Mühendislik Bilimleri ve Tasarım Dergisi
11(1), 68 – 80, 2023
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.1117976

Araştırma Makalesi Research Article

TIBBİ VERİ KÜMELERİNDE GENETİK ALGORİTMALARLA ÖZELLİK SEÇİMİ VE SINIFLANDIRMA BAŞARIMINA ETKİSİ

Ömer DEPERLİOĞLU*

Afyon Kocatepe Üniversitesi, Afyon Meslek Yüksek Okulu, Bilgisayar Teknolojileri Bölümü, Afyonkarahisar, Türkiye

Anahtar Kelimeler

Özellik Seçimi, Genetik Algoritmalar, Tıbbi Veri Kümeleri, Özellikler Alt Kümesi, Tıbbi Sınıflandırma.

Öz

Günümüzde çok büyük boyuttaki tıbbi veri tabanlarından, klinik karar destek sistemlerinin faydalı bilgiler elde etmesi oldukça zorlaşmıştır. Genetik algoritmalar (GA) yaygın olarak kullanılan bir özellik seçme yöntemidir ve en iyi çözümleri verebilir. Bu çalışmada, çok sayıda karmaşık verilere sahip olan tıbbi verilerden özellik seçimi yapmak ve en uygun özellik alt kümesini oluşturarak sınıflandırma başarısını artırmak için GA içeren bir model önerilmiştir. Önerilen yöntemin performansını değerlendirmek için çalışmada en çok bilinen ve rahatlıkla ulaşılabilen 5 tıbbi veri kümesi ve 7 farklı denetimli sınıflandırma yöntemi kullanılmıştır. Her veri kümesi ile her sınıflandırıcı için ayrı ayrı özellik seçimi ve sınıflandırma uygulamaları yapılmıştır. Bu uygulamalarda elde edilen sonuçlar, önerilen yaklaşımla yapılan sınıflandırmalarda, veri kümesine bağlı olarak, Doğruluk oranında dolayısıyla makine öğrenmesi modeli performansında ortalama %2 ile %21 arasında artış sağlandığını ortaya koymuştur. Ayrıca yapılan çalışmalarda denetimli sınıflandırma algoritmalarından Rastgele Ormanın bütün veri kümelerinde diğer algoritmalardan daha iyi sonuçlar verdiği görülmekte ve tıbbi veri kümelerindeki sınıflandırma başarısı ile öne çıktığı görülmüştür.

FEATURE SELECTION WITH GENETIC ALGORITHMS AND ITS EFFECT ON CLASSIFICATION PERFORMANCE IN MEDICAL DATASETS

Keywords

Future Selection, Genetic Algorithms, Medical Data Set, Futures Subset, Medical Classification.

Abstract

Nowadays, it has become very difficult for clinical decision support systems to obtain useful information from very large medical databases. Genetic algorithms (GA) are a widely used feature selection method and can give the best solutions. In this study, a model with GA is proposed to select features from medical data with a large number of complex data and to increase classification success by creating the most appropriate feature subset. In order to evaluate the performance of the proposed method, 5 most well-known and easily accessible medical data sets and 7 different supervised classification methods were used in the study. Feature selection and classification applications were made separately for each data set and each classifier. The results obtained in these applications revealed that, depending on the data set, in the classifications made with the proposed approach, an average of 2% to 21% increase was achieved in the accuracy rate and thus in the machine learning model performance. In addition, it is seen that the Random Forest, one of the supervised classification algorithms, gives better results in all data sets than other algorithms and it has been seen that it stands out with its classification success in medical datasets.

Alıntı / Cite

Deperlioğlu, Ö., (2023). Tıbbi Veri Kümelerinde Genetik Algoritmalarla Özellik Seçimi ve Sınıflandırma Başarımına Etkisi, Mühendislik Bilimleri ve Tasarım Dergisi, 11(1), 68-80.

* İlgili yazar / Corresponding author: deperlioglu@aku.edu.tr, +90-272-218-2936

Yazar Kimliği / Author ID (ORCID Number)	Makale Süreci / Article Process	
Ö. Deperlioğlu, 0000-0002-7241-5219	**Başvuru Tarihi / Submission Date**	17.05.2022
	Revizyon Tarihi / Revision Date	30.09.2022
	Kabul Tarihi / Accepted Date	04.10.2022
	Yayım Tarihi / Published Date	27.03.2023

FEATURE SELECTION WITH GENETIC ALGORITHMS AND ITS EFFECT ON CLASSIFICATION PERFORMANCE IN MEDICAL DATASETS

Ömer DEPERLİOĞLU†
Afyon Kocatepe Üniversitesi, Afyon Meslek Yüksek Okulu, Bilgisayar Teknolojileri Bölümü, Afyonkarahisar, Türkiye

Highlights

- Today it has become very difficult to obtain useful information from very large medical databases.
- Genetic algorithms is a widely used feature selection method and can give the best solutions.
- A model with GA is proposed to select features from medical data and to increase classification success
- The proposed approach increased the performance of the model between 2% and 21% on average.

Purpose and Scope

In this study, a model with genetic algorithm is proposed to select features from medical data with a large number of complex data or to increase the classification success by creating the most appropriate subset of features.

Design/methodology/approach

Medical datasets constitute the input of the proposed model. In the data preprocessing stage, the data is transferred to the genetic algorithm section for feature selection, after checking the data for lack or missing areas, removing the fields that will not be used, such as diagnostic columns, if any. Then, the classification process is performed with the selected feature subset.

In order to evaluate the performance of the proposed method, applications were made with 5 medical datasets consisting of Pima Indian diabetes, Wisconsin Breast Cancer, Hepatitis, Cleveland Heart Diseases and Parkinson datasets from the most known and easily accessible medical databases in the machine learning laboratory of the University of California Irvine (UCI). As the classification algorithm, 7 different methods such as artificial neural networks, linear and radial core support vector machines, decision tree, logistic regression, random forests, K nearest neighbors, among supervised learning methods were used.

Findings

Feature selection and classification applications were made separately for each data set and each classifier. The results obtained in these applications revealed that, depending on the data set, in the classifications made with the proposed approach, an average of 2% to 21% increase was achieved in the Accuracy rate and thus the machine learning model performance. In addition, it has been seen that Random Forest, one of the supervised classification algorithms, gives better results than other algorithms in all datasets and it has been seen to come to the fore with its classification success in medical datasets.

Originality

Genetic Algorithm (GA) is one of the heuristic approaches that can be applied to many real-world applications to obtain optimized solutions and quality feature subsets in medical datasets. The feature selection techniques of machine learning and data mining also take advantage of the genetic algorithm to extract meaningful features from high-dimensional datasets. GA is capable of giving exact or estimated best solutions. For this reason, in this study, feature selection from medical datasets with GA was proposed and it was shown that it increased the classification success.

† Corresponding author: deperlioglu@aku.edu.tr, +90-272-218-2936

1. Giriş (Introduction)

Yaşlanan nüfus ve yaşam tarzı değişiklikleri, dünya genelinde sağlık sistemleri üzerinde artan baskılar oluşturmaktadır. Tıbbi algılayıcılar da dahil olmak üzere bilgi teknolojisindeki gelişmeler yoluyla sağlık ve hasta verilerinin sayısallaştırılmasıyla birlikte bu eğilimler, sağlık alanında büyük hacimli birincil ve ikincil verilerin üretilmesine yol açmıştır. Büyük veri talebi, göreceli klinik kararların aksine kanıta dayalı tıbbi tanıya geçişle de desteklenmektedir. Veri hazinesi sağlık hizmeti sunumunu, yönetimini ve politika oluşturmayı iyileştirmek için önemli fırsatlar sunarken, büyük verilerin etkin bir şekilde kullanılması için yeni bilgi sistemlerine ve yaklaşımlara ihtiyaç vardır. Gerçekten de büyük veri, geleneksel bilgi işlem araçları tarafından analiz edilemeyecek ve yönetilemeyecek kadar büyük ve karmaşık veriler olarak adlandırılmıştır (Kankanhalli vd., 2016).

Günümüzde hastaneler, verileri hastane bilgi sistemlerinde toplamak ve depolamak için orantılı olarak makul araçlar tahsis eden kapsamlı veri toplama araçlarıyla iyi bir şekilde donatılmıştır. Tıbbi veri tabanlarında biriken büyük miktarlardaki veriler, verilerin etkin bir şekilde kullanılması için verilerin depolanması, bunlara erişilmesi ve analiz edilmesi için özel araçlar gerektirir. Anlatım, metin, sayısal ölçümler, kayıtlı sinyaller ve görüntüler gibi çeşitli tıbbi veriler vardır. Son zamanlarda, bu veri boyutlarındaki büyüme nedeniyle karar desteği için faydalı bilgiler çıkarmak zahmetli hale gelmiştir (Jothi vd., 2019).

