TİROİD VE KRONİK BÖBREK HASTALIĞI VERİLERİNİN SINIFLANDIRILMASINDA GENETİK ALGORİTMALAR VE PCA İLE HİBRİT ÖZELLİK SEÇİMİ

Ayşe Nagehan MAT[1], Onur İNAN[1]
[1]Necmettin Erbakan Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Konya Türkiye
nagehanmat@gmail.com (ORCID 0000-0003-4975-6418)
droinan@gmail.com (ORCID 0000-0003-4573-7025)

Özet

Bu çalışmada tiroid ve kronik böbrek hastalığının teşhisinde k-nearest neighbors sınıflandırıcının performansını arttırmak amacıyla genetik algoritmalar ve temel bileşenler analizi (PCA) hibrit şekilde kullanılmış ve yeni bir özellik seçimi yöntemi önerilmiştir. Hibrit özellik seçimi yönteminde elde edilen uygulama sonuçları, veri setlerinin özellik seçimi uygulanmamış başlangıç performansıyla karşılaştırılmıştır. Sonuç olarak önerilen hibrit metotla birlikte sınıflandırma başarısı tiroid veri seti için %93.44'ten %95.89'a, böbrek veri seti için %93.75'ten %98.25'e çıkarılmıştır. Sonuçların tutarlı olması için her iki veri setine 10-kat çapraz doğrulama yapılmıştır.

Anahtar Kelimeler: Genetik Algoritmalar, PCA, Özellik Seçimi, K-nearest neighbors

HYBRID FEATURE SELECTION USING GENETIC ALGORITHMS AND PCA IN CLASSIFICATION OF THYROID AND CHRONIC KIDNEY DISEASE DATA

Abstract

In this study, genetic algorithms and principal component analysis (PCA) were used in a hybrid way to increase the performance of the k-nearest neighbors classifier in the diagnosis of thyroid and chronic kidney disease, and a new feature selection method was proposed. The application results obtained in the hybrid feature selection method were compared with the initial performance of the data sets before the feature selection was applied. As a result, with the proposed hybrid method, the classification success

was increased from 93.44% to 95.89% for the thyroid data set and from 93.75%to 98.25%for the kidney data set. A 10-fold cross validation was applied to both data sets to ensure consistent results.

Keywords: Genetic Algorithms, PCA, Feature Selection, K-nearest neighbors

1. Giriş

Tiroid hastalığının erken teşhisi yaşam kalitesini arttırmak açısından çok önemlidir. Hastalık şüphesi olan kişilere yapılan bazı ölçüm ve anketler hastalığın teşhisinde büyük rol oynamaktadır. Tiroid hastalığının tehlikeli yönü, Tiroid hastalarının, kendilerinde bu hastalık olup olmadığını anlayamamalarıdır [1]. Bu nedenle, Tiroid hastalıklarının hızlı ve doğru şekilde tespit edilmesi büyük önem taşımaktadır. Pek çok tiroid problemi başarıyla tedavi edilebilmektedir. Tiroid fonksiyonlarının anormallikleri, genellikle çok az tiroid hormonu (hipotiroidizm) üretimi veya çok fazla tiroid hormonu (hipertiroidizm) üretimi ile ilgilidir [2 ve 3]. Tiroid bozukluklarının teşhisinde, klinik muayenenin yanı sıra tiroid verilerinin doğru şekilde yorumlanması da etkilidir. Tiroid tanısı önemli bir sınıflandırma problemidir. [2, 3 ve 4]. Hastalığın teşhisinde kalıp tanıma teknikleri, bulanık sınıflandırıcılar, yapay bağışıklık tanıma sistemi, sinir ağları vb. gibi çeşitli yeni yöntemler kullanılmıştır [5].

Kronik böbrek rahatsızlığı insan hayatını olumsuz etkileyen ve böbreklere zarar veren bir hastalıktır. Hastalık, azalan böbrek fonksiyonlarına bağlı olarak artan idrar yoğunluğu takip edilerek tespit edilebilmektedir. Kronik böbrek hastalığının tedavisi kan-damar rahatsızlıkları, tansiyon yüksekliği, anemi, kemik rahatsızlığı ve böbrek yetmezliği gibi diğer hastalıkları tetiklediği için hastalığa bağlı ölüm oranları artabilmektedir [6]. Ayrıca, hastalığın ilerleyen safhalarında diğer organlar etkilenmekte ve bir sağlık kuruluşuna bağlı olarak yaşayan hasta çeşitli sosyal problemler yaşamaktadır [7]. Literatürde kronik böbrek rahatsızlığı verilerinin sınıflandırılmasında sezgisel ve sezgisel olmayan tekniklerin kullanıldığı görülmüştür. Sınıflandırmada kullanılan sezgisel teknikler; yapay sinir ağları [8], naive bayes [9], radial tabanlı sinir ağları [10], destek vektör makinesi [11] olarak, sezgisel olmayan teknikler ise; rastgele orman [12], karar ağacı [13], kstar [13] ve k-nn [14] algoritması olarak karşımıza çıkmaktadır [15].

Genetik algoritma, (GA) optimizasyon yöntemi olarak kullanılan sezgisel bir algoritmadır. Doğal seçilim ve DNA kopyalanması gibi mekanizmaları taklit etmektedir. GA' da, çözüme rastgele bireylerle başlanır ve bir uygunluk fonksiyonu kullanılarak bireylerin performansı hesaplanır [16, 17]. Başlangıç aşamasında rastgele seçilen bireylerin optimum değere sahip olma ihtimali oldukça düşüktür. Bu nedenle en uygun değerli bireylerden oluşan popülasyon yapısını elde etmek için iteratif bir doğal seçim süreci kullanılır [18, 19]. Seçim sürecinde en iyi bireyler yeni jenerasyona aktarılır ve en kötü bireyler elenir [20, 21]. Bu evrimsel sürecin sonunda, kullanılan uygunluk fonksiyonu için en iyi adaptasyonu sağlayan değişken alt kümesi seçilmiş olur. GA, özellik seçimi için biyomedikal ve klinik veri setlerine uygulanmış ve başarılı sonuçlar elde edilmiştir [16, 22, 23].

