

Makine Öğrenmesi Algoritmalarının Performansını Artırmak için Metasezgisel Algoritma Tabanlı Öznitelik Seçimi

Sinan BALIBEY¹, Feyza ALTUNBEY ÖZBAY^{2*}, Erkan TANYILDIZI³

^{1,3} Fırat Üniversitesi Teknoloji Fakültesi Yazılım Mühendisliği Bölümü, Elazığ, Türkiye

² Fırat Üniversitesi Mühendislik Fakültesi Yazılım Mühendisliği Bölümü, Elazığ, Türkiye

¹ sinan.balibey@gmail.com, ^{2*} faltunbey@firat.edu.tr, ³ etanyildizi@firat.edu.tr

(Geliş/Received: 04/02/2025;

Kabul/Accepted: 10/03/2025)

Öz: Artan veri dünyasında, makine öğrenmesi (ML) algoritmalarının başarısı büyük ölçüde öznitelik seçimi süreçlerine bağlıdır. Öznitelik seçimi, gereksiz, etiketsiz veya gürültülü öznitelikleri eleyerek daha küçük ve anlamlı bir öznitelik kümesi oluşturmayı hedefler. Bu çalışmanın amacı, metasezgisel algoritmalar sayesinde veri kümelerinden en anlamlı öznitelikleri elde ederek ML algoritmalarının farklı kıyaslama ölçütlerine göre daha iyi performans elde etmesini sağlamaktır. Bu kapsamda, öznitelik seçimi için metasezgisel algoritmalar İkilili Parçacık Sürü Optimizasyonu ve Genetik Algoritma kullanılarak iki farklı veri seti üzerinde deneysel araştırmalar gerçekleştirilmiştir. Sonar ve Meme Kanseri veri setlerinde, K-En Yakın Komşu, Destek Vektör Makinesi ve Karar Ağaçları sınıflandırma algoritmaları amaç fonksiyonu olarak kullanılmıştır. Her iki algoritma için de farklı popülasyon büyüklüğünde yüksek iterasyonlarda öznitelik seçimi yapılmış, ardından seçilen en iyi öznitelikler farklı ML modeli (Lojistik Regresyon, Naive Bayes, Aşırı Gradyan Arttırma, Yapay Sinir Ağları, Evrimsel Sinir Ağları ve Uzun Kısa Süreli Bellek) ile değerlendirilmiştir. Sonuçlar, İkilili Parçacık Sürü Optimizasyon ve Genetik Algoritma yöntemlerinin öznitelik seçimi sonrası model doğruluklarında anlamlı artışlara yol açtığını göstermiştir. Elde edilen sonuçların metasezgisel algoritmalar kullanarak özellik seçim sürecini geliştiren araştırmacılara katkı sağlaması amaçlanmaktadır. Bu bulgular, metasezgisel algoritmaların öznitelik seçimi sürecinde etkinliğini ve çeşitli ML modelleri ile birleşmesinin performans iyileştirmedeki potansiyelini desteklemektedir.

Anahtar kelimeler: Metasezgisel algoritmalar, Öznitelik seçimi, Makine öğrenmesi algoritmaları.

Metaheuristic Algorithm Based Feature Selection for Improving the Performance of Machine Learning Algorithms

Abstract: In a world of increasing data, the success of machine learning (ML) algorithms depends heavily on the process of feature selection. Feature selection aims to create a smaller and more meaningful feature set by eliminating redundant, unlabeled, or noisy features. The aim of this study is to obtain the most meaningful features from data sets with metaheuristic algorithms and to ensure that ML algorithms achieve better performance according to different evaluation metrics. In this context, metaheuristic algorithms for feature selection, Binary Particle Swarm Optimization, and Genetic Algorithms are used for experimental research on two different datasets. In the Sonar and Breast Cancer datasets, K-Nearest Neighbor, Support Vector Machine, and Decision Tree classification algorithms were used as the objective function. For both algorithms, feature selection was performed at high iterations with different population sizes, and then the best-selected features were evaluated with different ML models (Logistic Regression, Naive Bayes, Extreme Gradient Boosting, Artificial Neural Networks, Convolutional Neural Networks, and Long Short Term Memory). The results show that Binary Particle Swarm Optimization and Genetic Algorithm methods lead to significant increases in model accuracy after feature selection. The results obtained are aimed at contributing to researchers using metaheuristic algorithms to improve the feature selection process. These findings support the effectiveness of metaheuristics in the feature selection process and the potential for performance improvement when combined with various ML models.

Key words: Metaheuristic algorithms, Feature selection, Machine learning algorithms.

1. Giriş

Günümüzde dijital veri üretimi her gün hızla artmakta olup, bu veriler metin, görüntü, sayılar, ses, video ve grafikler gibi çeşitli biçimlerde ortaya çıkmaktadır. Mühendislik, sağlık hizmetleri, bilim, endüstri, eğitim ve diğer insan faaliyetlerinin birçok alanında üretilen bu büyük veri setleri, bilgi ilerlemesine önemli katkılar sağlamaktadır [1]. Ancak, bu verilerin önemli bir kısmı ham, alakasız, kullanılabilir olmayan veya tekrarlayıcı nitelikte olup, modelleme için uygun hale getirilmesi gerekmektedir. Özellikle yüksek boyutlu veri setlerinde, gereksiz

* Sorumlu yazar: faltunbey@firat.edu.tr. Yazarların ORCID Numarası: ¹ 0009-0006-0083-6301, ² 0000-0003-0629-6888, ³ 0000-0003-2973-9389

özniteliklerin varlığı hem model performansını düşürmekte hem de hesaplama maliyetlerini artırmaktadır [2]. Bu bağlamda, öznitelik seçimi problemleri, büyük veri setlerinin boyutunu azaltırken doğruluk oranını korumak açısından kritik bir önem taşımaktadır [3].

Öznitelik seçimi, ML ve veri madenciliği alanlarında önemli bir araştırma konusu olmuştur. Özellikle, yüksek boyutlulukta özniteliklerin etkili bir şekilde seçilmesi hem modelin genel performansını artırmakta hem de hesaplama kaynaklarının verimli kullanılmasını sağlamaktadır [4]. Geleneksel öznitelik seçimi yöntemleri, yüksek boyutlulukta yerel optimumlara takılma, aşırı karmaşıklık ve yüksek hesaplama maliyeti gibi çeşitli zorluklarla karşılaşmaktadır [5]. Bu sorunların üstesinden gelmek için, metasezgisel algoritmalar giderek daha fazla önem kazanmaktadır. Metasezgisel algoritmalar, evrimsel hesaplama, sürü zekâsı, fizik tabanlı ve insan davranışına dayalı yaklaşımlar gibi çeşitli kategorilere ayrılmakta olup, büyük ve karmaşık arama uzaylarında etkili çözümler sunma kapasitesine sahiptir [2].

Metasezgisel algoritmalar, özelliği seçme sürecinde en uygun öznitelik alt kümesini bulmada oldukça verimli ve etkilidir. Genetik Algoritma, İkili Parçacık Sürü Optimizasyonu gibi doğadan esinlenen yöntemler, yüksek boyutlu veri setlerinde en iyi öznitelik kombinasyonlarını etkili bir şekilde seçme kapasitesine sahiptir [6]. Bu çalışmada, ikili öznitelik seçimi problemlerini çözmek amacıyla İkili Parçacık Sürü Optimizasyon ve Genetik algoritmalarının etkinliği incelenmiştir. Sonar ve Meme Kanseri veri setleri üzerinde gerçekleştirilen deneyler aracılığıyla, bu algoritmaların farklı amaç fonksiyonları kullanılarak öznitelik sayısını azaltırken model doğruluğunu artırmadaki başarıları değerlendirilmiştir. Elde edilen sonuçlar, çeşitli ML algoritmalarında da test edilerek, öznitelik seçiminin model performansını iyileştirme potansiyeli ortaya konmuştur.

Makalenin geri kalanı dört ana bölümden oluşacak şekilde tasarlanmıştır: İkinci bölümde ilgili literatür taraması sunulmakta, üçüncü bölümde kullanılan yöntemler ayrıntılı biçimde açıklanmakta, dördüncü bölümde deneysel çalışmalar ele alınmakta ve beşinci bölümde elde edilen sonuçlar değerlendirilmektedir.

2. İlgili Çalışmalar

Öznitelik seçimi, yüksek boyutlu veri setlerinden gereksiz ve gürültülü özellikleri eleyerek sınıflandırma doğruluğunu ve işlem verimliliğini artıran önemli bir adım olarak kabul edilmektedir. Bu bağlamda, çeşitli metasezgisel algoritmalar ile yapılan öznitelik seçimi ile ilgili yayınlar özetlenmiştir.