Sağlık alanında yaygın olan büyük miktarda yüksek boyutlu verilerin yorumlanması zorlu bir problem olarak devam etmekte olup, yüksek boyutlu ve düşük örneklem büyüklüğüne sahip olmaları nedeniyle aktif bir araştırma alanıdır. Bu problemler, yüksek doğruluk elde etmede mevcut sınıflandırma yöntemlerine önemli bir zorluk çıkarmaktadır. Bu nedenle, farklı hastalıkları doğru bir şekilde sınıflandırmak ve dolayısıyla tıp pratisyenlerine yardımcı olmak için bu durumda zorlayıcı bir özellik seçim yöntemi önemlidir.

Tıbbi veri kümelerinde makine öğrenmesinin başarısını birçok faktör etkiler. Verilerin kalitesi böyle bir faktördür. Bilgi alakasız veya fazla ise veya veriler gürültülü ve güvenilmez ise, eğitim sırasında bilgi keşfi daha zordur. Özellik seçimi, mümkün olduğunca alakasız ve gereksiz bilgilerin belirlenmesi ve çıkarılması işlemidir. Tıbbi verilerin işlenmesinde, en uygun özellik alt kümesini seçmenin iki büyük avantajı vardır. Birincisi veri tanımlamasının basitleştirilmesi, hekimlerin sağlıklı ve hızlı tanı koymasını kolaylaştırabilir. İkincisi daha az özelliğe sahip olmak, daha az veri toplanması gerektiği anlamına gelir. Bilindiği gibi, zaman alıcı ve maliyetli olan tıbbi uygulamalarda veri toplamak kolay bir iş değildir (Wang ve Ma, 2009).

Sınıflandırma, popüler makine öğrenimi işi, bilgisayar destekli teşhis sistemlerinin bir parçası ve çeşitli tıbbi veri analiz yazılımı paketleridir. Daha yüksek sınıflandırma doğruluğu elde etmek için sınıflandırma modeli için fonksiyonel küme ve uygun parametrelerin seçilmesi gereklidir. Tıbbi veri tabanları ayrıca birçok özelliğin diğerleriyle ilişkili olduğu geniş bir özellik kümesi içerir, bu nedenle özellik kümesini azaltmak esastır. Çoğu sınıflandırıcı, bir eğitim süreci boyunca verilerden öğrenebilecekleri şekilde yapılandırılmıştır, çünkü tam uzman deneyimi sınıflandırma parametrelerini değerlendirmek için gerçekçi değildir (Kumar, 2021). Özellik seçimi veya özellik azaltma, büyük veri işlemedeki en kritik adımlardan biridir. Bir özellik seçim algoritması, özellik vektöründen en alakalı özellikleri seçer ve alakasız nitelikleri bırakır. Bu aynı zamanda makine öğrenmesi ve örüntü tanımada aktif bir araştırma alanıdır. Bu büyük veri nedeniyle, özellik çıkarma prosedüründe daha büyük boyutlu bir özellik kümesi elde edilir. Çıkarılan tüm öznitelikler kullanışlı değildir ve daha yüksek boyutlu bir vektör, zaman maliyeti ve doğruluk açısından modelin performansını etkiler. Boyutluluğu azaltmak ve aynı zamanda bilgi çıkarma ve modelin anlaşılması maliyetlerini azaltmak için özellik seçim teknikleri kullanılır. (Naheed vd., 2020).

Veri sınıflandırmasında en iyi sonucu elde etmek için uygun özellikleri seçmek son yıllarda en zorlu konulardan biri olmuştur. Öğrenme teorisinden, daha fazla özelliğin kullanılması tahminin doğruluğunu artırsa da pratik kanıtlar bunun her zaman doğru olmadığını gösterir çünkü tüm özellikler veri sınıfı etiketini tespit etmek için önemli değildir veya bazıları veri etiketi ile ilgisizdir. Özellik seçim yöntemleri üç kategoriye ayrılabilir: filtreleme, sarmalama ve gömülü olanlar. Filtreleme yöntemleri, sınıfların ne kadar iyi ayrıldığını gösteren mesafe kriteri gibi dolaylı bir kritere dayalı olarak tahminlerin veya sınıflandırmaların doğruluğunu ölçer. Sarmalama yöntemi tamamen sınıflandırma modeline bağlıdır ve algoritma, sınıflandırıcı modelden elde edilen doğruluğa dayalı olarak en uygun alt kümeyi belirler. Seçim kriteri elde edilen doğrulukla aynıdır ve daha yüksek doğruluk sağlayan bir alt küme seçilir. Gömülü yöntemler, öğrenme sürecinde özellik seçimini gerçekleştirir ve genellikle bir öğreniciye atanır. Bu model ayrıca, farklı arama aşamalarında farklı değerlendirme kriterlerini kullanarak önceki her iki modelden de yararlanır (Abdollahi ve Nouri-Moghaddam, 2021; Mwadulo, 2016). Ancak, topluluk özellik seçimi üretmek için bu üç tür tekniği birleştirmeye odaklanan bir çalışma yapılmaması nedeniyle Chen ve arkadaşları kategorik, sayısal ve karma veri türleri dahil olmak üzere farklı tıbbi veri türleri için farklı türdeki özellik seçim algoritmalarının hangi kombinasyonunun en iyi performansı sunduğu sorusuna cevap aramışlardır. Deneysel

sonuçlar, birleşim yöntemiyle temel bileşen analizi gibi filtre ve genetik algoritmalar gibi sarmalayıcı tekniklerin bir kombinasyonunun, nispeten yüksek sınıflandırma doğruluğu ve oldukça iyi bir özellik azaltma oranı sağlayarak daha iyi bir seçim olduğunu göstermişlerdir (Chen vd., 2020).

Son zamanlarda, tıbbi veri kümelerinden en uygun özellik alt kümesinin seçilebilmesi için araştırmacılar tarafından çeşitli özellik seçim algoritmaları tanıtılmış ve kullanılmıştır. Bu algoritmalar, genetik algoritma (GA), entropi seçimi, Parçacık Sürü Optimizasyonu ve Karınca kolonisi gibi birçok yöntemi içermektedir. Bu teknikler hem doğruluğu hem de zaman performansını artırmıştır. GA günümüzde yaygın olarak kullanılan bir özellik seçme yöntemidir. GA, kesin veya tahmini en iyi çözümleri verme yeteneğine sahiptir (Naheed vd., 2020; Goldberg, 1989; Booker vd., 1989). Bu nedenle bu çalışmada tıbbi veri kümelerinden GA ile özellik seçimi ve sınıflandırma başarısına etkisi araştırılmıştır.

Makalenin bundan sonraki bölümlerin düzeni şöyle belirtilebilir. Bölüm 2'de, çalışmada kullanılan veri kümeleri ve yöntemler açıklanmıştır. Bölüm 3'te uygulamanın ayrıntıları verilmiştir. Bölüm 4'te, önerilen yöntemin performansı elde edilen sonuçlarla tartışılmaktadır. Makale 5. bölümdeki sonuç ve öneriler kısmı ile sonlandırılmaktadır.

2. Kaynak Araştırması (Literature Survey)

Son yıllarda, yararlı tıbbi bilgileri ve kuralları otomatik olarak keşfetmek için makine öğrenimi veya veri madenciliği tekniklerini kullanmaya yönelik önemli araştırmalar yapılmaktadır.

Yeniterzi ve arkadaşları kalp aritmileri üzerine yaptıkları çalışmalarında, genetik algoritmalar kullanılarak 278 öznitelikte veri kümesini çok daha iyi açıklayan özelliklerin keşfedilmesine odaklanmaktadır. GA sonuçlarından elde edilen özellikler kullanılarak sınıflandırma doğruluğunu %90'a kadar yükseltmişlerdir (Yeniterzi vd, 2007). Bir başka çalışmada, meme kanseri teşhisi için özellik seçimini ele almaktadır. Mevcut süreç, özellik seçimine ve PS sınıflandırıcıya dayalı GA kullanan bir sarmalayıcı yaklaşımı kullanmaktadır. Deney sonuçları, önerilen modelin Wisconsin meme kanseri veri kümelerindeki diğer modellerle karşılaştırılabilir verimliliğe sahip olduğunu göstermektedir (Aalaei vd, 2016). Algoritmaları karşılaştıran bir çalışmada, Temel Bileşen Analizi (TBA), Faktör Analizi (FA) ve Nitelik Sıralaması (NS) yöntemi olmak üzere üç farklı özellik seçme yöntemini karşılaştırmışlardır. TBA' nın yüksek performansını Naïve Bayes sınıflandırıcı ve K-en yakın komşular sınıflandırıcı kullanılarak bir veri kümesi üzerinde bir dizi deney yoluyla doğrulamışlardır (Samant ve Rao, 2013).

