Technol., 17(6), 666–673. doi: 10.1109/TST.2012.6374368.

[6] Gunavathi, C., Premalatha, K. (2014). Performance analysis of genetic algorithm with KNN and SVM for Feature Selection in Tumor Classification. World Academy of Science, Engineering and Technology, International Journal of Computer, Control, Quantum and Information Engineering, 8(8), 1390–1397.

[7] Hernandez, J. C. H., Duval, B., Hao, J.-K. (2007). A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Springer. doi: 10.1007/978-3-540-71783-6_9.

[8] Salem, H., Attiya, G., El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. Appl. Soft Comput., 50, 124–134. doi: 10.1016/j.asoc.2016.11.026.

[9] Nguyen, T., Khosravi, A., Creighton, D., Nahavandi, S. (2015). Hidden Markov models for cancer classification using gene expression profiles. Inform. Sci., 316, 293–307. doi: 10.1016/j.ins.2015.04.012.

[10] Hengpraprohm, S. (2013). GA-Based Classifier with SNR Weighted Features for Cancer Microarray Data Classification. International Journal of Signal Processing Systems, 1(1), 29–33. doi: 10.12720/ijsps.1.1.29-33.

[11] Gumaei, A., Sammouda, R., Al-Rakhami, M., AlSalman, H., & El-Zaart, A. (2021). Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression. Health Informatics J., 27(1), 1460458221989402. doi: 10.1177/1460458221989402

[12] Candan, H., Durmuş, A., Harman, G. (2019). Genetik Algoritma ve Sınıflandırıcı Yöntemler ile Kanser Tahmini. Veri Bilimi, 2(1), 30–34.

[13] Kahraman M., Kaya, M. (2010). Çok amaçlı genetik algoritma kullanarak DNA mikrodizi verilerinin kümelenmesi. (20 Ağustos 2021). Retrieved from https://tez.yok.gov.tr (tez no: 269977).

[14] Turgut S., Dağtekin M., Ensari T. (2017). Makine öğrenmesi yöntemleri kullanarak kanser teşhisi. (22 Ağustos 2021). Retrieved from https://tez.yok.gov.tr (tez no: 487852).

[15] Su, Q., Wang, Y., Jiang, X., Chen, F., Lu, W.-c. (2017). A Cancer Gene Selection Algorithm Based on the K-S Test and CFS. Biomed Res. Int., 2017, 1645619. doi: 10.1155/2017/1645619.

[16] Tinker, A. V., Boussioutas, A., & Bowtell, D. D. L. (2006). The challenges of gene expression microarrays for the study of human cancer. Cancer Cell, 9(5), 333–339. doi: 10.1016/j.ccr.2006.05.001

[17] Motieghader, H., Najafi, A., Sadeghi, B., & Masoudi-Nejad, A. (2017). A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Inf. Med. Unlocked, 9(C), 246–254. doi: 10.1016/j.imu.2017.10.004

[18] Hong, H., Hong, Q., Liu, J., Tong, W., & Shi, L. (2013). Estimating relative noise to signal in DNA microarray data. Int. J. Bioinf. Res. Appl., 24001721. doi: 10.1504/IJBRA.2013.056085

[19] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 1–13. doi: 10.1186/s12864-019-6413-7

[20] Roobaert et al.: Information Gain, Correlation and Support Vector Machines, StudFuzz 207, 463–470 (2006).

[21] Hall, M. 1999. Correlation-based Feature Selection for Machine Learning, The University of Waikato, PhD Thesis, Hamilton.

[22] Jadhav, S., He, H., Jenkins, K. (2018). Information Gain Directed Genetic Algorithm Wrapper Feature selection for Credit Rating. Appl. Soft Comput., 69. doi: 10.1016/j.asoc.2018.04.033.

[23] Budak, H. (2018). Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(zel), 10. doi: 10.19113/sdufbed.01653.

[24] Kira, K., Rendell, L. A. (1992). The feature selection problem: traditional methods and a new algorithm. AAAI'92: Proceedings of the tenth national conference on Artificial intelligence. AAAI Press. doi: 10.5555/1867135.1867155.

[25] Islam, M. J., Wu, Q. M. J., Ahmadi, M., Sid-Ahmed, M. A. (2007). Investigating the performance of naive-Bayes classifiers and K-nearest neighbor classifiers. 2007 International Conference on Convergence Information Technology (ICCIT 2007).

[26] Chicco, D. and Giuseppe J., "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", BMC genomics 21.1 (2020): 1-13.

[27] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. U.S.A., 96(12), 6745–6750. doi: 10.1073/pnas.96.12.6745.

[28] Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., ...Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression - Nature. Nature, 415, 436–442. doi: 10.1038/415436a.

[29] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ...Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439), 531–537. doi: 10.1126/science.286.5439.531.

[30] Feltes, B. C., Chandelier, E. B., Grisci, B. I., Dorn, M. (2019). CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. J. Comput. Biol., 26(4), 376–386. doi: 10.1089/cmb.2018.0238.

[31] Islam, Md. M., Iqbal, H., Haque, Md. R., Hasan, Md. K. (2017). Prediction of breast cancer using support vector machine and K-Nearest neighbors. 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). IEEE. doi: 10.1109/R10-HTC.2017.8288944.

[32] Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2010). Naïve bayes variants in classification learning. 2010 International Conference on Information Retrieval & Knowledge Management (CAMP). IEEE. doi: 10.1109/INFRKM.2010.5466902

[33] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. Morgan Kaufmann Publishers Inc. doi: 10.5555/1643031.1643047.