Yılmaz vd. yaptığı çalışmada, ilaç tasarımı için kullanılan mutajenitenin tahmin edilmesi amacıyla bir sınıflandırma problemine odaklanılmıştır. Bu bağlamda, öznitelik seçimi için Kelebek Optimizasyon Algoritması (BOA) ve Parçacık Sürü Optimizasyonu (PSO) kullanılmıştır. Çalışmada iki farklı veri seti değerlendirilmiştir: biri öznitelik seçimi ve modelleme için, diğeri ise doğrulama ve kıyaslama içindir. BOA ile öznitelik sayısı %68 oranında azaltılırken PSO ile %46 oranında bir azalma sağlanmıştır. Azaltılan öznitelikler, 6 farklı ML algoritması ile test edilmiş ve sonuçlar kıyaslanmıştır. Bu sonuçlar, metasezgisel algoritmaların hem öznitelik sayısını azaltmada hem de sınıflandırma başarımını artırmada etkili olduğunu göstermektedir [7].

Efeoğlu'nun yaptığı çalışmada, gömülü nesnelerin türünün belirlenmesi için yapay zekâ tabanlı bir yöntem geliştirilmiştir. Deneysel bir sistem oluşturularak farklı türlerde metalik ve metalik olmayan nesneler, çeşitli derinlik ve şekillerde bir kum havuzuna gömülmüştür. Ölçüm verilerinden elde edilen 2 boyutlu görüntüler, görüntü işleme teknikleri ile işlenmiştir. Öznitelik seçimi için temel bileşen analizi (PCA) ve korelasyona dayalı öznitelik seçimi (CFS) yöntemleri kullanılmış; CFS için Genetik Algoritma, PSO, Harmony Arama ve Evrimsel Arama gibi metasezgisel algoritmalar tercih edilmiştir. En yüksek doğruluk oranı (%95,8), PCA sonrası Rastgele Ağaç algoritmasıyla elde edilmiştir. Çalışma, gömülü nesnelerin tür tespitinde hibrit yapay zekâ yaklaşımlarının başarısını göstermektedir [8].

Gedikli ve Yurdusever'in yaptığı çalışmada epilepsi hastalığının teşhisi için sınıflandırma problemine öznitelik seçimi için metasezgisel algoritmalar kullanılmıştır. Veri seti EEG sinyallerinde oluşmakta ve 667 öznitelik bulunmaktadır. Tüm veri setini 1/3 oranında test geriye kalanları ise eğitim seti olacak şekilde k-en yakın komşu algoritmasıyla sınıflandırmış ve başarımını %60,05 olarak ulaşmıştır. Öznitelik seçimi için metasezgisel algoritmaların her birini 30 kez 50 popülasyonda 50 iterasyon ile çalıştırmıştır. Gri Kurt Optimizasyonu, Balina Optimizasyonu ve Harris Şahinler Optimizasyon algoritmaları ile öznitelik seçimi yapmıştır. Bu algoritmaların sınıflandırmalarının en yüksek başarımı sırasıyla %80,16, %79,20 ve %79,34 olacak şekilde görülmüş, doğrulukları yükselttiği ve öznitelik sayısını azalttığı tespit edilmiştir [9].

Açar ve Aydılek'in yaptığı çalışmada, 3 farklı veri setini 5 farklı ML algoritmasıyla 10 kat çapraz doğrulamadan geçirerek çalıştırmış doğruluklarını ve çalışma sürelerini hesaplamıştır. Öznitelik seçimi için Kolerasyon Analiz Yöntemi, Lasso Yöntemi, Hipotez Testi, PSO ve Genetik Algoritma kullanmıştır. Kullanılan öznitelik seçme yöntemlerini tekrar 5 farklı ML algoritmasıyla 10 kat çapraz doğrulamadan geçirerek

kıyaslamıştır. Elde edilen kıyaslamalar sonucunda metasezgisel algoritmalar öznelik seçiminde daha az öznelik ile daha yüksek doğruluğa ulaşmıştır [10].

Ciran ve Özbay'ın yaptığı çalışmada, dört farklı göz hastalığını sınıflandırmak için yeni bir derin öğrenme metodolojisi geliştirilmiştir. Bu metodolojide, ResNet101, DenseNet201 ve DarkNet53 gibi önceden eğitilmiş derin öğrenme mimarilerinden elde edilen özellikler birleştirilerek hibrit bir özellik havuzu oluşturulmuş ve görüntülerin daha etkili temsil edilmesi sağlanmıştır. Önemli özneliklerin seçimi için PSO algoritması kullanılmış, sınıflandırma işlemi ise Destek Vektör Makineleri ile gerçekleştirilmiştir. Ayrıca destek vektör makinelerinin hiperparametreleri, Bayesian optimizasyon yöntemi ile optimize edilerek %93,8 doğruluk oranına ulaşılmıştır [11].

Gümüş vd. yaptığı çalışmada, güç kalitesi olaylarının sınıflandırılmasında özellik seçimi için Denge Optimizasyon Algoritması (EO) ve Salp Sürü Optimizasyon Algoritması (SSA) kullanılmıştır. Güç kalitesi olaylarını içeren sentetik veri setleri, Ayrik Dalgacık Dönüşümü kullanılarak işlenmiş ve öznelikler çıkarılmıştır. Normalize edilmiş ve logaritmik olarak iki farklı veri seti oluşturulmuştur. EO ile %96,05, SSA ile %90,62 doğruluk oranlarına ulaşılmıştır. Ayrıca, en sık seçilen özelliklerin analizi histogram grafikleri ile gerçekleştirilmiş ve bu özellikler kullanılarak sınıflandırma başarımı artırılmıştır. Bu çalışma, metasezgisel algoritmaların özellik seçimi ve sınıflandırma performansı üzerindeki olumlu etkisini ortaya koymaktadır [12].

Mishra vd. yaptığı çalışmada, meme kanserinin erken teşhisi için bir öznelik seçimi yaklaşımı geliştirilmiştir. Çalışmada ikili parçacık sürü optimizasyon algoritması, özellik seçimi için ikili bir optimizasyon problemi olarak ele alınmıştır. Breast Cancer Coimbra Dataset (BCCD) ve Breast Cancer Wisconsin Diagnostic Dataset (BCWDD) üzerinde gerçekleştirilen deneylerde, öznelik seçimi sonrası BCCD için %80,83, BCWDD için %98,24 doğruluk oranlarına ulaşılmıştır. BCCD'de seçilen öznelikler arasında Yaş, BMI, Glukoz, Resistin ve Leptin bulunurken, BCWDD için 9 önemli özellik seçilmiştir. Öznelik seçimi, her iki veri setinde de sınıflandırma performansında %4-8 oranında iyileşme sağlamış ve bu yöntemin tanı sistemlerinde etkili olduğunu göstermiştir [13].

Abualigah ve Dulaimi'nin yaptığı çalışmada, öznelik seçimi problemlerini çözmek için Sinüs Kosinüs Algoritması ve Genetik Algoritmayı birleştiren hibrit bir yöntem geliştirilmiştir. SCAGA adlı bu yöntem, veri setlerindeki gereksiz ve gürültülü öznelikleri eleyerek sınıflandırma doğruluğunu artırmayı hedeflemiştir. Önerilen yöntem, UCI makine öğrenimi deposundaki veri setlerinde test edilmiş ve diğer algoritmalarla karşılaştırıldığında daha yüksek doğruluk oranlarına ve daha az seçilen özneliğe ulaşmıştır. Krvskp veri setinde doğruluk %97,97'ye, öznelik sayısı ise %50,13 oranında azalmıştır [14].

Ewees vd. yaptığı çalışmada, Aritmetik Optimizasyon Algoritması (AOA) ve Genetik Algoritma operatörleri birleştirilerek yüksek boyutlu veri setleri için AOAGA adlı bir yöntem sunulmuştur. Bu yöntem, AOA'nın arama stratejilerindeki zayıflıkları gidermek ve yerel optimumlardan kaçınmak için geliştirilmiştir. AOAGA, 20 UCI makine öğrenmesi deposundaki veri setleri ve iki gerçek dünya gen veri seti üzerinde test edilmiştir. Sonuçlar, AOAGA'nın diğer algoritmalarla kıyasla daha yüksek doğruluk, daha az öznelik ve daha iyi uygunluk değerleri sağladığını göstermiştir [15].

Maleki vd. yaptığı çalışmada, akciğer kanseri teşhisini geliştirmek amacıyla k-en yakın komşu algoritması, Genetik Algoritma ile birleştirilmiştir. Genetik Algoritma, veri setindeki gereksiz özellikleri eleyerek en iyi öznelik kombinasyonunu seçmiş ve k-en yakın komşu algoritmasının doğruluk oranını artırmıştır. Ayrıca, k parametresinin optimal değeri deneysel olarak belirlenmiştir. Sonuçlar, bu hibrit yöntemin sınıflandırma doğruluğunu %100'e kadar çıkardığını ve işlem süresini önemli ölçüde azalttığını göstermiştir. Bu, akciğer kanseri teşhisinde etkili bir çözüm sunmaktadır [16].