Tıbbi veri kümelerinde farklı algoritmalarla özellik seçimi üzerine yapılan bir çalışmada, ayarlanmış beyin fırtınası optimizasyon algoritması önermektedir. Sınıflandırma, parametrelerinin beyin fırtınası optimizasyon algoritması ile optimize edildiği destek vektör makinesi ile yapılmıştır. Önerilen yöntem, kamuya açık standart tıbbi veri kümeleri üzerinde test edilmiş ve diğer son teknoloji yöntemlerle karşılaştırılmıştır. Elde edilen sonuçlar analiz edilerek önerilen yöntemin daha yüksek doğruluk sağladığı ve ihtiyaç duyulan öznitelik sayısını azalttığı gösterilmiştir (Tuba vd, 2019). Uygulamalarda farklı algoritmalar kullanıldığı gibi karma yöntemler de özellik seçiminde kullanılmaktadır. Örneğin Khadir ve Amanullah makalelerinde, sınıflandırma veya kümeleme sonuçlarının iyileştirilmesi için aynı öznitelik seçim süreci için geliştirilmiş bir genetik algoritma sunmaktadır. Bir bilgi tahmin modeli oluşturmak için en iyi etkileyen niteliği belirlemek için uygunluk fonksiyonu olarak Çoklu Doğrusal Regresyon (MLR) tekniği kullanmışlardır. MLR-GA'nın sonuçları, doğruluk açısından mevcut özellik seçim algoritmalarından daha iyi performans sergilediğini göstermişlerdir (Khadir ve Amanullah, 2017).

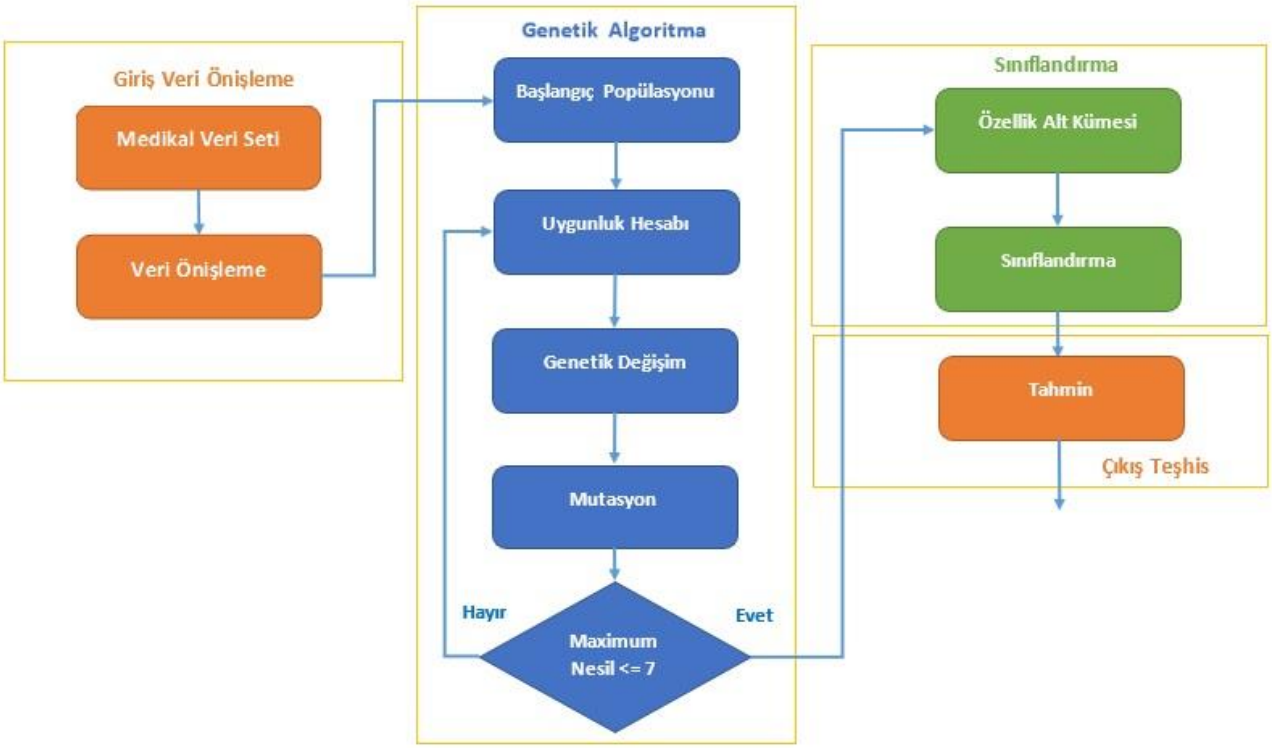
Veri kümelerinde özellik seçimi için Parçacık Sürü Optimizasyonunun (PSO) da sıklıkla kullanıldığı görülmektedir. Örneğin Rostami ve arkadaşları bir çalışmalarında, yeni bir Parçacık Sürü Optimizasyonu tabanlı çok amaçlı özellik seçim yöntemi önermiştir. Önerilen yöntem üç ana aşamadan oluşmaktadır. İlk aşamada, orijinal özellikler bir grafik temsil modeli olarak gösterilmektedir. Bir sonraki aşamada, grafikteki tüm düğümler için özellik merkezilikleri hesaplanmakta ve son olarak üçüncü aşamada, son özellik seçimi için geliştirilmiş bir PSO tabanlı arama süreci kullanılmaktadır. Beş tıbbi veri kümesi üzerinde elde edilen sonuçlarla, önerilen yöntemin verimlilik ve etkinlik açısından diğer ilgili yöntemleri iyileştirdiğini göstermişlerdir (Rostami vd, 2020). Benzer şekilde farklı hastalık türlerinin varlığını veya yokluğunu tespit etmek için sarmalayıcı tabanlı öznitelik seçim yöntemi olarak kabul görmüş çok amaçlı parçacık sürü optimizasyonunun kullanımını araştırmışlardır (Habib vd, 2020). Yine benzer bir çalışmada yazarlar özellik seçimi amacıyla üç adımda işlev gören topluluk algılamaya dayalı bir genetik algoritma önermektedir. İlk adımda özellik benzerlikleri hesaplanıyor. Özellikler, ikinci adım boyunca topluluk algılama algoritmaları tarafından kümeler halinde sınıflandırılıyor. Üçüncü adımda, özellikler, topluluk tabanlı yeni bir onarım işlemiyle genetik bir algoritma tarafından seçiliyor. Dokuz kıyaslamalı sınıflandırma problemi ile yaklaşımın performansı analiz edilmiştir. Ayrıca yazarlar, önerilen yaklaşımın verimliliğini özellik seçimi için parçacık sürü optimizasyonu, Karınca Kolonisi Optimizasyonu ve Yapay Arı Kolonisi algoritmalarından elde edilen bulgularla karşılaştırmışlardır. Buna göre GA ile daha iyi başarım elde ettiklerini belirtmişlerdir (Rostami, vd, 2021).

Aynı şekilde 30 veri kümesi ve 14 sınıflandırıcı ile yaptıkları çalışmada özelliklerin özdeğerleri ile en iyi özellik alt kümesini bulmak için Genetik algoritmanın ideal bir yöntem olduğunu ve 14 sınıflandırıcının tamamı arasında Radyal tabanlı fonksiyon ağının (RBF) en iyi ortalama performansa sahip olduğu belirtilmiştir (Ershadi ve Seifi, 2022).

Özellik seçimi, makine öğrenmesi analizinde çok önemli bir adımdır. Halihazırda, birçok özellik seçimi yaklaşımı, 'Omics' veri kümelerinde olduğu gibi, veri miktarı çok büyük olduğunda, doğruluk ve hesaplama süresi açısından tatmin edici sonuçlar sağlamamaktadır. Bu nedenle yazarlar çok sınıflı ve yüksek boyutlu veri kümelerinde bilgilendirici özelliklerin hızlı ve doğru bir şekilde tanımlanması için GARS adı verilen bir genetik algoritma ile özellik seçimi uygulamasını önermişlerdir. Tüm benzetimlerde, GARS' ın makul bir hesaplama süresinde yüksek sınıflandırma doğrulukları göstererek, iki standart filtre tabanlı, iki sarmalayıcı ve bir gömülü seçim yönteminden daha iyi performans gösterdiğini belirtmişlerdir (Chiesa vd., 2020). Bunun dışında küçük ve büyük tıbbi veri kümelerinden etkili özellikleri seçmek için ikili bir güve alevi optimizasyonu (B-MFO) (Nadimi-Shahraki vd., 2021), en uygun özellik alt kümesini seçmek için Gömülü Hibrit Filtre Sarıcı (HFWE) tabanlı özellik seçimi gibi özel algoritmalar da geliştirilmiştir (Parthiban, 2021).