Doğrusal dönüşüm tekniklerinden biri olan Temel Bileşenler Analizi (PCA) verideki gerekli bilgilerin çıkarılmasında oldukça etkili bir yöntemdir. PCA, eğitim, psikoloji, yüz tanıma, kalite kontrol, market araştırmaları, ekonomi, fotoğrafik bilimler, ziraat, haritacılık, görüntü sıkıştırma ve genetik gibi pek çok alanda kullanılan tekniklerden birisidir [24]. PCA' da asıl amaç, yüksek boyutlu verinin genel özelliklerini bularak boyut sayısını azaltmaktır. Azalan boyut sayısıyla bazı özelliklerin kaybedilmesi kaçınılmazdır. Ancak burada hedef, kaybolan özelliklerin veri popülasyonu hakkında çok az bilgi içermesidir. Genel olarak bu yöntemde yüksek korelasyonlu değişkenler bir araya getirilir. Verilerde en çok varyasyonu oluşturan ve "temel bileşenler" olarak adlandırılan daha az sayıda yapay bir değişken kümesi oluşturulur. Veri setlerindeki özellik sayısının fazla olması hasta verilerinin sınıflandırılmasını zorlaştırmakta, modeli karmaşık hale getirdiği için sınıflandırıcının performansını olumsuz etkilemektedir. Bu çalışmada, veri setlerinin sınıflandırılmasında kullanılan özellik sayısını azaltmak için Genetik Algoritmalar ve Temel Bileşenler Analizi (PCA) hibrit olarak kullanılmıştır. Şekil 1'de, önerilen metotta uygulanan aşamalar gösterilmektedir.

Şekil 1. Önerilen metot

1.1. Önceki çalışmalar

Sınıflandırma sistemleri, diğer klinik tanı problemleri için kullanıldığı gibi tiroid ve böbrek hastalıklarının teşhisi için de kullanılmaktadır. Bu bölümde, tiroid ve kronik böbrek hastalığı ile ilgili yapılan önceki çalışmalar ve sonuçlarına yer verilmiştir.

Tiroid veri kümesi için, Serpen ve diğ. [25] çalışmasında Multilayer Perceptron (MLP), Learning Vector Quantization (LVQ), Radial Basis Function (RBF) ve Probabilistic Potential Function Neural Network (PPFNN) algoritmalarını kullanmıştır. En yüksek sınıflandırma başarısı %81.86 değeriyle LVQ algoritmasında elde edilmiştir. Ozyilmaz ve Yildirim [2], MLP + bp (3 x FC), MLP + fbp (3 x FC), RBF (3 x FC) ve Adaptive Conic Section Function Neural Network (CSFNN) (3 x FC) kullanmış ve sırasıyla %86.33, %89.80, %79.08 ve %91.14 değerlerine ulaşmıştır. Pasi [26], Linear Discriminate Analysis (LDA), C4.5, MLP, DIMLP algoritmalarını kullanarak sınıflandırma yapmıştır. Çalışma sonuçlarına göre en iyi performans MLP algoritmasında %96.24 değeri olarak görülmektedir. Polat ve diğ. [3], aynı veri kümesi için bulanık ağırlıklı ön işleme içeren Artificial Immune Recognition System (AIRS) (10 x FC) ve AIRS (10 x FC) kullanmış olup %85.00 ve %81.00 sınıflandırma doğruluklarını elde etmiştir. Keles ve diğ. [27], Expert System for Thyroid Diseases Diagnosis (ESTDD) ile %95.33 (10 x FC) sonucunu elde etmiştir. Temurtas [28] çalışmasında, Multilayer Neural Network (MLNN) ve Levenberg-Marquardt (LM), Probabilistic Neural Network (PNN) ve LVQ algoritmalarını kullanarak (3 x FC) için sırasıyla %92.96, %94.43, %89.79 değerlerini, (10 x FC) için sırasıyla %93.19, %94.81, %90.05 değerlerini elde etmiştir. Esin Dogantekin ve diğ. [29], Generalized

Discriminant Analysis ve Wavelet Support Vector Machine System (GDA-WSVM) kullandığı çalışmasında %91.86, Chen ve diğ. [30] Fisher Score (FS) - Particle Swarm Optimization (PSO) - Support Vector Machine (SVM) sistemiyle %97.40 ve Li ve diğ. [31] Principle Component Analysis (PCA) ve Extreme Learning Machine (ELM) algoritmalarını kullanarak %98.10 (10 x FC) oranında başarılı bir sınıflandırma gerçekleştirmiştir. Shen ve diğ. [32], PSO-SVM, Grid search technique-SVM, Genetic Algorithm-SVM, Bacterial Forging Optimization (BFO)-SVM ve Fruit Fly Optimization Algorithm (FOA)-SVM kullanarak sınıflandırma yapmıştır. FOA-SVM %96.38 sınıflandırma başarısıyla çalışmanın en iyi sonucu olarak görülmektedir. Inbarani ve diğ. [33], Hybrid Rough-bijective Soft Set (RBISO), Bijective soft set theory (BISO), Decision Table (DT), Naive Bayes (NB) ve MLP algoritmalarını kullanmıştır. Sırasıyla %96.13 (10 x FC), %84.89 (10 x FC), 76.11 (10 x FC), %68.83 (10 x FC) ve 79.85 (10 x FC) doğruluk oranlarını elde etmiştir.