Donuk vd. yaptığı çalışmada, yüz ifadelerinden duygu tanıma işlemi için İkili Parçacık Sürü Optimizasyonu ve Destek Vektör Makineleri yöntemlerini kullanarak derin özellik seçimi tabanlı bir sistem geliştirmiştir. İlk aşamada, FER+ veri seti ile eğitilen bir Evrişimsel Sinir Ağı kullanılarak görüntülerden öznelikler çıkarılmıştır. Daha sonra, evrişimsel sinir ağından elde edilen öznelikler arasından en anlamlı olanları seçmek için ikili parçacık sürü optimizasyon algoritması uygulanmıştır. Son olarak, seçilen öznelikler destek vektör makineleri ile sınıflandırılmıştır. Önerilen sistem, doğruluk oranını %85,74'e çıkarmış ve işlem hızını artırarak, duygu tanıma performansına katkı sağladığını göstermiştir [17].

Luo vd. yaptığı çalışmada, büyük ölçekli veri setlerinde öznelik seçimi için Hibrit Kaba Hiperküp Yaklaşımı ve İkili Parçacık Sürü Optimizasyonu birleştirilerek yeni bir yöntem geliştirilmiştir. RH-BPSO olarak adlandırılan bu yöntem, özelliklerin anlamlılığını, bağımlılığını ve önemini dikkate alarak yüksek doğruluk oranları ve daha kompakt öznelik alt kümeleri elde etmeyi hedeflemiştir. Ayrıca, Apache Spark çerçevesinde paralel bir strateji uygulanarak DiRH-BPSO adı verilen dağıtılmış bir versiyon geliştirilmiş ve büyük veri kümeleri üzerinde ölçeklenebilirliği artırılmıştır. Yapılan deneysel çalışmalar, önerilen yöntemin sınıflandırma doğruluğu, öznelik alt küme boyutu ve hesaplama verimliliği açısından mevcut yaklaşımlardan daha üstün olduğunu göstermiştir [18].

3. Yöntem

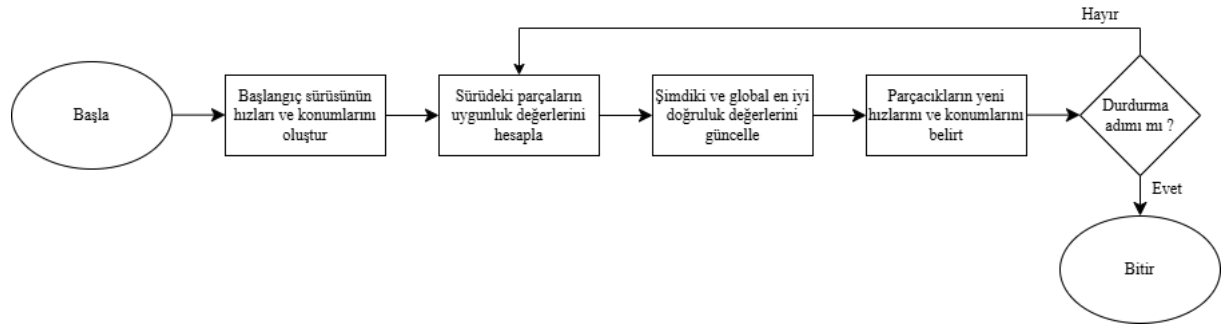
Bu bölümde, öncelikle öznitelik seçimi için kullanılan iki farklı metasezgisel algoritma olan İkili Parçacık Sürü Optimizasyon Algoritması ve Genetik Algoritmanın temel prensipleri tanıtılmaktadır. Ardından, sınıflandırma problemlerinin çözümünde yararlanılan dokuz farklı makine öğrenmesi algoritması kısaca açıklanmıştır. Son olarak, deneysel çalışmada kullanılan değerlendirici ölçütler ve çapraz doğrulama tekniğinin uygulama yöntemleri ele alınarak, model performansının nasıl kıyaslandığı ortaya konmuştur.

3.1 Metasezgisel algoritmalar

3.1.1 İkili parçacık sürü optimizasyon algoritması (BPSO)

PSO, Eberhart ve Kennedy [19] tarafından geliştirilen ve doğadaki sürü davranışlarını taklit eden popülasyon tabanlı bir optimizasyon algoritmasıdır. Bu yöntem, kuşların, balıkların veya böceklerin ortak hareket ederek en uygun çözümü bulma prensibine dayanır. PSO’da her bir birey, yani parçacık, arama uzayında rastgele başlatılır ve zamanla daha iyi çözümlere ulaşmak için kendi en iyi konumunu (*pbest*) ve tüm sürünün en iyi konumunu (*gbest*) referans alarak hareket eder. BPSO, özellikle özellik seçimi gibi ikili karar gerektiren problemlerde tercih edilir. Bu yöntemde parçacıkların konumları 0 veya 1 şeklinde güncellenerek belirli bir veri kümesinde hangi özelliklerin daha önemli olduğu tespit edilir [20].

BPSO’nun çalışma mekanizması, belirli adımlardan oluşan iteratif bir süreçtir. Bu adımlar Şekil 1’deki akış şeması ile görselleştirilmiştir.



Şekil 1. BPSO akış diyagramı.

BPSO’nun temel avantajı, yüksek boyutlu veri kümelerinde gereksiz özellikleri elemeye yardımcı olması ve hesaplama maliyetlerini azaltarak daha etkili bir model oluşturmasıdır. Çalışmamızda, BPSO farklı amaç fonksiyonlarıyla test edilerek çeşitli ML algoritmaları ile birlikte kullanılmış ve belirli bir problem için en uygun özellik alt kümesi belirlenmiştir.

3.1.2 Genetik algoritma (GA)

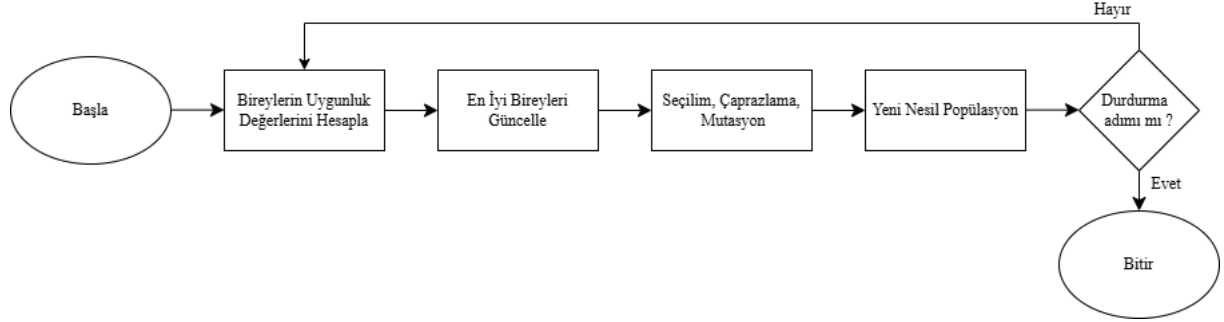
GA, evrimsel biyolojiden ilham alarak geliştirilmiş bir optimizasyon tekniğidir. Seçim, çaprazlama, mutasyon ve kalıtım gibi mekanizmaları kullanarak en iyi çözümü bulmayı hedefler. GA, özellikle yüksek boyutlu veri setlerinde en iyi öznitelik alt kümesini seçmek için kullanılan güçlü bir yöntemdir.

GA’nın temel yapısı, bireyler (kromozomlar) ve popülasyondan oluşur. Her birey, ikili kodlarla temsil edilen bir dizi özellikten oluşur, burada “1” ilgili özelliğin seçildiğini, “0” ise elendiğini gösterir. Genetik algoritma, popülasyondaki bireylerin uygunluklarını değerlendirerek en iyi olanları seçer ve nesilden nesile geliştirir.

İlk adımda, rastgele oluşturulan bireyler bir popülasyon halinde başlatılır ve her bireyin uygunluk değeri hesaplanır. Daha başarılı bireylerin yeni nesilde yer alma olasılığı artırılır. Seçim aşamasında, iyi bireylerin ebeveyn olarak seçilmesi için rulet tekerleği yöntemi gibi teknikler kullanılır.

Daha sonra, çaprazlama işlemi gerçekleştirilir, burada ebeveyn bireylerden alınan genetik bilgiler yeni bireylere aktarılır. Yeni bireyler oluşturulduktan sonra, mutasyon süreci uygulanır, yani rastgele genlerde küçük değişiklikler yapılarak çözüm çeşitliliği artırılır. Bu süreç belirlenen durdurma kriterine kadar devam eder [21].