3. Materyal ve Yöntem (Material and Method)

Yukarıda örneklerle belirtildiği gibi optimize edilmiş çözümlerin elde edilmesi için birçok gerçek dünya uygulamasına uygulanabilen sezgisel yaklaşımlardan birisi Genetik Algoritmadır (GA). Makine öğrenmesi ve veri madenciliğinin özellik seçme teknikleri, yüksek boyutlu veri kümelerinden anlamlı özellikleri çıkarmak için genetik algoritmanın avantajlarından da yararlanmaktadır. Aynı amaçla bu çalışmada, çok sayıda karmaşık verilere sahip olan tıbbi verilerden özellik seçimi yapmak veya en uygun özellikler alt kümesini oluşturarak sınıflandırma başarısını artırmak için genetik algoritma içeren bir model önerilmiştir. Önerilen yöntemin genel yapısı Şekil 1'de verilmiştir. Şekilde görüldüğü gibi sistemin girişini tıbbi veri kümeleri oluşturmaktadır. Veri önişleme aşamasında veriler kontrol edilerek eksiklik veya boşluk olup olmadığı kontrol edildikten sonra veri kümesinde varsa tanı sütunları gibi kullanılmayacak olan alanlar çıkartıldıktan sonra veriler özellik seçimi için genetik algoritma bölümüne aktarılıyor.



Şekil 1. Önerilen yöntemin genel yapısı (general structure of the proposed method)

Genetik algoritma bölümünde, öncelikle başlangıç popülasyonu özellik kümelerinin örnek uzayından rastgele oluşturuluyor. Oluşturulan rastgele popülasyon, en yüksek doğruluk veren en iyi ebeveynleri döndüren uygunluk işlevinden geçiriliyor. Burada, hedef uygunluk değeri belirtilen sınıflandırıcının doğruluk değerini içeren performansıdır. Böylece en iyi ebeveynlerin seçimi gerçekleştiriliyor. Hangi üyelerin bir sonraki nesle devam edeceği belirlendikten sonra, önce genetik değişim (çaprazlama) ardından mutasyon işlemleri yürütülüyor. Genetik değişim, birinci ebeveynin bir kısmı ile ikinci ebeveynin bir kısmı rastgele seçilerek en uygun iki

ebeveynden gelen genlerin birleştirilmesiyle oluşturuluyor. Mutasyon ise genetik değişim çocuğundan seçilen bitlerin rastgele çevrilmesiyle elde ediliyor. Bir önceki nesilden en uygun ebeveynler seçilerek genetik değişim ve mutasyon uygulanarak yeni nesil oluşturulma işlemi belirtilen 7 nesil boyunca yineleniyor. Böylece sınıflandırıcı için uygun özellikler seçilmiş oluyor.

GA ile özellik seçimi ile oluşturulan özellikler alt kümesi ile sınıflandırma yapılarak ilgili veri kümesi için teşhis veya tanı tahmini gerçekleştiriliyor.

3.1. Tıbbi Veri Kümeleri (Medical Data Sets)

Önerilen yöntemin performansını değerlendirmek için en çok bilinen ve rahatlıkla ulaşılabilen 5 tıbbi veri kümesi ile uygulamalar yapılmıştır. Çalışmada kullanılan Kaliforniya Üniversitesi Irvine (UCI) makine öğrenme laboratuvarındaki tıbbi veri tabanlarından yaygın kullanılan, Pima Indian diyabet, Wisconsin Göğüs Kanseri, Hepatitler, Cleveland Kalp Hastalıkları ve Parkinson veri kümelerinin özellikleri Tablo 1'de verilmiştir (UCI, 2007).

Tablo 1. Tıbbi veri kümelerinin özellikleri (attributes of medical data sets)

Veri Kümesi	Örnek Sayısı	Özellik Sayısı	Çıkış Sınıfı Sayısı
Pima Indian Diabet	768	8	2
Wisconsin Göğüs Kanseri	103	9	6
Hepatit	155	19	2
Cleveland Kalp Hastalıkları	303	13	5
Parkinson Hastalığı	195	22	2

3.1.1. Pima Indian Diyabet Veri kümesi (Pima Indian Diabetes Data Set)

Pima Indian diyabet veri kümesindeki bu örneklerin, daha büyük bir veri tabanından seçilmesine ilişkin çeşitli kısıtlamalar yapılmıştır. Özellikle, tüm hastalar en az 21 yaşında Pima Indian kökenli kadınlardan seçilmiştir. Veri kümesinde Toplam 768 örnek bulunmaktadır ve her bir örnek için 8 özellik mevcuttur. Bu özellikleri şunlardır: 1) gebelik sayısı, 2) plazma glikoz konsantrasyonu, 3) diyastolik tansiyon, 4) kol kası cilt deri kalınlığı, 5) serum insülin, 6) vücut kütleendeksi, 7) diyabet soyağacı fonksiyonu ve 8) yaş. Veri kümesindeki verilerin, sağlıklı-diyabet negatif ve diyabetli-diyabet pozitif olmak üzere 2 sınıftan oluşmaktadır (Jaganathan ve Kuppuchamy, 2013).

3.1.2. Wisconsin Göğüs Kanseri Veri kümesi (Wisconsin Breast Cancer Dataset)

Wisconsin Göğüs kanseri Veri kümesi, Wisconsin Üniversitesi Madison Hastaneleri'nde Dr. William H. Wolberg tarafından 1989-1991 yılları arasında toplanmıştır. Dokuz özellik ile karakterize 699 örnek içermektedir: 1) Küme Kalınlığı, 2) Hücre Boyutunun Eşsizliği, 3) Hücre Şeklinin Eşbiçimi, 4) Marjinal Yapışma, 5) Tek Epitelyal Hücre Boyutu, 6) Çıplak Çekirdekler, 7) Bland Kromatin, 8) Normal Nükleoliant ve 9) İyi huylu veya habis büyümeleri tahmin etmek için kullanılan mitozlar. Bu veri kümesinde, elektrikle empedans ölçümleri kullanılarak altı yeni çıkarılan doku sınıfı çalışılmıştır. Bu dokular kanser 21, Fibro adenom 15, mastopatileri 18, beze gibi 16, bağlayıcı 14, Yağ 22 adettir (Salama vd., 2012).

3.1.3. Hepatit Veri kümesi (Hepatitis Data Set)

Hepatitler veri kümesi Carnegie-Mellon Üniversitesi'nden alınmıştır. Her bir örnek 19 özellik içermektedir. Bu özellikler: 1) yaş, 2)cinsiyet, 3) steroid, 4) antiviraller, 5) yorgunluk, 6) halsizlik, 7) anoreksiya, 8) karaciğer, 9) karaciğer filmi, 10) dalak bulguları, 11) örümcekler ,12) asitler, 13) varisler, 14) bilirubin, 16) alkfosfat, 17) SGOT, 18) albumin, 19) protime ve 20) histolojidir.Veri kümesinde canlı veya ölü olmak üzere toplam 155 örnek vardır (Ba-Alwi ve Hintaya, 2013).

3.1.4. Cleveland Kalp Hastalıkları Veri kümesi (Cleveland Heart Diseases Data Set)

Veri kümesi Cleveland Kliniği Vakfı'ndan toplanmış ve her biri başlangıçta 76 ham özellikten seçilmiş 13 özelliği olan yaklaşık 303 örnek içermektedir. Özellikleri şunlardır:1) yaş, 2) cinsel ilişki, 3) göğüs ağrısı türü, 4) istirahat tansiyonu,5) kolestrol, 6) açlık kan şekeri, 7) dinlenme elektrokardiyografik sonuçlar, 8) maksimum kalp atışı, 9) egzersiz indükteanjina, 10) segmente göre egzersiz tarafından indüklenen depresyon, 11) zirve egzersizinin eğimi, 12) büyük damarların sayısı ve 13) tal. Veri kümesindeki veriler, kalp hastalığının varlığını gösteren değerler 1, 2, 3, 4 ve yokluğunu gösteren değer 0 olmak üzere 5 sınıftan oluşmaktadır (Detrano vd., 1989).