Böbrek veri kümesi için literatürde Chen ve diğ. [34], KNN, SVM ve Soft Independent Modeling of Class Analogy (SIMCA) algoritmalarını kullanarak sınıflandırma yapmıştır. En iyi sınıflandırma sonucu %99.7 olarak KNN ve SVM algoritmalarında elde edilmiştir. Polat ve diğ. [35], SVM sınıflandırıcısını farklı tekniklerle birlikte kullanmış ve en iyi performans SVM ile FilterSubsetEval ile Best First (10 x FC) algoritmasıyla %98.5 olarak bulunmuştur. Zhang ve diğ. [36], Fuzzy Rule-building Expert System (FuRES) (10 x FC) algoritmasıyla %99.6, Fuzzy Optimal Associative Memory (FOAM) (10 x FC) ile %98.0 ve Partial Least Squares Discriminant Analysis (PLS-DA) (10 x FC) ile %95.5 değerlerini elde etmiştir.

2. Materyal ve Metot

2.1. Veri Seti

Bu çalışmada UCI (University of California, Irvine) Machine Learning Repository veri tabanında bulunan “Thyroid Disease“ ve “Chronic Kidney Disease” veri setleri kullanılmıştır. Tablo 1’de tiroid, Tablo 2’de kronik böbrek veri setinde yer alan özellik tip ve aralıkları paylaşılmıştır.

Tiroid hastalığı veri setinde; normal, hipertiroid ve hipotiroid hastalarını temsil eden 215 veri bulunmaktadır. Verilerin sınıf dağılımı, normal 150 örnek, hipertiroid 35

örnek ve hipotiroid 30 örnek şeklindedir. Her bir veri 5 özellikten oluşmaktadır. Bu özellikler:

- Özellik 1: T3-resin uptake testi (yüzde olarak),
- Özellik 2: İzotopik deplasman yöntemiyle ölçülen toplam serum thyroxin miktarı (T4),
- Özellik 3: Radioimmuno assay yöntemiyle ölçülen toplam serum triiodothyronine (T3),
- Özellik 4: Radioimmuno assay yöntemiyle ölçülen bazal tiroit-uyarıcı hormonu (TSH),
- Özellik 5: 200 mg thyrotropin-releasing hormon enjeksiyonundan sonra TSH değerinin bazal değerle kıyaslandığında en yüksek mutlak fark değeri.

Tablo 1. Tiroid veri seti özellik tip ve aralıkları

Özellikler	Açıklama	Tip	Min	Max	Ortalama	Eksik Değer
Özellik 1	T3 (yüzde)	integer	65	144	109.595	Yok
Özellik 2	T4 (serum)	real	0.5	25.3	9.805	Yok
Özellik 3	T3 (serum)	real	0.2	10	2.050	Yok
Özellik 4	TSH	real	0.1	56.4	2.880	Yok
Özellik 5	TSH (fark değer)	real	-0.7	56.3	4.199	Yok

Kronik böbrek rahatsızlığı veri seti toplam 400 veriden oluşmaktadır. Veriler ckd (250 örnek) ve notckd (150 örnek) olmak üzere 2 sınıfla temsil edilmektedir. Her bir veri 24 özellikten oluşmaktadır. Böbrek veri setindeki eksik değerler için, ilgili özellikte en sık kullanılan değerler kullanılmıştır.

2.2. Genetik Algoritmalar ile Özellik Seçimi

İstatistik ve makine öğrenmesinde bir model hazırlanırken ilgili özelliklerden bir alt küme oluşturulmasına özellik seçimi denir. Özellik seçimi diğer adıyla değişken seçimi birçok farklı amaçla kullanılmaktadır. Bu amaçlardan bazıları; modellerin basitleştirilerek kullanıcı ve araştırmacılar tarafından yorumlanmasını kolaylaştırmak, aşırı öğrenmeyi sınırlayarak genellemeyi artırmak ve eğitim sürelerini kısaltmak olarak sıralanabilir [17, 20]. Özellik seçimindeki esas amaç, veri setindeki ilgisiz ya da gereksiz özelliklerin herhangi bir bilgi kaybı yaşanmadan elenmesidir [18]. Burada özellik seçim yöntemleri ile özellik çıkarımı karıştırılmamalıdır. Özellik seçiminde bir özellik alt kümesi oluşturulurken, özellik çıkarımında mevcut özelliklerden yeni özellikler üretilir [19, 21].

Tablo 2. Böbrek veri seti özellik tip ve aralıkları

Özellikler	Açıklama	Tip	Min	Max	Ortalama	Eksik Değer
Özellik 1	Yaş	integer	2	90	51.472	Var
Özellik 2	Kan basıncı	integer	50	180	76.455	Var
Özellik 3	Özgül ağırlık	real	1.005	1.025	1.017	Var
Özellik 4	Albümin	integer	0	5	1.015	Var
Özellik 5	Şeker	integer	0	5	0.395	Var
Özellik 6	Kırmızı kan hücreleri	binominal	normal	abnormal	-	Var
Özellik 7	İrin hücresi	binominal	normal	abnormal	-	Var
Özellik 8	İrin hücre kümeleri	binominal	present	notpresent	-	Var
Özellik 9	Bakteriler	binominal	present	notpresent	-	Var
Özellik 10	Kan şekeri	integer	22	490	148.032	Var
Özellik 11	Kandaki üre miktarı	integer	1.5	391	57.406	Var
Özellik 12	Serum kreatinin	real	0.4	76	3.072	Var
Özellik 13	Sodyum	integer	4.5	163	137.631	Var
Özellik 14	Potasyum	real	2.5	47	4.627	Var
Özellik 15	Hemoglobin	real	3.1	17.8	12.526	Var
Özellik 16	Paketlenmiş hücre hacmi	integer	9	54	38.905	Var
Özellik 17	Beyaz kan hücresi sayımı	integer	2200	26400	8406.090	Var
Özellik 18	Kırmızı kan hücresi sayısı	real	2.1	8	4.707	Var
Özellik 19	Hipertansiyon	binominal	yes	no	-	Var
Özellik 20	Şeker hastalığı	binominal	yes	no	-	Var
Özellik 21	Koroner arter hastalığı	binominal	yes	no	-	Var
Özellik 22	İştah	binominal	good	poor	-	Var
Özellik 23	Pedal ödem	binominal	yes	no	-	Var
Özellik 24	Kansızlık	binominal	yes	no	-	Var

Bir optimizasyon tekniği olarak Genetik Algoritmalar (GA), özellik seçiminde sıklıkla kullanılmaktadır. Farklı özellik vektörlerine bakarak en uygun özellikleri seçmek GA' nın çalışma prensibine oldukça uygundur [22, 23]. Burada GA' nın işlevi sınıflandırıcıdan gelen hata değerini minimize etmek şeklinde açıklanabilir. GA, akış diyagramı Şekil 2'de, sözde kodu Tablo 3'te gösterilmiştir.