GA'nın çalışma adımları Şekil 2'de verilen akış diyagramında gösterildiği gibi ilerlemektedir. Bu yapısı sayesinde, özellik seçimi ve optimizasyon problemlerinde etkili bir yöntem olarak öne çıkmaktadır.



Şekil 2. GA akış diyagramı.

3.2 Makine öğrenmesi algoritmaları

3.2.1 K-en yakın komşu algoritması (KNN)

KNN algoritması, bir veri noktasının sınıfını belirlemek için en yakın komşularının sınıflarına bakar. Bu algoritma genellikle öklid mesafesini kullanarak komşuların yakınlığını hesaplar. Basitliği ve yüksek doğruluk potansiyeli nedeniyle sınıflandırma problemlerinde sıklıkla kullanılır [22].

3.2.2 Karar ağaçları (DT)

Karar Ağaçları, verileri dallar ve yapraklar şeklinde yapılandırarak sıralı kararlarla sınıflandırma yapar. Kolay yorumlanabilir olması nedeniyle yaygın olarak tercih edilir. Her bir düğüm, veri özelliklerine göre bölünmüş bir alt küme temsil eder ve algoritma bu şekilde homojen gruplar oluşturmayı hedefler [22].

3.2.3 Destek vektör makinesi (SVM)

SVM, veri noktalarını sınıflar arasında ayırmak için en iyi hiper düzlemi bulan güçlü bir algoritmadır. Marjinal boşluğu maksimize ederek sınıflandırma yapar. Özellikle doğrusal olmayan problemlerde çekirdek fonksiyonları ile etkili sonuçlar sağlar [22].

3.2.4 Lojistik regresyon (LR)

Lojistik Regresyon, sınıflandırma problemlerinde kullanılan, kategorik bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkiyi modellemek için geliştirilmiş parametrik bir yöntemdir. Bu algoritma, veriyi Sigmoid fonksiyonu üzerinden sınıflandırır ve çıktı olasılığını 0 ile 1 arasında bir değerde ifade eder. LR, özellikle çoklu doğrusal bağlantı durumunda etkilenebilir, bu nedenle veri setindeki bağımsız değişkenlerin ilişki durumu kontrol edilmelidir [23].

3.2.5 Naive bayes (NB)

Naive Bayes algoritması, Bayes teoremi üzerine kurulu olup, sınıflandırma problemlerinde genellikle veri setindeki özellikler arasında bağımsızlık varsayımı yaparak çalışır. Bu algoritma, koşullu olasılıklara dayanarak yeni bir verinin ait olduğu sınıfı belirler. Naive Bayes, hızlı ve hafif bir yöntem olup genellikle metin sınıflandırma ve duygu analizi gibi uygulamalarda kullanılır. Ancak, bu algoritmanın performansı, özellikle yüksek boyutlu veri setlerinde doğrusal bağlantı veya aşırı veri dengesizliği durumunda düşebilir [23].

3.2.6 Aşırı gradyan artırma (XGBoost)

XGBoost, karar ağaçları ve gradyan artırma yöntemine dayanan, performans odaklı bir sınıflandırma algoritmasıdır. Yüksek doğruluk ve hız sağlamak için düzenleme yöntemlerini entegre eden bu algoritma, büyük ölçekli veri setlerinde etkili sonuçlar verir. XGBoost, özellikle kayıp fonksiyonunu optimize etmek için Taylor serisinin ikinci derecesini kullanarak öğrenme sürecini hızlandırır. Bu algoritma, doğruluk ve F1-ölçütü gibi metriklerde diğer algoritmalarından genellikle üstün performans göstermektedir [23].

3.2.7 Yapay sinir ağları (ANN)

Yapay Sinir Ağları (ANN), insan beynindeki öğrenme ve karar verme süreçlerini modelleyen matematiksel bir algoritmadır. ANN'ler genellikle üç temel katmandan oluşur: giriş katmanı, gizli katmanlar ve çıkış katmanı. Giriş katmanı, modelin işlediği ham veriyi alır, gizli katmanlar bu veriyi işleyerek örüntüleri tanımlar ve çıkış katmanı nihai sonucu üretir. Her bir nöron, kendisine bağlı diğer nöronlardan gelen bilgileri alır ve aktivasyon fonksiyonları kullanarak işleme tabi tutar. Bu işlem, sistemin doğrusal olmayan ilişkileri öğrenmesini ve karmaşık veri örüntülerini tanımasını sağlar [24].

3.2.8 Evrimsel sinir ağları (CNN)

CNN, görüntü tabanlı verilerle çalışan bir derin öğrenme modelidir. Ancak, bu çalışmada bir boyutlu veri ile çalışıldığından, 1D-CNN mimarisi tercih edilmiştir. CNN, temel olarak evrim katmanları, havuzlama katmanları ve tamamen bağlı katmanlardan oluşur. Evrim katmanları, giriş verisindeki önemli özellikleri filtrelemek ve çıkartmak için evrimsel filtreler kullanır. Havuzlama katmanları ise veri boyutunu azaltarak modelin hesaplama yükünü hafifletir ve öğrenmeyi hızlandırır. Modelin sonunda, sınıflandırma için bir veya daha fazla tamamen bağlı katman yer alır [25].

3.2.9 Uzun kısa süreli bellek ağları (LSTM)

LSTM, tekrarlayan sinir ağlarının (RNN) bir türevidir ve uzun vadeli bağımlılıkları öğrenmede etkili bir yaklaşımdır. LSTM, giriş verileri arasındaki zaman bağımlılıklarını işlemek için “unutma kapısı”, “giriş kapısı” ve “çıkış kapısı” gibi özel mekanizmalara sahiptir. Bu mekanizmalar, geçmiş bilgilerin hücre durumunda tutulmasını ve gerekli bilgilerle güncellenmesini sağlar. LSTM'ler özellikle sıralı verilerin analizinde güçlü bir araç olarak karşımıza çıkmaktadır [25,26].

3.3 Değerlendirme metrikleri

ML algoritmalarının performansını değerlendirmek için bu çalışmada doğruluk, kesinlik, duyarlılık ve F1-ölçütü gibi temel metrikler kullanılmıştır. Ayrıca modelin genelleştirilebilirliğini ölçmek için k-kat çapraz doğrulama yöntemi uygulanmıştır. Bu metriklerin hesaplanmasında, sınıflandırma sonuçlarını görselleştirmek ve yorumlamak için karmaşıklık matrisinden yararlanılmıştır.

Karmaşıklık matrisi, bir sınıflandırma modelinin tahmin sonuçlarını görselleştiren bir tablodur. Bu matris, doğru pozitif (TP), doğru negatif (TN), yanlış pozitif (FP) ve yanlış negatif (FN) olmak üzere dört temel bileşeni içerir. Bu bileşenler, sınıflandırma performansını değerlendirmek için kullanılan metriklerin temelini oluşturur:

- TP: Modelin doğru şekilde pozitif sınıfa atadığı örnekler.
- TN: Modelin doğru şekilde negatif sınıfa atadığı örnekler.
- FP: Modelin yanlış şekilde pozitif sınıfa atadığı örnekler.
- FN: Modelin yanlış şekilde negatif sınıfa atadığı örnekler.

Karmaşıklık matrisi, sınıflandırma sonuçlarının detaylı bir analizi için vazgeçilmez bir araçtır ve özellikle sınıf dengesizliklerinin olduğu veri setlerinde metriklerin yorumlanmasını kolaylaştırır [27].

3.3.1 Doğruluk

Doğruluk, modelin toplam doğru tahmin oranını ölçen bir metriktir. Özellikle sınıf dengesi sağlanmış veri setlerinde etkili bir ölçüm olarak kullanılır ve Denklem 1 ile hesaplanır.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Bu oran, modelin genel başarısını değerlendirirken temel bir referans noktası sağlar.

3.3.2 Kesinlik

Kesinlik, modelin pozitif tahminlerinin ne kadarının doğru olduğunu ölçen Denklem 2 ile hesaplanan bir metriktir. Yanlış pozitiflerin (FP) az olduğu durumlarda yüksek kesinlik değerine ulaşılır.

$$Kesinlik = \frac{TP}{TP + FP} \quad (2)$$

3.3.3 Duyarlılık

Duyarlılık, Denklem 3'te verildiği gibi modelin pozitif sınıfı doğru bir şekilde tanıma oranını ifade eder. Yanlış negatiflerin (FN) etkisini ölçerek, modelin hassasiyetini ortaya koyar:

$$Duyarlılık = \frac{TP}{TP + FN} \quad (3)$$

3.3.4 F1-ölçütü

F1-ölçütü, Denklem 4 ile hesaplanan ve kesinlik ile hassasiyet arasındaki dengeyi ölçmek için kullanılan harmonik bir ortalamadır. Özellikle sınıf dengesizliğinin olduğu veri setlerinde kullanışlıdır:

$$F1 - \text{ölçütü} = 2 \times \frac{Kesinlik \times Hassasiyet}{Kesinlik + Hassasiyet} \quad (4)$$

3.3.5 Çapraz doğrulama

Çapraz doğrulama veri setini k eşit parçaya ayırır ve her bir parça sırasıyla test verisi olarak kullanılırken diğer parçalar eğitim için kullanılır. Bu süreç k kez tekrarlanır ve sonuçlar ortalaması alınarak modelin performansı değerlendirilir. Çapraz doğrulama, veri setinin küçük olduğu durumlarda ezberleme (aşırı öğrenme) riskini azaltmada etkili bir yöntemdir [28].