3.1.5. Parkinson Hastalığı Veri Kümesi (Parkinson's Disease Data Set)

Bu veri kümesi, 23'ü Parkinson hastalığı olan 31 kişiden alınan çeşitli biyomedikal ses ölçümlerinden oluşmaktadır. Tablodaki her sütun belirli bir ses ölçüsüdür ve her satır bu kişilerden gelen 195 ses kaydından birine karşılık gelir. Veri kümesindeki özellik bilgileri, name- ASCII konu adı ve kayıt numarası, MDVP: Fo(Hz) - Ortalama vokal temel frekansı, MDVP: Fhi(Hz) - Maksimum vokal temel frekansı, MDVP: Flo(Hz) - Minimum vokal temel frekansı, MDVP: Jitter(%), MDVP: Jitter(Abs), MDVP: RAP, MDVP:PPQ, Jitter: DDP - Temel frekansta çeşitli varyasyon ölçütleri, MDVP: Işıltı, MDVP: Işıltı(dB), Işıltı: APQ3, Işıltı: APQ5, MDVP: APQ, Işıltı: DDA - Genlikte çeşitli varyasyon ölçütleri NHR, HNR - Sesteki gürültünün ton bileşenlerine oranının iki ölçüsü, state - Deneğin sağlık durumu (bir) - Parkinson, (sıfır) – sağlıklı, RPDE, D2 - İki doğrusal olmayan dinamik karmaşıklık ölçüsü, DFA - Sinyal fraktal ölçekleme üssü, spread1, spread2, PPE - Temel frekans değişiminin üç doğrusal olmayan ölçümü oluşmaktadır. Veri kümesinde hasta başına yaklaşık altı kayıt vardır. Veri kümesi, sağlıklı insanlar için 0 ve Parkinson hastaları için 1 olmak üzere iki çıkış sınıfına sahiptir (Little vd., 2008).

3.2. Kullanılan Yapay Zekâ Yöntemleri (Used Artificial Intelligence Methods)

Önerilen modelde sarmalayıcı özellik seçme yöntemi olarak Genetik Algoritmalar (GA) kullanılmıştır. Önerilen yöntemin performansını değerlendirmek için yapay zekâda denetimli öğrenme yöntemlerinden Yapay Sinir Ağları (YSA), doğrusal ve radyal çekirdekli Destek Vektör Makineleri (DVM), Karar Ağacı (KA), Lojistik Regresyon (LR), Rastgele Ormanlar (RO), K en yakın Komşu (K-YK) olmak üzere 7 adet denetimli sınıflandırma yöntemi kullanılmıştır. Bu yöntemler aşağıda kısaca tanıtılmıştır.

3.2.1. Genetik Algoritmalar (Genetic Algorithms)

Genetik algoritmalar, belirli bir problem için en uygun çözümü bulmaya çalışır. Genetik algoritmalar, belirli bir problem için bireyler adı verilen bir aday çözüm popülasyonu sağlar. Bu aday çözümler yinelemeli olarak değerlendirilir ve yeni nesil çözümler oluşturmak için kullanılır. Bu sorunu çözmede daha iyi olanların seçilme ve niteliklerini yeni nesil aday çözümlere aktarma şansı daha yüksektir. Bu şekilde, nesiller geçtikçe, aday çözümler eldeki sorunu çözmede daha iyi hale gelir (Wirsansky, 2020).

Genetik algoritmaların altında yatan yapı taşı hipotezi, eldeki problemin optimal çözümünün küçük yapı taşlarından bir araya getirilmesidir. Bu yapı taşlarından daha fazlası bir araya getirildikçe ideal çözüme yaklaşılmış olur. İstenen yapı taşlarından bazılarını içeren popülasyondaki bireyler, üstün puanlarıyla tanımlanır. Tekrarlanan seçim ve genetik değişim işlemleri, bu yapı taşlarını sonraki nesillere daha iyi aktaran ve muhtemelen bunları diğer başarılı yapı taşlarıyla birleştiren daha iyi bireylerle sonuçlanır. Bu, genetik baskı yaratır, böylece popülasyonu ideal çözümü oluşturan yapı taşlarına sahip daha fazla bireye sahip olmaya yönlendirir. Sonuç olarak, her nesil bir öncekinden daha iyidir ve ideal çözüme daha yakın olan daha fazla birey içerir.

3.2.2. Yapay Sinir Ağları (Artificial Neural Network)

YSA aşağıdaki özelliklere sahip paralel dağınık bilgi işleme yapısıdır: Sinirlerden esinlenmiş matematik modelidir ve çok sayıda birbirine çok iyi bağlı işlem elemanlarını kapsar. Ağırlıklar veya bağlantılar bilgiyi tutar, işlem elamanları giriş uyarıcılarına dinamik olarak tepki gösterebilir. Tepki tamamen işlem elemanına bağlantılar ve bağlantıların ağırlıkları ile gelen giriş sinyallerinden oluşan bölgesel bilgisine bağlıdır. Öğrenme, hatırlama ve görünen veya ayarlanan bağlantı ağırlıklarıyla öğrenme verisinden genelleştirme özelliklerine sahiptir. Ortaklaşa davranışı hesaplanabilir güç gösterir ve yayılmış temsil özelliği nedeniyle tek sinir özel bilgi taşıyıcı değildir. Ayrıca bir sinir ağı yalnızca sayısal, sürekli bilgileri işleyebilir. Bir kaliteyi belirten bir görüntüde kırmızı, mavi veya yeşil etiketler gibi nitel değişkenleri işleyemez. Niteliksel değişkenleri, bir dizi ikili değer gibi sürekli bir sayısal değere dönüştürerek işlenebilir.

3.2.3. Destek Vektör Makinesi (Support Vector Machines)

Destek Vektör Makinesi (DVM), çoğunlukla sınıflandırma çalışmalarında yaygın olarak kullanılan denetimli öğrenme algoritmasıdır. Ancak regresyon problemleri için de kullanılabilmektedir. Ancak, öncelikle makine öğreniminde sınıflandırma problemleri için kullanılır. DVM algoritmasının amacı, gelecekte yeni veri noktasını kolayca doğru kategoriye koyabilmemiz için n-boyutlu uzayı sınıflara ayırabilen en iyi çizgi veya karar sınırını oluşturmaktır. Bu en iyi karar sınırına hiperdüzlem denir. DVM, hiper düzlemi oluşturmaya yardımcı olan uç noktaları/vektörleri seçer. Bu uç durumlara destek vektörleri denir ve bu nedenle algoritma Destek Vektör Makinesi olarak adlandırılır (Machine Learning Notes, 2020). DVM yapısına göre Doğrusal DVM ve Doğrusal olmayan DVM olmak üzere ikiye ayrılır.

Ayrıca, DVM algoritmaları öğrenme aşamasında bazı matematiksel fonksiyonlar kullanır ve bunlar çekirdek olarak tanımlanır. Çekirdeğin ana görevi girdi olarak sunulan verileri alarak, bunları gerekli forma dönüştürmektir. Farklı DVM algoritmaları, farklı türde çekirdek işlevleri kullanır. Bu fonksiyonlar farklı tiplerde olabilir. Örneğin lineer, lineer olmayan, polinom, radyal temel fonksiyon ve sigmoid. En çok kullanılan çekirdek işlevi türü radyaldır. Çünkü tüm x ekseni boyunca lokalize ve sonlu bir tepkiye sahiptir. Çekirdek işlevleri, uygun bir özellik uzayında iki nokta arasındaki iç çarpımı döndürür. Böylece, çok yüksek boyutlu uzaylarda bile çok az hesaplama maliyeti ile bir benzerlik kavramı tanımlayabilir.

3.2.4. Karar Ağaçları (Decision Trees)

Karar Ağaçları, verilerin belirli bir parametreye göre sürekli olarak bölündüğü bir denetimli makine öğrenmesi türüdür. Yani, eğitim verilerinde girdinin ne olduğunu ve karşılık gelen çıktının ne olduğunu belirten örneklerle öğrenir. Temel olarak, bağımsız değişkenlere dayalı özyinelemeli bölüm olarak ifade edilen bir sınıflandırıcıdır. Ağaç, karar düğümleri ve yapraklar olmak üzere iki varlıkla açıklanabilir. Yapraklar kararlar veya nihai sonuçlardır. Karar düğümleri, verilerin bölündüğü yerdir. Başka bir deyişle karar ağacı, köklü ağacı oluşturan düğümlere sahiptir. Köklü ağaç, kök adı verilen bir düğüme sahip yönlendirilmiş bir ağaçtır. Kökün herhangi bir gelen kenarı yoktur ve diğer tüm düğümlerin bir gelen kenarı vardır. Bu düğümlere yaprak veya karar düğümleri denir (Tutorials Point, 2016).

3.2.5. Lojistik Regresyon (Logistic Regresyon)

Lojistik regresyon, sınıflandırma problemlerini çözmek için kullanılan bir başka denetimli öğrenme algoritmasıdır ve logit regresyon olarak da bilinir. Temel olarak lojistik regresyon, belirli bir bağımsız değişken kümesine dayalı olarak 0 veya 1, doğru veya yanlış, evet veya hayır gibi ayrık değerleri tahmin etmek için kullanılan bir sınıflandırma algoritmasıdır. Temel olarak, olasılığı tahmin eder, dolayısıyla çıktısı 0 ile 1 arasında bulunur.