GA' da tek bir çözümün optimize edilmesi yerine, kromozom olarak kodlanan aday çözümlerin oluşturduğu popülasyon kullanılır. Bu aday çözümler başlangıç aşamasında genellikle rastgele oluşturulur ve eşit boyutlu vektörleri temsil eder. Her nesilde genetik işlemler uygulanarak başlangıç popülasyonu geliştirilir.

Çözümün ilk aşamasında, amaç fonksiyonundan elde edilen sonuçlar yeterli olurken, ilerleyen aşamalarda daha iyi çözümler ile iyi çözümler arasındaki farkı ayırt etmek zorlaşmaktadır. Bu durumda algoritma sadece belli bir bölgeye odaklanacak ve arama uzayının tümünü taramayacaktır. Bu yerel optimuma takılma durumu erken yakınsama (premature convergence) olarak adlandırılmaktadır. Algoritmanın global bir arama yapabilmesi araştırmaya yeteri kadar ıraksama sağlanmasıyla mümkündür. GA' da bu amaçla bazı kromozomlara mutasyon ve çaprazlama gibi operatörler uygulanır.

Çaprazlama işleminde, mevcut popülasyondan iki bireye (ebeveyn) bilgi değişimi (gen takası) yapılarak iki yeni birey (çocuk) elde edilir. Bunun için eşleştirme havuzundan seçilen iki bireyin belirli gen dizileri yer değiştirilir. Mutasyon işlemi, oluşan bu yeni kromozomlardaki bazı genlerin olasılık tabanlı olarak değiştirilmesidir. Oldukça küçük bir değeri olan mutasyon oranı kullanılarak mutasyona uğrayacak genler rastgele olarak belirlenir. Mutasyon işlemi sayesinde popülasyona yeni bilgiler eklendiği için arama uzayındaki farklı bölgelerin taranması sağlanır ve erken yakınsama probleminin çözümü kolaylaşır.

Yeni neslin bireyleri oluşturulurken farklı yöntemler kullanılmaktadır. Örneğin, Rulet tekerleği yönteminde, bir kromozomun uygunluk değerinin popülasyondaki bütün bireylerin uygunluk değerlerinin toplamına oranı bulunur. Böylece her birey için seçilme olasılığını temsil eden [0,1] aralığında bir değer elde edilir. Bu hesaplamadan sonra her birey Rulet tekerleğine seçilme olasılıklarına göre yerleştirilir ve rastgele üretilen bir sayıya göre birey seçimi yapılır. Bu yöntemde yüksek uygunluk değerine sahip bireylerin seçilme ihtimali daha fazladır.

Güçlü olan hayatta kalır ilkesine uygun şekilde, her nesilde en zayıf bireyler popülasyondan çıkarılır ve ilerleyen her nesille birlikte problem çözümüne yaklaşılmış olur. Sonuncu nesilde ulaşılan en iyi birey optimum çözüm olarak kabul edilir. Bu birey kesin olarak optimum çözüm değilse de optimuma en yakın çözümdür.

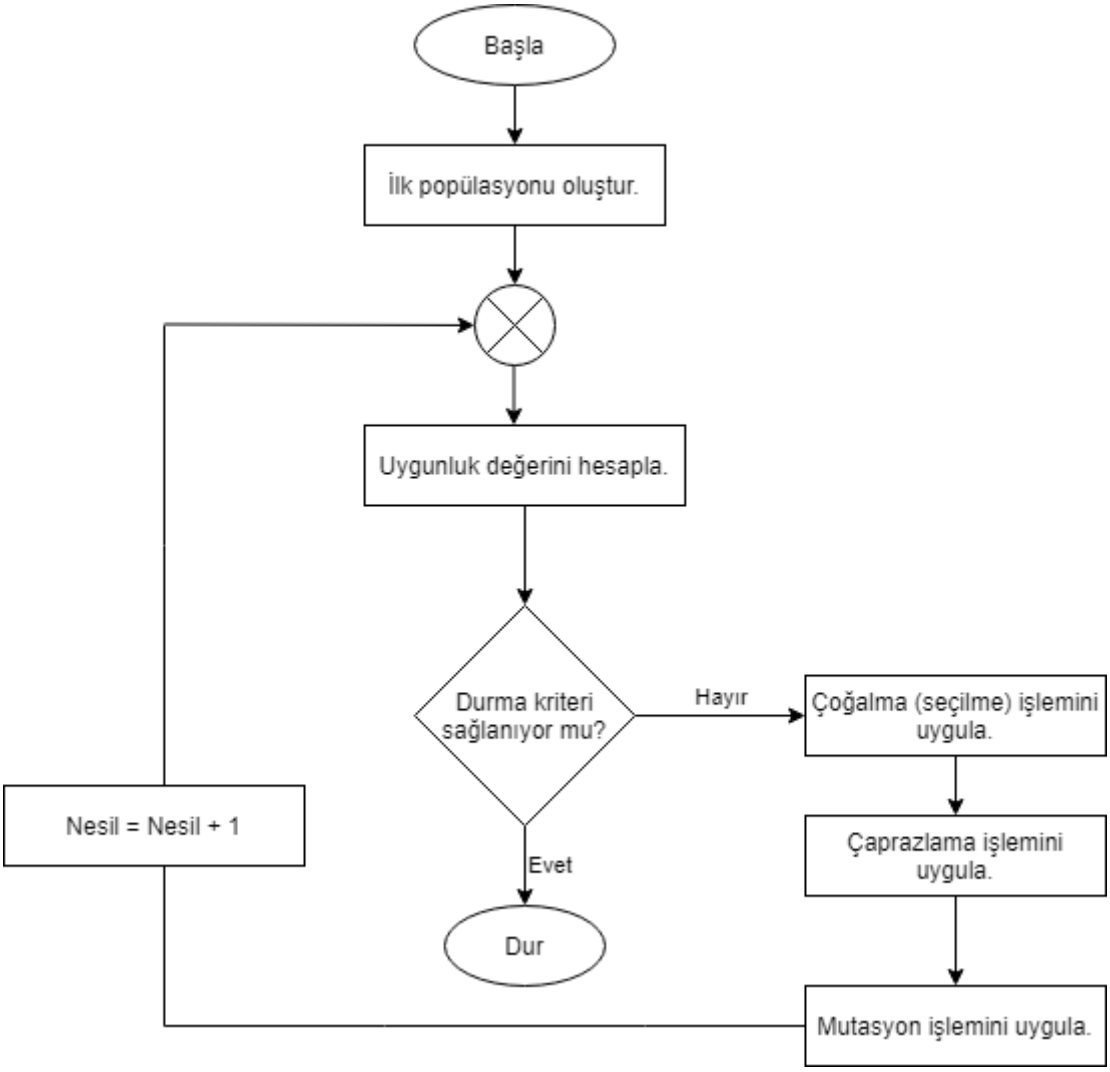
Tablo 3. GA sözde kod

```
function GENETIK_ALGORITMA (populasyon, UYGUNLUK_FN) return birey
    input : populasyon, bir dizi birey
        UYGUNLUK_FN, bireyin kalitesini belirleyen fonksiyon
    repeat
        yeni_populasyon ← boş küme
        loop for i from 1 to BOYUT (populasyon) do
            x ← RASTGELE_SECIM (populasyon, UYGUNLUK_FN)
            y ← RASTGELE_SECIM (populasyon, UYGUNLUK_FN)
            cocuk ← YENIDEN_URET (x,y)
            if (rastgele olasılıktan küçükse) then cocuk ← MUTASYON (cocuk)
            cocuk yeni_populasyona eklenir
        populasyon ← yeni_populasyon
    bir birey uygun değere ulaşana / yeterli zaman geçene kadar devam et
    en iyi bireyi döndür
```