4. Deneysel Çalışma

Bu çalışmada gerçekleştirilen tüm deneyler, Python programlama dili kullanılarak Jupyter Notebook ortamında geliştirilmiş ve iki farklı bilgisayar sistemi üzerinde yürütülmüştür. Kullanılan veri setleri, sınıflandırma modellerinin başarısını değerlendirmek amacıyla 0,1 ile 0,9 aralığında normalize edilmiştir. Bu normalizasyon işlemi MinMaxScaler yöntemi ile gerçekleştirilirken, çıkış sınıfları 0 ile 1 arasında normalize edilmiş ve LabelEncoder kullanılmıştır. Meme Kanseri veri setinde, anlamlı bilgi taşımayan ve model performansına katkı sağlamayan iki sütun veri setinden kaldırılmıştır. Bu işlem, veri setinin daha odaklı hale getirilmesi ve gereksiz özniteliklerin etkisinin minimize edilmesi amacıyla gerçekleştirilmiştir.

Her iki veri seti için öznitelik seçimi ve model değerlendirme süreçleri aşağıdaki adımlar çerçevesinde yürütülmüştür:

- Öznitelik Seçimi: Metasezgisel algoritmalar olan BPSO ve GA, 1000 iterasyon boyunca 10 kez çalıştırılarak öznitelik seçimi yapılmıştır. Her iki algoritma, 20 ve 30 popülasyon büyüklükleri ile üç farklı amaç fonksiyonu (SVM, KNN, DT) kullanılarak optimize edilmiştir.
- Çapraz Doğrulama: Seçilen öznitelikler ve tüm veri seti, 10 katlı çapraz doğrulama yöntemi ile değerlendirilmiştir. Bu yöntem, modelin genelleştirilebilirliğini ve doğruluğunu ölçmek için veri setini farklı alt kümelerle ayırarak kullanılmıştır. Her çapraz doğrulama döngüsü 10 kez tekrarlanmış ve sonuçlar ortalama alınarak raporlanmıştır.

• Makine Öğrenmesi Modelleri: Öznitelik seçimi sonrası elde edilen alt kümeler, hem kendi amaç fonksiyonları (KNN, SVM, DT) ile hem de farklı makine öğrenmesi algoritmalarına (LR, NB, XGBoost, ANN, CNN ve LSTM) uygulanmış ve her algoritmanın doğruluk, kesinlik, duyarlılık ve F1-ölçütü metrikleri hesaplanmıştır.

ML modelleri ve derin öğrenme mimarileri, Python programlama dili ve ilgili kütüphaneler kullanılarak uygulanmıştır. KNN, SVM, DT, LR ve NB algoritmaları için Scikit-learn kütüphanesi tercih edilmiş ve varsayılan hiperparametre değerleri kullanılmıştır. XGBoost modeli için XGBoost kütüphanesi, derin öğrenme modelleri olan ANN, CNN ve LSTM için ise TensorFlow ve Keras kütüphaneleri kullanılmıştır. Derin öğrenme modellerinde gizli katmanlarda sırasıyla 32-64-64-32 nöronlar ve ReLU aktivasyon fonksiyonu, çıkış katmanında ise 1 nöron ve Sigmoid aktivasyon fonksiyonu kullanılmıştır. Eğitim süreci 32 batch boyutu ve 50 epoch ile gerçekleştirilmiş olup, tüm modellerin performansları doğruluk, kesinlik, duyarlılık ve F1-ölçütü metrikleri ile değerlendirilmiştir.

4.1 Sonar veri seti

Sonar veri seti [29], deniz tabanına gönderilen sonar sinyallerini kullanarak mineral ve kaya türlerinin sınıflandırılmasını amaçlayan 208 örnekten oluşmaktadır. Veri seti 60 öznitelige sahip olup, iki sınıfa (%54 kaya, %46 maden) ayrılmıştır. Sonar veri setine ait özellikler Tablo 1’de listelenmiştir.

Tablo 1. Sonar veri setinin özellikleri.

Veri Seti	Öznitelik Sayısı	Sınıf Sayısı	Sınıf Dağılımı	Örnek Sayısı
Sonar	60	2	%54, %46	208

Sonar veri seti üzerinde gerçekleştirilen deneylerde, tüm özniteliklerin kullanıldığı durumlarda SVM algoritması en yüksek doğruluk oranını (0,8500) sergilemiştir (Tablo 2). BPSO ve GA algoritmaları kullanılarak gerçekleştirilen öznitelik seçimi sonuçları ise Tablo 3 ve Tablo 4’te detaylandırılmıştır.

Tablo 2. Sonar veri setinin farklı algoritmalar ile doğrulukları.

Algoritma	Öznitelik Sayısı	Doğruluk
KNN	60	0,8310
SVM	60	0,8500
DT	60	0,7043

Veri seti üzerinde farklı popülasyon ve farklı amaç fonksiyonlarıyla yapılan 10 tekrar 1000 iterasyon sonuçlarının ortalamaları Tablo 3’te gösterildiği gibidir. Tablodaki SS standart sapmayı göstermektedir.

Tablo 3. Sonar veri seti için öznitelik seçim ortalamaları.

Algoritma	Amaç Fonksiyonu	Popülasyon	En iyi	En Kötü	Ortalama	SS	Ortalama Öznitelik Sayısı
BPSO	SVM	20	0,9273	0,8980	0,9066	0,0084	28,3
BPSO	SVM	30	0,9226	0,8983	0,9105	0,0069	30,8
BPSO	KNN	20	0,9130	0,8985	0,9048	0,0053	29,1
BPSO	KNN	30	0,9085	0,8942	0,9029	0,0053	30,5
BPSO	DT	20	0,8459	0,8307	0,8392	0,0046	25,9
BPSO	DT	30	0,8645	0,8300	0,8414	0,0103	26,4
GA	SVM	20	0,8890	0,8690	0,8810	0,0058	29,6
GA	SVM	30	0,8983	0,8792	0,8888	0,0058	30,3
GA	KNN	20	0,8983	0,8695	0,8847	0,0097	27,2
GA	KNN	30	0,9033	0,8745	0,8869	0,0096	32,2
GA	DT	20	0,8640	0,8066	0,8280	0,0172	29,3
GA	DT	30	0,8461	0,8166	0,8330	0,0103	30,0

Sonar veri seti için gerçekleştirilen deneylerde, Tablo 4’te yer alan sonuçlar, Tablo 3’teki öznitelik seçim ortalamalarına göre en yüksek doğruluk oranlarına sahip kombinasyonlar seçilerek belirlenmiştir. Bu sonuçlar, BPSO algoritması ile SVM amaç fonksiyonu kullanıldığında 0,9273 doğruluk oranına ve 30 öznitelik sayısına,

GA algoritması ile KNN amaç fonksiyonu kullanıldığında ise 0,9033 doğruluk oranına ve 30 öznitelik sayısına ulaşıldığını göstermektedir. Tablo 4'te elde edilen yeni öznitelikler ile Tablo 5 ve Şekil 3'teki algoritmalar denenmiştir. Bu bulgular, her iki metasezgisel algoritmanın öznitelik sayısını azaltırken model doğruluğunu artırmada etkili olduğunu ortaya koymaktadır.