3.2.6. Rasgele Ormanlar (Random Forests)

Rastgele orman, birden fazla karar ağacının oluşturulduğu ve daha doğru bir tahmin elde etmek için birleştirildiği bir topluluk öğrenme yöntemidir. Makine öğreniminde son birkaç yılda popülaritesi artan bir yöntem varsa, o da rastgele ormanlar fikridir. Temel olarak karar ağaçlarının toplanmasıdır yani orman veya karar ağaçlarının topluluğu denilebilir. Rastgele ormanın temel konsepti, her ağacın bir sınıflandırma vermesi ve ormanın bunlardan en iyi sınıflandırmayı seçmesidir. Rastgele orman algoritmasının avantajları çoktur. Bunlardan bazıları: Rastgele orman sınıflandırıcısı hem sınıflandırma hem de regresyon görevleri için kullanılabilir, eksik değerleri işleyebilirler.

3.2.7. K-En Yakın Komşu (K-Nearest Neighbor)

Denetimli öğrenme tekniğine dayalı en basit makine öğrenimi algoritmalarından biridir. K-YK algoritması, yeni durum/veriler ile mevcut durumlar arasındaki benzerliği varsayar ve yeni durumu mevcut kategorilere en çok benzeyen kategoriye yerleştirir. K-YK algoritması, mevcut tüm verileri saklar ve benzerliğe göre yeni bir veri noktasını sınıflandırır. Sınıflandırmanın yanı sıra regresyon için de kullanılabilir ancak daha çok sınıflandırma problemleri için kullanılır.

K-YK, parametrik olmayan bir algoritmadır, yani temel veriler üzerinde herhangi bir varsayımda bulunmaz. Tembel öğrenen algoritması olarak da adlandırılır, çünkü eğitim kümesinden hemen öğrenmez, bunun yerine veri kümesini depolar ve sınıflandırma anında veri kümesi üzerinde bir işlem gerçekleştirir. K-YK algoritması eğitim aşamasında sadece veri kümesini saklar ve yeni veri aldığında bu veriyi yeni veriye çok benzeyen bir kategoride sınıflandırır.

3.3. Sınıflandırma Performans Değerlendirmesi (Evaluation of Performance Classification)

Tıbbi sınıflandırma çalışmalarının performansını değerlendirmek için çoğunlukla 7 farklı performans ölçütleri tercih edilmektedir. Bu performans ölçütleri doğruluk (Accuracy), kesinlik (precision), özgüllük (specificity), duyarlılık (sensitivity), geri çağırma (recall-hatırlama), F ölçümü veya F1 Puanı (F measure or F1 Score) ve AUC'dir. Bu ölçütlerden önerilen modelin tahmin kabiliyetini ve doğruluğunu değerlendirmek için Doğruluk kullanılmıştır ve aşağıdaki gibi hesaplanabilir (Sokolova ve Lapalme, 2009; Deperlioglu, 2019):

$$\text{Doğruluk} = \frac{tp+tn}{tp+fp+fn+tn} \quad (1)$$

Bu denklemde, *tp* gerçek ve gerçek pozitiflerin sayısıdır. *fp*, yanlış pozitif tanı sayısına karşılık gelir. *tn* gerçek negatif sayıdır. *fn*, yanlış negatif tanıların sayısına karşılık gelir.

4. Deneysel Sonuçlar (Experimental Results)

Tıbbi veri kümelerinden genetik algoritmalarla özellik seçimi yaparak özellik alt kümesinin oluşturulması çok kullanılan halka açık 5 veri kümesi üzerinde yaygın kullanılan 7 ayrı denetimli sınıflandırıcı üzerinde denenmiştir. Tüm uygulamalarda Python programlama dili ve Scikit learn kitaplığı kullanılmıştır. Kullanılan sınıflandırıcıların özellikleri bu kitaplığa göre seçilmiştir. YSA' da her katmanda 256, 128, 64, 32 farklı nöronlar bulunan dört gizli katman modellenmiştir. Girdi ve çıktı katmanı düşünüldüğünde modelde toplam 6 katman bulunmaktadır. Aktivasyon işlevi 'relu', çözücü işlevi 'adam', yineleme sayısı (max_iter) 1000 olarak seçilmiştir. DVM' de doğrusal ve radyal iki ayrı çekirdekle (kernel='linear', kernel='rbf') sınıflandırma yapılmıştır. Lojistik regresyonda maksimum yineleme sayısı 1000 olarak seçilmiştir. Rastgele orman sınıflandırıcıda tahmin edici (n_estimators) 200, rasgele durum (random_state) 0 olarak alınmıştır. Karar ağacında rasgele durum (random_state) 0 olarak alınmıştır. Bunlarında dışında özel bir düzenleme yapılmamış varsayılan sınıflandırıcı yapısı kullanılmıştır. Genetik algoritmada özellik seçimi için özellik sayısı kadar popülasyon kullanılmış ve 7 nesil boyunca tekrar edilerek en iyi nesile ulaşılmaya veya en iyi özellik alt kümesi elde edilmeye çalışılmıştır.

Uygulamada tüm veri kümelerinde ilk önce veri denetimi yapılmış, ardından isim, tanıtım ve tanı verileri gibi kullanılmayacak özellikler çıkarılmıştır. Daha sonra karşılaştırma yapabilmek amacıyla tüm özelliklerle 10 kez sınıflandırma yapılarak ortalama başlangıç Doğruluk oranı hesaplanmıştır. Ardından GA ile her veri tabanında özellik seçimi yapılarak yine her sınıflandırıcı için 10 kez sınıflandırma işlemi gerçekleştirilmiştir. Bu sınıflandırma işlemlerinden elde edilen sonuçlar aşağıda verilmiştir.

4.1. Pima Indian Diyabet Veri kümesi (Pima Indian Diabetes Data Set)

Pima Indian Diyabet Veri kümesinde sınıflandırmada kullanılan 8 özellik ve toplamda 768 örnek ile yapılan özellik seçimi uygulamasında elde edilen sonuçlar Tablo 2' de verilmiştir.

Tablo 2. Pima Indian diyabet veri kümesi ile elde edilen sonuçlar (Results obtained with the pima Indian diabetes dataset)

Sınıflandırma Yöntemi	Başlangıç Doğruluk (%)	Seçilen Özellik Sayısı	En Düşük Doğruluk (%)	Ortalama Doğruluk (%)	En Yüksek Doğruluk (%)
Yapay Sinir Ağları	69,27	5	73,43	74,47	75,52
Destek Vektör Doğ.	72,91	5	74,47	75,52	76,04
Destek Vektör Rad.	72,91	7	75	75	75
Rastgele Orman	**73,43**	7	75,52	75,52	75,52
Lojistik Regresyon	72,39	7	75,52	**75,84**	76,04
Karar Ağacı	68,75	5	70,83	71,08	71,35
K en yakın Komşular	65,62	7	73,43	73,43	73,43

Yukarıdaki tabloda elde edilen en yüksek değerler koyu olarak gösterilmiştir. Özellik seçimi öncesi 8 özellikle yapılan sınıflandırmalarda başlangıç ortalama Doğruluk oranı %73,43 ile RO sınıflandırıcısı ile elde ediliyor. GA ile özellik seçiminden sonra ise %75,84 ortalama Doğruluk oranı ile LR ile elde ediliyor. Seçilen özellik sayısı sınıflandırıcıya göre değişmektedir. Örneği LR'de seçilen özellik sayısı 7'dir ve seçilen özellikler: 1) gebelik sayısı, 2) plazma glikoz konsantrasyonu, 3) diyastolik tansiyon, 4) kol kası cilt deri kalınlığı, 5) serum insülin, 6) vücut kütle endeksi, 7) diyabet soyağacı fonksiyonudur (True, True, True, True, True, True, True, False). 5 özellikle ortalama %75,52 Doğruluk sağlayan doğrusal çekirdekli DVM'de ise seçilen özellikler: 1) gebelik sayısı, 2) plazma glikoz konsantrasyonu, 6) vücut kütle endeksi, 7) diyabet soyağacı fonksiyonu, 8) yaş (True, True, False, False, False, True, True, True) şeklindedir.

GA ile özellik seçiminin sınıflandırma başarısına etkisine bakıldığında bütün sınıflandırıcılarda doğruluk oranının yükseldiği görülmektedir. Ortalama değerlere bakıldığında sınıflandırma performansının YSA'da yaklaşık %5, D-DVM'de yaklaşık %2,5, R-DVM'de yaklaşık %3, RO'da yaklaşık %2, LR'da yaklaşık %3, KA'da yaklaşık %3, K-YK'da yaklaşık %8'lik bir artış görülmektedir.

4.2. Wisconsin Göğüs Kanseri Veri Kümesi (Wisconsin Breast Cancer Dataset)

Wisconsin Göğüs Kanseri Veri kümesinde sınıflandırmada kullanılan 30 özellik ve toplamda 569 kayıt ile yapılan özellik seçimi uygulamasında elde edilen sonuçlar Tablo 3' de verilmiştir.