2.3. Temel Bileşenler Analizi (PCA)

PCA, çok değişkenli veri setlerinde bilgiyi daha az değişkenle ve minimum kayıpla temsil etmek için geliştirilmiş matematiksel bir tekniktir. Genellikle veri kümelerini sadeleştirmek ve boyut azaltmak amacıyla kullanılmaktadır. İlk etapta birden fazla boyutun birbiriyle ilişkili olup olmadığının tespiti yapılır. Böylece aralarında bağlantı olan iki bilgiden birini ve aralarındaki bağlantıyı tutmak iki bilgiye de ulaşılmasını sağlamaktadır. PCA, veri sadeleştirmenin yanı sıra verilerin birbiri ile olan ilişkisini ve sonuca olan etki ağırlıklarını hesaplamada da kullanılmaktadır.

PCA metodundaki birinci amaç, değişkenler arasındaki bağımlılık yapısının ortadan kaldırılması, ikincisi amaç ise yüksek boyuttaki verinin indirgenmesidir. Veri çeşitliliğini daha iyi yakalayabilecek yeni boyut takımı bulunurken ilk boyut mümkün olan çok çeşitliliği temsil edecek şekilde belirlenir. 2. Boyut ilk boyuta dikey olacak ve yine mümkün olan çok çeşitliliği temsil edecek şekilde belirlenir [37].



Şekil 2. Genetik algoritmalar akış diyagramı

Boyut indirgeme işlemi, aralarında korelasyon olan değişkenleri bazı linear dönüşümler kullanarak, aynı sayıda ve korelasyon içermeyen değişkenlere dönüştürmek olarak ifade edilebilir. Elde edilen bu yeni değişkenler "temel bileşenler" olarak adlandırılır. PCA tekniği, değişkenler arası yüksek korelasyona sahip veri setlerinde oldukça işe yaramaktadır.

PCA analizinin temelinde, değişkenler arası kovaryans veya korelasyon matrisinin spektral özellikleri vardır. Yapısı gereği pozitif ve simetrik olan bu matrisin özdeğerleri pozitiftir ve verilerin varyansları ile özdeştir. Başka bir deyişle PCA, verisetlerinin kovaryans veya korelasyon matrislerinin özdeğerlerini ve özvektörlerini bulma problemidir.

PCA tekniğinin uygulaması genel olarak 5 temel adımdan oluşur:

1. Veriyi merkezleme

2. Kovaryans/korelasyon matrisini oluşturma

3. Kovaryans/korelasyon matrisinin özdeğerlerini ve özvektörlerini hesaplama

4. Temel bileşenleri seçme

5. Yeni veri setini hesaplama

2.4. K-Nearest Neighbors (K-NN) Sınıflandırıcı

Bu çalışmada tiroid ve böbrek hastalığı verilerinin sınıflandırılmasında örnek tabanlı bir öğrenme algoritması olan k-nearest neighbor (K-NN) kullanılmıştır. K-NN makine öğrenme algoritmaları içerisinde en çok bilinen ve kullanılan algoritmalardan biridir. Sınıflandırma ve regresyonda kullanılan parametrik olmayan bir yöntem olan K-NN sınıflamada çıktı bir sınıf üyeliğidir. Bir nesne komşularının çoğunluk oyuyla bir sınıfa atanır [38, 39]. Yani temel olarak nesne, diğer nesnelerle arasındaki yakınlığa göre sınıflandırılır. Nesneler arasındaki mesafe Denklem 1'deki formül kullanılarak hesaplanmaktadır [40].

$$d_{(i,j)} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2} \quad (1)$$

Algoritmadaki k değeri eşitlik durumlarına karşı genellikle tek sayı seçilir. Sınıflandırılacak yeni nesne ile diğer nesneler arasındaki mesafe hesaplanırken Kosinüs, Öklid veya Manhattan uzaklığı gibi yöntemler kullanılır [40].