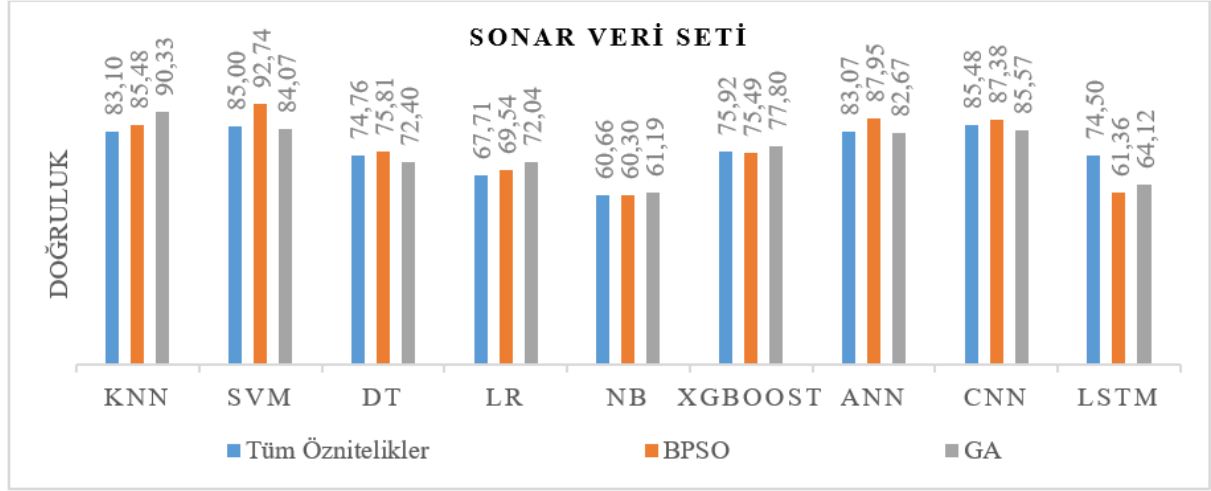
Tablo 4. Sonar veri seti için en iyi öznitelikler ve doğrulukları.

Algoritma	Popülasyon Sayısı	Amaç Fonksiyonu	Doğruluk	Öznitelik Sayısı
BPSO	20	SVM	0,9273	30
BPSO	30	SVM	0,9226	31
GA	20	KNN	0,8983	25
GA	30	KNN	0,9033	30

Tablo 5. Sonar veri seti için makine öğrenmesi algoritmalarının performansları.

Algoritma	Öznitelik	Doğruluk	Kesinlik	Duyarlılık	F1-Ölçütü
LR	Normal(60)	0,6771	0,6934	0,6771	0,6571
	BPSO(30)	0,6954	0,7052	0,6954	0,6820
	GA(30)	0,7204	0,7415	0,7204	0,7052
NB	Normal(60)	0,6066	0,6470	0,6066	0,5857
	BPSO(30)	0,6030	0,6272	0,6030	0,5840
	GA(30)	0,6119	0,6422	0,6119	0,5979
XGBoost	Normal(60)	0,7592	0,7904	0,7592	0,7497
	BPSO(30)	0,7549	0,7742	0,7549	0,7494
	GA(30)	0,7780	0,8082	0,7780	0,7650
ANN	Normal(60)	0,8307	0,8691	0,7589	0,7971
	BPSO(30)	0,8795	0,8900	0,8544	0,8698
	GA(30)	0,8267	0,8349	0,8022	0,8108
CNN	Normal(60)	0,8548	0,8530	0,8422	0,8445
	BPSO(30)	0,8738	0,8878	0,8422	0,8607
	GA(30)	0,8357	0,8545	0,7989	0,8189
LSTM	Normal(60)	0,7450	0,7905	0,6467	0,6870
	BPSO(30)	0,6136	0,6537	0,2167	0,2967
	GA(30)	0,6412	0,5189	0,4033	0,4253

Tablo 5'te Sonar veri seti üzerinde gerçekleştirilen deneylerde, farklı makine öğrenmesi algoritmalarının 60 öznitelik ve BPSO ve GA tarafından seçilen 30 öznitelik üzerinden elde edilen doğruluk, kesinlik, duyarlılık ve F1-ölçütü değerleri karşılaştırılmıştır. LR modelinde, normal durumda 0,6771 doğruluk oranı elde edilirken, BPSO ile 0,6954 ve GA ile 0,7204 oranına ulaşılması, öznitelik sayısının azaltılmasıyla gereksiz bilgi ve gürültünün elenmesinin modele olumlu yansıdığını göstermektedir. Buna karşın, NB modelinde normal durumda 0,6066 doğruluk oranı elde edilirken, BPSO ile 0,6030 ve GA ile gerçekleştirilen öznitelik seçimi sonrası 0,6119 doğruluk oranı elde edilmiştir. XGBoost modelinde ise normal durumda 0,7592 doğruluk elde edilirken, GA ile öznitelik seçimi sonrasında 0,7780 doğruluk oranına ulaşılması, gereksiz özniteliklerin temizlenmesinin model performansına pozitif katkı sağladığını göstermektedir. ANN modeli, normal öznitelik setiyle 0,8307 doğruluk oranına ulaşırken, BPSO ile seçilen öznitelik setinde 0,8795 ile en yüksek doğruluk oranına çıkmış ve GA ile elde edilen öznitelik setinde 0,8267 doğruluk değeri gözlemlenmiştir. CNN modelinde, normal durumda 0,8548 doğruluk oranı elde edilirken, BPSO uygulaması ile 0,8738 oranına ulaşmıştır. Buna karşın, LSTM modelinde normal öznitelik kullanımı 0,7450 doğruluk verirken, öznitelik seçimi sonrası BPSO ile 0,6136 ve GA ile 0,6412 doğruluk değeri kaydedilmiştir.



Şekil 3. Sonar veri setinin tüm algoritmalar ile kıyaslanması.

4.2 Meme Kanseri veri seti

Meme Kanseri veri seti [30], hastaların tıbbi görüntüleme sonuçlarını içeren ve meme kanseri teşhisi için kullanılan 569 örnekten oluşmaktadır. Veri seti, 30 öznitelikle iki sınıfa (%63 kanserli, %37 kanser olmayan) ayrılmıştır. Meme Kanseri veri setine ait özellikler Tablo 6’da verilmiştir.

Tablo 6. Meme kanseri veri setinin açıklanması.

Veri Seti	Öznitelik Sayısı	Sınıf Sayısı	Sınıf Dağılımı	Örnek Sayısı
Meme Kanseri	30	2	%63, %37	569

Meme Kanseri veri seti üzerinde gerçekleştirilen deneylerde, tüm özniteliklerin kullanıldığı durumlarda SVM algoritması en yüksek doğruluk oranını (0,9754) sergilemiştir (Tablo 7). Öznitelik seçimi sonuçları Tablo 8 ve Tablo 9’da sunulmuştur.

Tablo 7. Meme kanseri veri setinin farklı algoritmalar ile doğrulukları.

Algoritma	Öznitelik Sayısı	Doğruluk
KNN	30	0,9666
SVM	30	0,9754
DT	30	0,9209

Veri seti üzerinde farklı popülasyon ve farklı amaç fonksiyonlarıyla yapılan 10 tekrar 1000 iterasyon sonuçlarının ortalamaları Tablo 8’de gösterildiği gibidir. Tablodaki SS standart sapmayı göstermektedir.

Meme Kanseri veri seti için yapılan öznitelik seçimi sonuçlarında, Tablo 9’da yer alan değerler, Tablo 8’deki öznitelik seçim ortalamalarına göre en yüksek doğruluk oranlarına sahip kombinasyonlar seçilerek belirlenmiştir. BPSO algoritması ile SVM amaç fonksiyonu kullanıldığında 0,9842 doğruluk oranına ve 14 öznitelik sayısına, GA algoritması ile SVM amaç fonksiyonu kullanıldığında ise 0,9824 doğruluk oranına ve 13 öznitelik sayısına ulaşılmıştır. Tablo 9’da elde edilen yeni öznitelikler ile Tablo 10 ve Şekil 4’teki algoritmalar denenmiştir. Bu sonuçlar, BPSO ve GA algoritmalarının öznitelik sayısını azaltırken model doğruluğunu artırmada etkili olduğunu göstermektedir.

Tablo 8. Meme kanseri veri seti için öznitelik seçim sonuçları.

Algoritma	Amaç Fonksiyonu	Popülasyon	En iyi	En Kötü	Ortalama	SS	Ortalama Öznitelik Sayısı
BPSO	SVM	20	0,9842	0,9807	0,9821	0,0011	16,0
BPSO	SVM	30	0,9842	0,9807	0,9824	0,0014	16,0
BPSO	KNN	20	0,9789	0,9771	0,9778	0,0009	16,7
BPSO	KNN	30	0,9806	0,9771	0,9787	0,0009	16,9
BPSO	DT	20	0,9630	0,9560	0,9594	0,0023	12,9
BPSO	DT	30	0,9631	0,9561	0,9603	0,0024	13,1
GA	SVM	20	0,9824	0,9771	0,9796	0,0014	15,9
GA	SVM	30	0,9824	0,9789	0,9803	0,0011	16,1
GA	KNN	20	0,9789	0,9754	0,9771	0,0008	15,6
GA	KNN	30	0,9789	0,9754	0,9771	0,0011	14,7
GA	DT	20	0,9613	0,9543	0,9578	0,0024	12,9
GA	DT	30	0,9648	0,9543	0,9587	0,0035	14,0

Tablo 9. Meme kanseri veri seti için en iyi öznitelikler ve doğrulukları.