Tablo 3. Wisconsin göğüs kanseri veri kümesi ile elde edilen sonuçlar (Results obtained with the Wisconsin breast cancer dataset)

Sınıflandırma Yöntemi	Başlangıç Doğruluk (%)	Seçilen Özellik Sayısı	En Düşük Doğruluk (%)	Ortalama Doğruluk (%)	En Yüksek Doğruluk (%)
Yapay Sinir Ağları	86,01	20	96,50	97,90	98,60
Destek Vektör Doğ.	95,80	23	98,60	**99,30**	100
Destek Vektör Rad.	95,10	19	98,60	98,97	99,30
Rastgele Orman	**97,20**	17	98,60	**99,30**	100
Lojistik Regresyon	96,50	23	99,30	**99,30**	99,30
Karar Ağacı	93,00	19	96,50	97,90	98,60
K en yakın Komşular	96,50	21	97,20	97,20	97,20

Özellik seçimi öncesi 22 özellikle yapılan sınıflandırmalarda başlangıç ortalama Doğruluk oranı %97,20 ile RO sınıflandırıcısı ile elde ediliyor. GA ile özellik seçiminden sonra ise %99,30 en yüksek ortalama Doğruluk oranı RO, LR ve D-DVM ile elde ediliyor. Seçilen özellik sayısı yine sınıflandırıcıya göre değişmektedir. Örneğin D-DVM'de seçilen özellik sayısı 23, LR'de seçilen özellik sayısı 23 ve RO'da seçilen özellik sayısı 17'dir ve bu 3 sınıflandırıcıda ortalama %99,30 doğruluk elde etmişlerdir.

GA ile özellik seçiminin sınıflandırma başarısına etkisine bakıldığında bütün sınıflandırıcılarda doğruluk oranının yükseldiği görülmektedir. Ortalama değerlere bakıldığında sınıflandırma performansının YSA'da yaklaşık %11, D-DVM'de yaklaşık %4, R-DVM'de yaklaşık %3, RO'da yaklaşık %2, LR'da yaklaşık %3, KA'da yaklaşık %5, K-YK'da yaklaşık %1'lik bir artış görülmektedir.

4.3. Hepatit Veri Kümesi (Hepatitis Data Set)

Wisconsin Göğüs Kanseri Veri kümesinde sınıflandırmada kullanılan 19 özellik ve toplamda 155 kayıt ile yapılan özellik seçimi uygulamasında elde edilen sonuçlar Tablo 4' de verilmiştir.

Tablo 4. Hepatit veri kümesi ile elde edilen sonuçlar (Results obtained with the Hepatitis data set dataset)

Sınıflandırma Yöntemi	Başlangıç Doğruluk (%)	Seçilen Özellik Sayısı	En Düşük Doğruluk (%)	Ortalama Doğruluk (%)	En Yüksek Doğruluk (%)
Yapay Sinir Ağları	61,53	13	76,92	81,07	84,61
Destek Vektör Doğ.	**74,35**	15	79,48	79,48	79,48
Destek Vektör Rad.	69,23	13	71,79	74,35	76,92
Rastgele Orman	**74,35**	**13**	79,48	**82,35**	84,61
Lojistik Regresyon	**74,35**	13	79,48	81,15	82,05
Karar Ağacı	66,67	11	79,01	82,12	84,61
K en yakın Komşular	61,53	11	66,66	71,79	74,35

Özellik seçimi öncesi 19 özellikle yapılan sınıflandırmalarda başlangıç ortalama Doğruluk oranı %74,35 ile D-DVM, RO, LR sınıflandırıcıları ile elde ediliyor. GA ile özellik seçiminden sonra ise %82,35 ile en yüksek ortalama Doğruluk oranı RO'da elde ediliyor. Seçilen özellik sayısı yine sınıflandırıcıya göre değişmektedir. Örneğin en yüksek ortalama Doğruluk oranlarının elde edildiği sırasıyla RO' da seçilen özellik sayısı 13, LR'de seçilen özellik sayısı 13 ve KA'da seçilen özellik sayısı 11'dir.

GA ile özellik seçiminin sınıflandırma başarısına etkisine bakıldığında bütün sınıflandırıcılarda doğruluk oranının yükseldiği görülmektedir. Ortalama değerlere bakıldığında sınıflandırma performansının YSA'da yaklaşık %21, D-DVM'de yaklaşık %5, R-DVM'de yaklaşık %5, RO'da yaklaşık %8, LR'da yaklaşık %7, KA'da yaklaşık %16, K-YK'da yaklaşık %10'lik bir artış görülmektedir. Ayrıca özellik seçiminden sonra YSA'da en düşük sınıflandırma oranının %76,92 olmasına rağmen başlangıçtan çok daha yüksek olduğu görülmektedir. Aynı durum KA içinde geçerlidir.

4.4. Cleveland Kalp Hastalıkları Veri kümesi (Cleveland Heart Diseases Data Set)

Cleveland Kalp Hastalıkları Veri kümesinde sınıflandırmada kullanılan 13 özellik ve toplamda 1025 kayıt ile yapılan özellik seçimi uygulamasında elde edilen sonuçlar Tablo 5' de verilmiştir.

Özellik seçimi öncesi 13 özellikle yapılan sınıflandırmalarda başlangıç ortalama Doğruluk oranı %97,66 ile RO ve KA sınıflandırıcıları ile elde ediliyor. GA ile özellik seçiminden sonra ise %100 ortalama Doğruluk oranı ile RO ve KA ile elde ediliyor. Seçilen özellik sayısı yine sınıflandırıcıya göre değişmektedir. En yüksek ortalama Doğruluk oranlarının elde edildiği sırasıyla RO' da ve KA'da seçilen özellik sayısı 10'dur.

Tablo 5. Cleveland kalp hastalıkları veri kümesi ile elde edilen sonuçlar (Results obtained with the Cleveland heart diseases data set)

Sınıflandırma Yöntemi	Başlangıç Doğruluk (%)	Seçilen Özellik Sayısı	En Düşük Doğruluk (%)	Ortalama Doğruluk (%)	En Yüksek Doğruluk (%)
Yapay Sinir Ağları	81,71	10	84,04	87,93	94,16
Destek Vektör Doğ.	78,98	8	81,71	82,10	82,87
Destek Vektör Rad.	67,70	10	69,64	80,54	86,38
Rastgele Orman	**97,66**	10	100	**100**	100
Lojistik Regresyon	78,21	8	82,49	82,87	83,26
Karar Ağacı	**97,66**	10	100	**100**	100
K en yakın Komşular	71,98	8	82,87	85,21	88,32

GA ile özellik seçiminin sınıflandırma başarısına etkisine bakıldığında bütün sınıflandırıcılarda doğruluk oranının yine yükseldiği görülmektedir. Ortalama değerlere bakıldığında sınıflandırma performansının YSA'da yaklaşık %6, D-DVM'de yaklaşık %4, R-DVM'de yaklaşık %12, RO'da yaklaşık %2,5, LR'da yaklaşık %4, KA'da yaklaşık %2,5, K-YK'da yaklaşık %13'lük bir artış görülmektedir.

4.5. Parkinson Hastalığı Veri kümesi (Parkinson's Disease Data Set)

Parkinson Hastalığı Veri kümesinde sınıflandırmada kullanılan 22 özellik ve toplamda 195 kayıt ile yapılan özellik seçimi uygulamasında elde edilen sonuçlar Tablo 6' da verilmiştir.

Tablo 6. Parkinson hastalığı veri kümesi ile elde edilen sonuçlar (Results obtained with the Parkinson's disease data set)

Sınıflandırma Yöntemi	Başlangıç Doğruluk (%)	Seçilen Özellik Sayısı	En Düşük Doğruluk (%)	Ortalama Doğruluk (%)	En Yüksek Doğruluk (%)
Yapay Sinir Ağları	83,67	16	91,83	92,89	93,87
Destek Vektör Doğ.	87,75	13	93,87	93,87	93,87
Destek Vektör Rad.	83,67	16	85,71	86,52	89,79
Rastgele Orman	**91,83**	16	93,87	94,43	95,91
Lojistik Regresyon	89,79	14	91,83	91,83	91,83
Karar Ağacı	87,75	12	93,87	**95,91**	97,95
K en yakın Komşular	83,67	16	91,83	91,83	91,83
Ada Boost	85,71	16	93,87	94,26	95,91

Özellik seçimi öncesi 22 özellikle yapılan sınıflandırmalarda başlangıç ortalama Doğruluk oranı %91,83 ile RO sınıflandırıcı ile elde ediliyor. GA ile özellik seçiminden sonra ise %95,91 ortalama Doğruluk oranı ile KA ile elde ediliyor. Seçilen özellik sayısı yine sınıflandırıcıya göre değişmektedir. Örneğin en yüksek ortalama Doğruluk oranlarının elde edildiği sırasıyla KA' da seçilen özellik sayısı 12, RO'da seçilen özellik sayısı 16 ve D-DVM'de seçilen özellik sayısı 13'dür.