2.5. 10-Kat Çapraz Doğrulama

Çalışmada, rastgele oluşturulan eğitim ve test verilerinden elde edilen sonuçların tutarlılığını sağlamak amacıyla k-katlı çapraz doğrulama kullanılmıştır. K-katlı çapraz doğrulamada, tüm veriler k boyutunda rastgele oluşturulmuş eşit alt kümelere bölünür. Her durumda, katlardan biri test verisi olarak alınır ve kalan katlar, eğitim kümesi oluşturulmak üzere birleştirilir. Elde edilen doğruluk oranlarının ortalaması, algoritma doğruluk oranını ifade etmektedir. Bu çalışmada kullanılan sınıflandırma algoritması, 10 kez eğitilmiş ve test edilmiştir.

3. Uygulama Sonuçları ve Tartışma

Bu çalışmada önerilen hibrit özellik seçim yöntemi tiroid ve böbrek veri setlerine uygulanmıştır. Uygulama esnasında GA ve PCA hibrit olarak kullanılarak özellik sayısı azaltılmış bir veri seti oluşturulmuştur. Sınıflandırıcıya, bu veri seti giriş olarak verilmiş ve eğitilen sınıflandırıcının başarısı cross-validation (çapraz doğrulama) işlemi ile hesaplanmıştır. Sınıflandırıcının hibrit özellik seçimindeki başarısı, veri setinin başlangıç durumundaki ve sadece GA kullanılarak özellik seçimi yapılmış haliyle karşılaştırılmıştır. Tablo 4'te GA kontrol parametreleri, Tablo 5'te tiroid ve böbrek veri setine ait sınıflandırma doğruluk oranları görülmektedir.

Tablo 4. GA parametre değerleri

Parametre	**Tiroid**	**Böbrek**
Popülasyon Büyüklüğü	20	30
Seçim Fonksiyonu	Rulet Çemberi	Turnuva Seçimi
Mutasyon Oranı	-1.0	-1.0
Çaprazlama Oranı	0.5	0.5
Çaprazlama Şekli	Tek Noktalı	Karma

Tablo 5. Tiroid ve böbrek veri seti sınıflandırma doğruluk oranları

Veri Seti	**Metod**	**Sınıflandırma Doğruluk Oranı**
Tiroid	KNN	93.44 (10-fold-CV)
	GA-KNN	95.43 (10-fold-CV)
	GA-PCA-KNN	95.89 (10-fold-CV)
Böbrek	KNN	93.75 (10-fold-CV)
	GA-KNN	94.75 (10-fold-CV)
	GA-PCA-KNN	98.25 (10-fold-CV)

Tiroid veri setinde K-NN sınıflandırıcının başlangıç performansı %93.44, GA ile özellik seçimi yapılmış veri seti için sınıflandırma peformansı %95.43 olarak hesaplanmıştır. Uygulanan hibrit metoda göre özelliklerin sınıflandırma üzerindeki etkisine bakılarak Özellik 1 ve Özellik 3 veri setinden çıkarılmıştır. Özellik 2, Özellik 4 ve Özellik 5 kullanılarak K-NN sınıflandırıcının performansı yeniden hesaplanmış ve sınıflandırma başarısının %95.89'a yükseldiği görülmüştür.

Böbrek veri seti için başlangıç performansı %93.75, GA uygulanmış veri setinde %94.75 olarak hesaplanmıştır. Hibrit metoda göre 4, 5, 6, 7, 15, 18, 19, 22 ve 24 nolu özellikler seçilmiş, geri kalan 15 özellik elenmiştir. Seçilen 9 özellik ile sınıflandırıcı performansı büyük ölçüde iyileştirilmiş ve %98.25'e çıkarılmıştır. Özellik

seçimi sayesinde her iki veri setinde daha yüksek sınıflandırma doğrulukları elde edilmiştir.

Tablo 6'da tiroid, Tablo 7'de böbrek veri seti için sınıflandırma sonuçları detaylı olarak verilmiştir. Her iki veri seti için F-Score, Sensitivity, Specificity, Precision, Recall ve Accuracy değerleri hesaplanmıştır.

Tablo 6. Tiroid veri seti ayrıntılı doğruluk oranları

Metod	**F-Score**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**Accuracy**
KNN	90.75	86.50	93.28	95.80	86.50	93.44
GA-KNN	94.52	93.00	96.32	96.60	93.0	95.43
GA-PCA-KNN	94.85	93.50	96.59	96.66	93.50	95.89

Tablo 7. Böbrek veri seti ayrıntılı doğruluk oranları

Metod	**F-Score**	**Sensitivity**	**Specificity**	**Precision**	**Recall**	**Accuracy**
KNN	94.74	90.0	100	100	90.0	93.75
GA-KNN	95.62	91.60	100	100	91.60	94.75
GA-PCA-KNN	98.61	99.20	96.67	98.02	99.20	98.25

Tablo 8'de tiroid veri seti için yapılan önceki çalışmalar ve mevcut çalışmanın sonuçları karşılaştırılmaktadır. Tiroid veri seti kullanılan çalışmalarda, genellikle YSA algoritmaları ile sınıflandırma yapıldığı görülmektedir. Li ve diğ. [31], PCA-ELM (10xFC) algoritmasının %98.1'lik başarı oranıyla önceki çalışmalar arasında en iyi sonuç olduğu tespit edilmiştir.

Tablo 9'da böbrek veri seti için yapılan önceki çalışmalar ve mevcut çalışmanın sonuçları karşılaştırılmaktadır. Zhang ve diğ.[36], FuRES algoritmasını kullanarak en iyi sınıflandırma sonucunu %99.6 olarak bulmuştur. Bu sonucun mevcut çalışmada elde edilen %98.25'lik sınıflandırma başarısıyla oldukça yakın olduğu gözlemlenmiştir.