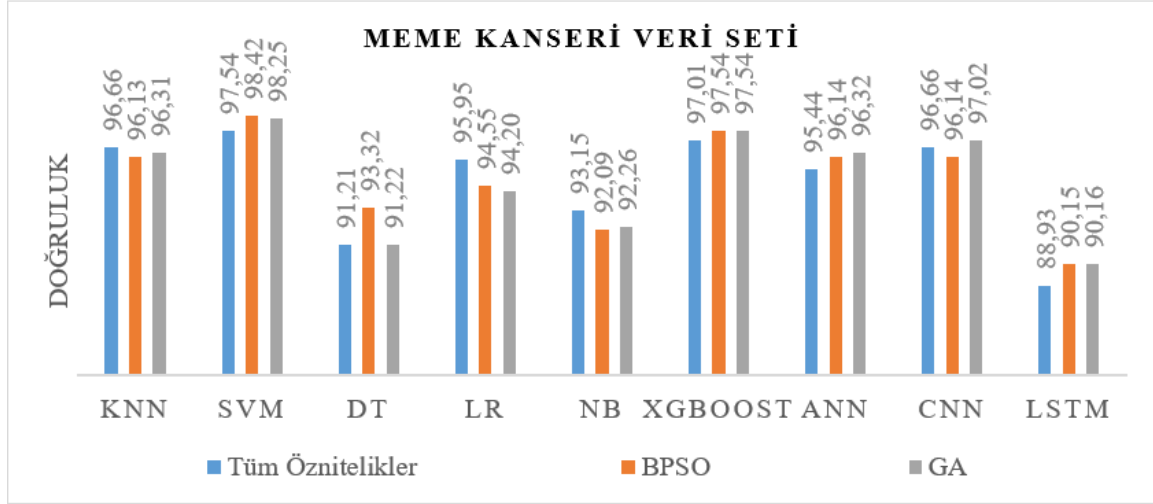
Algoritma	Popülasyon Sayısı	Amaç Fonksiyonu	Doğruluk	Öznitelik Sayısı
BPSO	20	SVM	0,9842	17
BPSO	30	SVM	0,9842	14
GA	20	SVM	0,9824	13
GA	30	SVM	0,9824	18

Tablo 10. Meme kanseri veri seti için makine öğrenmesi algoritmalarının performansları.

Algoritma	Öznitelik	Doğruluk	Kesinlik	Duyarlılık	F1-Ölçütü
LR	Normal(30)	0,9595	0,9614	0,9595	0,9590
	BPSO(14)	0,9455	0,9491	0,9455	0,9444
	GA(13)	0,9420	0,9454	0,9420	0,9408
NB	Normal(30)	0,9315	0,9343	0,9315	0,9312
	BPSO(14)	0,9209	0,9223	0,9209	0,9204
	GA(13)	0,9226	0,9235	0,9226	0,9222
XGBoost	Normal(30)	0,9701	0,9706	0,9701	0,9701
	BPSO(14)	0,9754	0,9759	0,9754	0,9752
	GA(13)	0,9754	0,9759	0,9754	0,9752
ANN	Normal(30)	0,9544	0,9598	0,9245	0,9375
	BPSO(14)	0,9614	0,9520	0,9481	0,9484
	GA(13)	0,9632	0,9688	0,9340	0,9488
CNN	Normal(30)	0,9666	0,9705	0,9433	0,9547
	BPSO(14)	0,9614	0,9631	0,9383	0,9473
	GA(13)	0,9702	0,9735	0,9478	0,9588
LSTM	Normal(30)	0,8893	0,8433	0,8955	0,8564
	BPSO(14)	0,9015	0,8936	0,8433	0,8613
	GA(13)	0,9016	0,8948	0,8392	0,8637

Meme Kanseri veri seti üzerinde yapılan deneysel çalışmalarda ise, 30 öznitelik, BPSO ile seçilen 14 öznitelik ve GA ile seçilen 13 öznitelik kullanılarak gerçekleştirilen deneyler sonucunda elde edilen doğruluk, kesinlik, duyarlılık ve F1-ölçütü değerleri karşılaştırılmıştır. LR modelinde normal öznitelik kullanıldığında 0,9595 doğruluk elde edilirken, öznitelik seçimi sonrasında BPSO ve GA ile sırasıyla 0,9455 ve 0,9420 doğruluk değerleri gözlemlenmiştir. NB modelinde ise normal durumda 0,9315 doğruluk oranı kaydedilirken, BPSO ve GA uygulamaları sonrası doğruluk değerleri sırasıyla 0,9209 ve 0,9226'dır. XGBoost modelinde normal öznitelik setiyle elde edilen 0,9701 doğruluk oranının, öznitelik seçimi sonrasında 0,9754'e yükselmesi en yüksek doğruluğa ulaşmıştır, gereksiz veya gürültülü özniteliklerin elenmesinin modelin karar verme mekanizmasını güçlendirdiğini ortaya koymaktadır. ANN modelinde normal öznitelik kullanımı 0,9544 doğruluk oranı verirken, BPSO ve GA ile

seçilen öznitelik setlerinde sırasıyla 0,9614 ve 0,9632 doğruluk oranlarına ulaşılmıştır. CNN modelinde ise normal durumda 0,9666 doğruluk oranı elde edilirken, GA uygulaması sonrası 0,9702 oranı gözlemlenmiştir. LSTM modelinde normal öznitelik setiyle 0,8893 doğruluk oranı ile en düşük doğruluk elde edilirken, öznitelik seçimi uygulamalarında BPSO ile 0,9015 doğruluk oranı elde edilmiştir.



Şekil 4. Meme kanseri veri setinin tüm algoritmalar ile kıyaslanması.

Şekil 4, Meme Kanseri veri setinde öznitelik seçimi öncesi ve sonrası farklı ML algoritmalarının doğruluk oranlarını görsel olarak karşılaştırmaktadır. BPSO ve GA ile öznitelik seçimi yapıldıktan sonra, özellikle LR, XGBoost, ANN ve CNN modellerinde doğruluk oranlarında belirgin artışlar gözlemlenmektedir. Bu bulgular, öznitelik seçiminin ML algoritmalarının performansını iyileştirmede etkili olduğunu ortaya koymaktadır.

5. Sonuç

Bu çalışmada, öznitelik seçimi için metasezgisel algoritmalar olan BPSO ve GA yöntemleri kullanılarak Sonar ve Meme Kanseri veri setlerinde gerçekleştirilen deneyler incelenmiştir. Elde edilen sonuçlar, öznitelik seçiminin model performansı üzerindeki olumlu etkilerini açıkça ortaya koymaktadır.

Sonar veri seti üzerinde yapılan deneylerde, BPSO ve GA algoritmaları ile gerçekleştirilen öznitelik seçimi süreçleri, temel sınıflandırma algoritmalarının doğruluk oranlarında anlamlı artışlar sağlamıştır. Özellikle BPSO algoritması ile SVM amaç fonksiyonu kullanılarak seçilen 30 öznitelik, doğruluk oranını %85,00'den %92,73'e yükseltmiştir. GA algoritması ile KNN amaç fonksiyonu kullanılarak seçilen 30 öznitelik ise %90,33 doğruluk oranına ulaşmıştır. Bu iyileşmeler, metasezgisel algoritmaların öznitelik sayısını azaltırken model doğruluğunu artırmada ne kadar etkili olduğunu göstermektedir. Ayrıca, oluşturulan doğrulama grafiklerinde, öznitelik seçiminin ML algoritmalarının performansını nasıl iyileştirdiği görsel olarak desteklenmiştir.

Meme Kanseri veri seti üzerinde gerçekleştirilen deneylerde ise, BPSO ve GA algoritmaları ile yapılan öznitelik seçimi süreçleri, sınıflandırma modellerinin doğruluk oranlarında da artışlar sağlamıştır. BPSO algoritması ile SVM amaç fonksiyonu kullanılarak seçilen 14 öznitelik, doğruluk oranını %97,54'ten %98,42'ye artırmıştır. GA algoritması ile SVM amaç fonksiyonu kullanılarak seçilen 13 öznitelik ise %98,24 doğruluk oranına ulaşmıştır. Bu sonuçlar, öznitelik seçiminin model performansını iyileştirebileceğini göstermektedir. Meme Kanseri veri seti için oluşturulan doğrulama grafikleri, öznitelik seçiminin ML algoritmalarının doğruluk oranları üzerindeki etkisini net bir şekilde ortaya koymaktadır.