GA ile özellik seçiminin sınıflandırma başarısına etkisine bakıldığında bütün sınıflandırıcılarda doğruluk oranının yine yükseldiği görülmektedir. Ortalama değerlere bakıldığında sınıflandırma performansının YSA'da yaklaşık %9, D-DVM'de yaklaşık %6, R-DVM'de yaklaşık %4, RO'da yaklaşık %3, LR'da yaklaşık %2, KA'da yaklaşık %8, K-YK'da yaklaşık %8'lük bir artış görülmektedir.

5. Sonuç ve Tartışma (Result and Discussion)

Bu çalışmada, çok sayıda karmaşık verilere sahip olan tıbbi verilerden özellik seçimi yapmak ve en uygun özellik alt kümesini oluşturarak sınıflandırma başarısını artırmak için genetik algoritma (GA) içeren bir model önerilmiştir. Önerilen yöntemin performansını değerlendirmek için en çok bilinen ve rahatlıkla ulaşılabilen Çalışmada Kaliforniya Üniversitesi Irvine (UCI) makine öğrenme laboratuvarındaki tıbbi veri tabanlarından, Pima Indian diyabet, Wisconsin Göğüs Kanseri, Hepatitler, Cleveland Kalp Hastalıkları ve Parkinson veri kümesi olmak üzere 5 tıbbi veri kümesi ile uygulamalar yapılmıştır. Bu uygulamalarda Yapay Sinir Ağları (YSA), doğrusal ve radyal çekirdekli Destek Vektör Makineleri (DVM), Karar Ağacı (KA), Lojistik Regresyon (LR), Rastgele Ormanlar (RO), K en yakın Komşu (K-YK) olmak üzere 7 adet denetimli sınıflandırma yöntemi kullanılmıştır.

Yapılan uygulamalarda elde sonuçlar GA ile yapılan özellik seçimi ile yapılan sınıflandırmalarda Doğruluk oranının dolayısıyla sınıflandırma performansının veri kümesine bağlı olarak ortalama %2 ile %21 arasında artış sağladığını ortaya koymuştur. Çalışmalarda özellik sayısı arttıkça özellik seçiminin sağladığı performans artışının da yükseldiği görülmektedir. Farklı veri kümeleri ile yapılan çalışmada kategorik, sayısal ve karma veri türleri dahil olmak üzere farklı tıbbi veri türleri için genetik algoritmanın özellik seçiminde çok iyi performansı

sunmaktadır. Burada özellik seçimi ile özelliklerin azaltılmasıyla sınıflandırma hızının da azaldığı göz önüne alınarak yöntemin oldukça başarılı olduğu görülmektedir. Ayrıca yapılan çalışmalarda RO sınıflandırıcının bütün veri kümelerinde iyi sonuçlar verdiği görülmekte ve tıbbi veri kümelerindeki sınıflandırma başarısı ile öne çıkmaktadır.

Bundan sonra yapılacak çalışmalarda önerilen modelle derin öğrenme yöntemleri kullanılarak sınıflandırma başarısına yapacağı katkılar araştırılabilir.

Çıkar Çatışması (Conflict of Interest)

Yazar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. No conflict of interest was declared by the author.

Kaynaklar (References)

Aalaei, S., Shahraki, H., Rowhanimanesh, A., Eslami, S., 2016. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iranian journal of basic medical sciences, 19(5), 476.

Abdollahi, J., Nouri-Moghaddam, B., 2021. Feature selection for medical diagnosis: Evaluation for using a hybrid Stacked-Genetic approach in the diagnosis of heart disease. arXiv preprint arXiv:2103.08175. Booker, L. B., Goldberg, D. E., Holland, J. H., 1989. Classifier systems and genetic algorithms. Artificial intelligence, 40(1-3), 235-282.

Tutorials Point, 2016. Artificial Intelligence and Python, www.tutorialspoint.com.

Ba-Alwi, F. M., Hintaya. H. M., 2013. Comparative study for analysis the prognostic in hepatitis data: data mining approach. Spinal Cord, 11(12).

Chen, C. W., Tsai, Y. H., Chang, F. R., Lin, W. C., 2020. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems, 37(5), e12553.

Chiesa, M., Maioli, G., Colombo, G. I., & Piacentini, L. 2020. GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. BMC bioinformatics, 21(1), 1-11.

Deperlioglu, O., 2019. Classification of segmented phonocardiograms by convolutional neural networks. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 10(2), 5-13.

Ershadi, M. M., & Seifi, A. 2022. Applications of dynamic feature selection and clustering methods to medical diagnosis. Applied Soft Computing, 126, 109293.

Goldberg, D. E. 1989. Genetic algorithms in search. Optimization, and machine learning.

Habib, M., Aljarah, I., Faris, H., & Mirjalili, S. 2020. Multi-objective particle swarm optimization: theory, literature review, and application in feature selection for medical diagnosis. Evolutionary machine learning techniques, 175-201.

Jaganathan, P., Kuppuchamy, R., 2013. A threshold fuzzy entropy-based feature selection for medical database classification. Computers in Biology and Medicine, 43, 2222–2229.

Jothi, N., Husain, W., Rashid, N. A., Syed-Mohamad, S., 2019. Feature Selection Method using Genetic Algorithm for Medical Dataset. International Journal on Advanced Science Engineering Information Technology, 9(6), 1907-1912.

Kankanhalli, A., Hahn, J., Tan, S., Gao, G., 2016. Big data and analytics in healthcare: Introduction to the special section. Information Systems Frontiers, 18(2), 233-235.

Khadir, D.A., Amanullah, K.M. 2017. An Implementation of genetic algorithm-based feature selection approach over medical datasets.

Kumar, C. S., Thangaraju, P. 2021. Optimal Feature Subset Selection Method for Improving Classification Accuracy of Medical Datasets. Annals of the Romanian Society for Cell Biology, 3892-3913.

Little, M. A., McSharry, P. E., Hunter, E. J., Ramig L.O., 2008. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering.

Naheed, N., Shaheen, M., Khan, S. A., Alawairdhi, M., & Khan, M. A., 2020. Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review. Computer Modeling in Engineering & Sciences, 125(1), 314-344.

Nadimi-Shahraki, M. H., Banaie-Dezfouli, M., Zamani, H., Taghian, S., & Mirjalili, S. 2021. B-MFO: a binary moth-flame optimization for feature selection from medical datasets. Computers, 10(11), 136.

Machine Learning Notes, Jawaharlal Nehru Technological University, Kakinada, https://www.studocu.com/in/document/jawaharlal-nehru-technological-university-kakinada/computer-science-engineering/machine-learning-notes/17339474. Son erişim: 16.05.2022.

Mwadulo, M. W., 2016. A review on feature selection methods for classification tasks.

Parthiban, R., Usharani, S., Saravanan, D., Jayakumar, D., Palani, D. U., StalinDavid, D. D., & Raghuraman, D. (2021). Prognosis of chronic kidney disease (CKD) using hybrid filter wrapper embedded feature selection method. European Journal of Molecular & Clinical Medicine, 7(9), 2511-2530.

Rostami, M., Forouzandeh, S., Berahmand, K., & Soltani, M. (2020). Integration of multi-objective PSO based feature selection and node centrality for medical datasets. Genomics, 112(6), 4370-4384.

Rostami, M., Berahmand, K., Forouzandeh, S., 2021. A novel community detection based genetic algorithm for feature selection. Journal of Big Data, 8(1), 1-27.

Salama, G. I., Abdelhalim, M., Abd-elghany Zeid. M., 2012. Breast cancer diagnosis on three different datasets using multi-classifiers." Breast Cancer (WDBC) 32(569): 2.

Samant, R., Rao, S., 2013. A study on Feature Selection Methods in Medical Decision Support Systems. International Journal of Engineering Research & Technology (IJERT). 2(11), 615-619.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information processing & management, 45, 4, 427-437.

Tuba, E., Strumberger, I., Bezdan, T., Bacanin, N., Tuba, M., 2019. Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. Procedia Computer Science, 162, 307-315.

UCI Machine Learning Repository, 2007, https://archive.ics.uci.edu/ml/index.php, Irvine, CA: University of California, School of Information and Computer Science, Son erişim15 Mayıs 2022.

Wang, Y., Ma, L. (2009, January). Feature selection for medical dataset using rough set theory. In WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering (No. 3). World Scientific and Engineering Academy and Society.

Wirsansky, E., 2020. Hands-on genetic algorithms with Python: applying genetic algorithms to solve real-world deep learning and artificial intelligence problems. Packt Publishing Ltd.

Yeniterzi, S., Yeniterzi, R., Kücükural, A., & Sezerman, U., 2007. Feature selection with genetic algorithms on cardiac arrhythmia database. In the 2nd International Symposium on Health Informatics and Bioinformatics (HIBIT).