Tablo 8. Tiroid veri seti sınıflandırma sonuçları

Çalışma	**Metod**	**Sınıflandırma Başarısı (%)**
Serpen ve diğ.(1997) [25]	MLP	36.74 (test data)
	LVQ	81.86 (test data)
	RBF	72.09 (test data)
	PPFNN	78.14 (test data)
Ozyilmaz and Yildirim (2002) [2]	MLP ile back-propagation	86.33 (average-3-fold-CV)
	MLP ile fast back-propagation	89.80 (average-3-fold-CV)
	RBF	79.08
	CSFNN	91.14
Pasi (2004) [26]	LDA	81.34 (test data)
	C4.5-1	93.26 (test data)
	C4.5-2	92.81 (test data)
	C4.5-3	92.94 (test data)
	MLP	96.24 (test data)
	DIMLP	94.86 (test data)
Polat ve diğ. (2007) [3]	AIRS	81.00 (average-10-fold-CV)
	AIRS ile Fuzzy weighted pre-processing	85.00 (average-3-fold-CV)
Keles ve diğ. (2008) [27]	ESTDD	95.33 (10-fold-CV)
Temurtas (2009) [28]	MLNN ile LM	92.96 (3-fold-CV)
	PNN	94.43 (3-fold-CV)
	LVQ	89.79 (3-fold-CV)
	MLNN ile LM	93.19 (10-fold-CV)
	PNN	94.81 (10-fold-CV)
	LVQ	90.05 (10-fold-CV)
Esin Dogantekin ve diğ. (2011) [29]	GDA-WSVM	91.86 (test data)
Chen ve diğ. (2011) [30]	FS-PSO-SVM	97.40 (average-10-fold-CV)
Li ve diğ. (2012) [31]	PCA-ELM	97.73 (average-10-fold-CV)
		98.10 (10-fold-CV)
Shen ve diğ. (2016) [32]	PSO-SVM	0.9526 ± 0.0080
	Grid-SVM	0.9499 ± 0.0092
	GA-SVM	0.9594 ± 0.0106
	BFO-SVM	0.9440 ± 0.0082
	FOA-SVM	0.9638 ± 0.0062
Inbarani ve diğ. (2016) [33]	RBISO	96.13 (10-fold-CV)
	BISO	84.89 (10-fold-CV)
	DT	76.11 (10-fold-CV)
	NB	68.83 (10-fold-CV)
	MLP	79.85 (10-fold-CV)
Mevcut Çalışma	GA-KNN	95.43 (10-fold-CV)
	GA-PCA-KNN	95.89 (10-fold-CV)

Tablo 9. Böbrek veri seti sınıflandırma sonuçları

Çalışma	**Metod**	**Sınıflandırma Başarısı (%)**
Chen ve diğ. (2016) [34]	KNN	99.7 ± 0.1
	SVM	99.7 ± 0.2
	SIMCA	93.5 ± 1.0
Polat ve diğ. (2017) [35]	SVM	97.75 (10xFC)
	SVM ile CfsSubsetEval ile Greedy stepwise	98 (10xFC)
	SVM ile WrapperSubsetEval ile Best First	98.25 (10xFC)
	SVM ile CfsSubsetEval ile Greedy stepwise	98.25 (10xFC)
	SVM ile FilterSubsetEval ile Best First	98.5 (10xFC)
Zhang ve diğ. (2016) [36]	FuRES	99.6 ± 0.2 (10xFC)
	FOAM	98.0 ± 0.7 (10xFC)
	PLS-DA	95.5 ± 0.6 (10xFC)
Mevcut Çalışma	GA-KNN	94.75 (10xFC)
	GA-PCA-KNN	98.25 (10xFC)

4. Sonuç

Hekimin olmadığı veya yetersiz kaldığı durumlarda hastalığa tanı koymak için geliştirilecek uzman sistemlerin insan hayatı açısından önem arz ettiği bilinmektedir. Tıp alanındaki teşhislerde daha doğru ve kesin sonuçlar elde etmek, insan kaynaklı hataları en aza indirmek ve hekime yardımcı olmak adına uzman sistemler tasarlanmaktadır.

Bu çalışmada, genetik algoritmalar ve temel bileşenler analizi hibrit bir şekilde kullanılarak özellik seçimi yapılmıştır. Böylece, tiroid ve kronik böbrek hastalığı teşhisine yeni bir yaklaşım sunularak az sayıda özellik ve düşük boyutla yüksek sınıflandırıcı başarısı elde edilmiştir. Önerilen hibrit özellik seçimi tekniğiyle, tiroid veri setinde K-NN sınıflandırıcının performansı %93.44'ten %95.89'a, kronik böbrek veri setinde ise, sınıflandırıcının performansı %93.75'ten %98.25'e çıkarılmıştır. Her iki veri setinde sınıflandırıcı doğruluk oranlarının tutarlı olması için 10-kat çapraz doğrulama yöntemi kullanılmıştır. Bu çalışmaya başka sınıflandırma algoritmaları dâhil edilerek performanslarının karşılaştırılması, sınıflandırma algoritmalarının problem üzerindeki etkinlik analizinin yapılması açısından faydalı olacaktır.

Kaynaklar

[1] Zhang G, Berardi V. An investigation of neural networks in thyroid function diagnosis. Health Care Management Science 1998; 1.1: 29-37.

[2] Ozyilmaz, L., Yildirim, T., 2002, Diagnosis of thyroid disease using artificial neural network methods. In Neural Information Processing, *Proceedings of the 9th International Conference -ICONIP'02*, 2033-2036.

[3] Polat K, Şahan S, Güneş S. A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. Expert Systems with Applications 2007; 32.4: 1141-1147.

[4] Hoshi K, Kawakami J, Kumagai M, Kasahara S, Nishimura N, Nakamura H, Sato K. An analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks. Chemical and pharmaceutical bulletin 2005; 53.12: 1570-1574.

[5] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine 2005; 34.2: 113-127.