Genel olarak, BPSO ve GA algoritmaları her iki veri setinde de öznitelik sayısını azaltırken model doğruluğunu artırmada başarılı olmuştur. Popülasyon büyüklüğünün (20 ve 30) performans üzerindeki etkisini göstermiştir. Her iki büyüklükte de farklı sonuçlar elde edilmiştir. Ayrıca, farklı ML algoritmalarının performans değerlendirmeleri, öznitelik seçiminin bazı modeller üzerinde daha olumlu etkiler yarattığını, bazı modellerde ise negatif etkiler gözlemlendiğini ortaya koymuştur. Bu durum, öznitelik seçiminin modelinin amaç fonksiyonuna ve veri setinin özelliklerine bağlı olarak farklı sonuçlar doğurabileceğini göstermektedir.

Sonuç olarak, bu çalışma metasezgisel öznitelik seçimi yöntemlerinin ML modellerinin performansını iyileştirmede etkin olduğunu doğrulamış olup, özellikle BPSO ve GA algoritmalarının öznitelik sayısını azaltırken doğruluk oranlarını artırmada başarılı olduğunu göstermektedir. Gelecek çalışmalarda, gerçek hayattaki problemlere yönelik büyük ve karmaşık veri setleri kullanılarak bu yöntemlerin uygulanması hedeflenmektedir. Sağlık, finans ve pazarlama gibi alanlarda karşılaşılan gerçek dünya verileri üzerinde öznitelik seçimi süreçlerinin optimize edilmesi, bu alanlarda daha doğru ve güvenilir tahmin modellerinin geliştirilmesine katkı sağlayabilir. Ek olarak, optimizasyon algoritmalarının paralel, çok amaçlı, ikili versiyonları önerilebilir, diğer algoritmalar ile hibrit versiyonları geliştirilerek özellik seçimi problemlerine uygulanabilir. Ayrıca, öznitelik seçimi sonrası modellerin hiperparametre optimizasyonu ile birleşik yaklaşımlar incelenerek daha yüksek performans elde edilmesi hedeflenebilir. Bu tür entegrasyonlar, gerçek hayat problemlerine yönelik çözümlerin etkinliğini artırarak, pratik uygulamalarda ML modellerinin kullanılabilirliğini ve başarısını artıracaktır.

Kaynaklar

- [1] Akinola OO, Ezugwu AE, Agushaka JO, Zitar RA, Abualigah L. Multiclass feature selection with metaheuristic optimization algorithms: a review. *Neural Comput Appl* 2022; 34(22): 19751-19790.
- [2] Kaur S, Kumar Y, Koul A, Kamboj SK. A systematic review on metaheuristic optimization techniques for feature selections in disease diagnosis: open issues and challenges. *Arch Comput Methods Eng* 2023; 30(3): 1863-1895.
- [3] El-Kenawy EM, Mirjalili S, Alassery F, Zhang Y, Eid MM, El-Mashad SY, Aloyaydi BA, Ibrahim A ve diğerleri. Novel meta-heuristic algorithm for feature selection, unconstrained functions and engineering problems. *IEEE Access* 2022; 10: 40536-40555.
- [4] Nssibi M, Manita G, Korbaa O. Advances in nature-inspired metaheuristic optimization for feature selection problem: A comprehensive survey. *Comput Sci Rev.* 2023; 49: 100559.
- [5] Sharma M, Kaur P. A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem. *Arch Comput Methods Eng* 2021; 28: 1103-1127.
- [6] Kumar L, Bharti KK. A novel hybrid BPSO-SCA approach for feature selection. *Natural Computing* 2021; 20(1): 39-61.
- [7] Yılmaz Ö, Gumustas E, Pehlivanlı AÇ. A Hybrid Metaheuristic based Feature Selection Framework for In-silico Mutagenicity Prediction. *Journal of Intelligent Systems: Theory and Applications* 2024; 7(2): 116-128.
- [8] Efeoğlu E. An Artificial Intelligence-Based Hybrid Approach to Detect the Type of Buried Objects with Broad Frequency Band Antenna Systems. *Firat University Journal of Experimental and Computational Engineering* 2024; 3(3): 362-378.
- [9] Gedikli E, Yurdusever T. EEG Sinyallerinden Meta-Sezgisel Optimizasyon Algoritmalarına Dayalı Özellik Seçimi. *Black Sea Journal of Engineering and Science* 2024; 7(4): 717-723.
- [10] Açar İ, Aydılek İB. Geleneksel Makine Öğrenmesi Yöntemleri e Metasezgisel Yöntemlerle Öznitelik Seçim Yöntemlerinin Karşılaştırılması. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi* 2024; 15(2): 397-409.
- [11] Cıran A, Özbay E. Derin Öğrenme ve Özellik Seçimi Yaklaşımları Kullanılarak Göz Hastalıkları Tespiti. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi* 2024; 15(2): 421-433.
- [12] Gümüş B, Çoban M, Tezcan SS. Metasezgisel Tabanlı Özellik Seçim Algoritmalarının Güç Kalitesi Olaylarının Sınıflandırılmasına Etkisinin İncelenmesi. *IJERD* 2024; 16(2): 646-658.
- [13] Mishra AK, Roy P, Bandyopadhyay S. Binary particle swarm optimization based feature selection (bpsi-fs) for improving breast cancer prediction. In *Proceedings of International Conference on Artificial Intelligence and Applications* 2021; 2020: 373-384.
- [14] Abualigah L, Dulaimi AJ. A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm. *Clust Comput* 2021; 24 2161-2176.
- [15] Ewees AA, Al-qaness MAA, Abualigah L, Oliva D, Algarni ZY, Anter AM, Ibrahim RA, Ghoniem RM ve diğerleri. Boosting arithmetic optimization algorithm with genetic algorithm operators for feature selection: case study on cox proportional hazards model. *Mathematics* 2021; 9(18): 2321.
- [16] Maleki N, Zeinali Y, Niaki STA. A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. *Expert Syst Appl* 2021; 164: 113981.
- [17] Donuk K, Arı A, Özdemir MF, Hanbay D. Deep feature selection for facial emotion recognition based on BPSO and SVM. *Politeknik Dergisi* 2023; 26(1): 131-142.
- [18] Luo C, Wang S, Li T, Chen H, Lv J, Yi Z. Large-scale meta-heuristic feature selection based on BPSO assisted rough hypercuboid approach. *IEEE Trans Neural Netw Learn* 2022; 34(12): 10889-10903.
- [19] Kennedy J, Eberhart R. Particle swarm optimization. In: *Proceedings of IEEE international conference on neural networks* 1995; 4: 1942-1948.
- [20] Salman OKM, Aksoy B. Rasgele orman ve ikili parçacık sürü zekası yöntemiyle kalp yetmezliği hastalığındaki ölüm riskinin tahminlenmesi. *International Journal of 3D Printing Technologies and Digital Industry* 2022; 6(3): 416-428.
- [21] Molajou A, Nourani V, Afshar A, Khosravi M, Brysiewicz A. Optimal design and feature selection by genetic algorithm for emotional artificial neural network (EANN) in rainfall-runoff modeling. *Water Resour Manag* 2021; 35(8): 2369-2384.

- [22] Bülbül MA. A New Artificial Intelligence Supported Approach to Diagnosis and Prediction of Psoriasis. *Gazi Journal of Engineering Sciences* 2024; 10(2): 405-418.
- [23] Diler S, Demir Y. Çoklu Doğrusal Bağlantı Olması Durumunda Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması. *Nicel Bilimler Dergisi* 2024; 6(1): 40-67.
- [24] Çiftci F. Türkiye'de Buğday Üretim Miktarının Geleceğini Tahmin Etmede Veri Madenciliği Yöntemlerinin Etkinliğinin Karşılaştırılması. *Wheat Studies* 2024; 13(1): 1-11.
- [25] Berus Y, Yakut YB. Derin Öğrenme (1D-CNN, RNN, LSTM, BiLSTM) ile Enerji Tüketim Tahmini: Diyarbakır AVM Örneği. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi* 2024; 15(2): 311-322.
- [26] Özbay E, Çınar A, Özbay FA. 3D Human Activity Classification with 3D Zernike Moment Based Convolutional, LSTM-Deep Neural Networks. *Traitement du Signal* 2021; 38(2): 269-280.
- [27] Özbay E, Özbay FA. Derin Öğrenme ve Sınıflandırma Yaklaşımları ile BT görüntülerinden Covid-19 Tespiti. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi* 2021; 12(2): 211-219.
- [28] Abbasoğulları EG, Gunay FB. Talasemi Hastalığı Tahmini İçin Farklı Makine Öğrenmesi Yöntemlerinin Kullanılması ve Karşılaştırılması. *Karadeniz Fen Bilimleri Dergisi* 2024; 14(4): 1990-2007.
- [29] Sonar 2022. <https://www.kaggle.com/datasets/rupakroy/sonarcsv> .
- [30] Breast Cancer Wisconsin (Diagnostic) Data Set 2017. <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>.