[6] Go A, Chertow G, Fan D, McCulloch C, Hsu C. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. New England Journal of Medicine 2004; 35.13: 1296-1305.

[7] Topbaş E. Kronik Böbrek Hastalığının Önemi, Evreleri Ve Evrelere Özgü Bakımı, Nefroloji Hemşireliği Dergisi 2015; 53-59.

[8] Jena L, Kamila K. Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease, International Journal of Emerging Research in Management &Technology 2015; 93594: 2278–9359.

[9] Kunwar, V., Chandel, K., Sabitha, A. S., Bansal, A., 2016, Chronic Kidney Disease analysis using data mining classification techniques, *In IEEE 6th International Conference Cloud System and Big Data Engineering (Confluence)*, 300-305.

[10] İlkuçar M. Kronik Böbrek Hastalarının Yapay Sinir Ağı ve Radyal Temelli Fonksiyon Ağı ile Teşhisi. Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2015; 6.2: 82-88.

[11] Polat H, Mehr H, Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, Journal of Medical Systems 2017; 41.4: 55.

[12] Kumar M. Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm Running Title: Prediction of Chronic Kidney Disease, International Journal of Computer Science and Mobile Computing 2016; 52522: 24–33.

[13] Baby P, Vital P. Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms, International Journal of Engineering Research & Technology 2015; 407: 206–210.

[14] Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM, International Journal of Engineering Research and Technology 2015; 4.12: 608-12.

[15] Kılıçarslan S, Çelik M. (2019). Rotasyon orman sınıflandırma algoritması kullanarak kronik böbrek rahatsızlığının tahmini.

[16] Ghaheri A, Shoar S, Naderan M, Hoseini S. The applications of genetic algorithms in medicine. Oman medical journal 2015; 30.6: 406.

[17] Welikala R, Fraz M, Dehmeshki J, Hoppe A, Tah V, Mann S, Barman S. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. Computerized Medical Imaging and Graphics 2015; 43: 64-77.

[18] Khan A, Baig A. Multi-objective feature subset selection using non-dominated sorting genetic algorithm. Journal of applied research and technology 2015; 13.1: 145-159.

[19] Xu L, Redman C, Payne S, Georgieva A. Feature selection using genetic algorithms for fetal heart rate analysis. Physiological measurement 2014; 35.7: 1357.

[20] Singh D, Leavline E, Priyanka R, Priya P. Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. International Journal of Intelligent Systems and Applications 2016; 8.1: 67.

[21] Erguzel T, Ozekes S, Tan O, Gultekin S. Feature selection and classification of electroencephalographic signals: an artificial neural network and genetic algorithm based approach. Clinical EEG and neuroscience 2015; 46.4: 321-326.

[22] Cerrada M, Sánchez R, Cabrera D, Zurita G, Li C. Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal. Sensors 2015; 15.9: 23903-23926.

[23] Hsu W. Improving classification accuracy of motor imagery EEG using genetic feature selection. Clinical EEG and neuroscience 2014; 45.3: 163-168.

[24] Çilli M, Arıtan S. Temel Bileşenler Analizi Yardımı İle Elde Edilen Daha Az Sayıda Değişken Kullanılarak Farklı Hızlarda İnsan Koşusunun Fourier Tabanlı Modelinin Oluşturulması. Spor Bilimleri Dergisi 2010; 21.1: 1-12.

[25] Serpen, G., Jiang, H., Allred, L., 1997, Performance analysis of probabilistic potential function neural network classifier, *In Proceedings of artificial neural networks in engineering conference*, St. Louis-United States, 471-476.

[26] Pasi, L., 2004, Similarity classifier applied to medical data sets, *In International conference on soft computing- Fuzziness in Finland'04*, Helsinki-Finland.

[27] Keleş A, Keleş A. ESTDD: Expert system for thyroid diseases diagnosis. Expert Systems with Applications 2008; 34.1: 242-246.

[28] Temurtas F. A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications 2009; 36.1: 944-949.

[29] Dogantekin E, Dogantekin A, Avci D. An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases. Expert Systems with Applications 2011;, 38.1: 146-150.

[30] Chen H, Yang B, Wang G, Liu J, Chen Y, Liu, D. A three-stage expert system based on support vector machines for thyroid disease diagnosis. Journal of medical systems 2012; 36.3: 1953-1963.

[31] Li L, Ouyang J, Chen H, Liu D. A computer aided diagnosis system for thyroid disease using extreme learning machine. Journal of medical systems 2012; 36.5: 3327-3337.

[32] Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, Liu D. Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Systems 2016; 96: 61-75.

[33] Inbarani H, Kumar S, Azar A, Hassanien A. Hybrid rough-bijective soft set classification system. Neural Computing and Applications 2018; 29.8: 67-78.

[34] Chen Z, Zhang X, Zhang Z. Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. International urology and nephrology 2016; 48.12: 2069-2075.

[35] Polat H, Mehr H, Cetin A. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of medical systems 2017; 41.4: 55.

[36] Zhang Z, Chen Z, Zhu R, Xiang, Y, Harrington P. Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers. Chemometrics and Intelligent Laboratory Systems 2016; 153: 140-145.

[37] Yıldız K, Çamurcu Y, Doğan, B. Veri madenciliğinde temel bileşenler analizi ve Negatifsiz matris çarpanlarına ayırma tekniklerinin karşılaştırmalı analizi. Akademik Bilişim 2010; 10-12.

[38] Hilda, G. T., Rajalaxmi, R. R., 2015, Effective feature selection for supervised learning using genetic algorithm, *2nd International Conference In Electronics and Communication System-ICECS 2015*, 909-914.

[39] Baur B, Bozdag S. A Feature Selection Algorithm to Compute Gene Centric Methylation from Probe Level Methylation Data. PloS one 2016; 11.2: e0148977.

[40] Kuang, Q., Zhao, L., 2009, A practical GPU based kNN algorithm. In Proceedings, *The 2009 International Symposium on Computer Science and Computational Technology-ISCSCI 2009*, Huangshan-China, 151.