BŞEÜ Fen Bilimleri Dergisi
7(2), 1106-1123, 2020

BSEU Journal of Science
DOI: 10.35193/bseufbd.742456

e-ISSN: 2458-7575 (http://dergipark.gov.tr/bseufbd)

Araştırma Makalesi - Research Article

Kanser Teşhisi için Makine Öğrenmesi Tekniklerine Dayalı Yeni Bir Sınıflandırma Metodu

Can Eyüpoğlu[1*], Erdem Yavuz[2]

Geliş / Received: 25/05/2020 ***Revize / Revised: 24/06/2020*** ***Kabul / Accepted: 29/06/2020***

ÖZ

İnsan ölümlerinin en büyük nedenlerinden biri kanserdir. Kadınlar arasındaki kanser ölümlerinin başlıca sebebi ise meme kanseridir. Bu kanser türü sebebiyle yaşanan ölümleri azaltmanın yolu erken teşhistir. Uzman sistemler, yapay zeka ve makine öğrenmesi tekniklerinin tıp alanında kullanılmasının temel amaçlarından biri hastalıkları erken teşhis etmede doktorlara yardımcı olmaktır. Kanser türleri arasında özellikle meme kanserinde erken teşhis sayesinde ölüm riski büyük oranda düşürülebilir. Bu çalışmada temel bileşen analizi (Principal Component Analysis-PCA) ve ileri beslemeli sinir ağı (Feed Forward Neural Network-FFNN) temelli yeni bir kanser teşhisi yöntemi önerilmiştir. Önerilen yöntemin performansı Meme Kanseri Coimbra Veri Seti (Breast Cancer Coimbra Dataset-BCCD) üzerinde sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü metrikleri ile test edilmiştir. Ayrıca önerilen yöntemin klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile ayrıntılı olarak karşılaştırmalı performans analizi yapılmıştır. Deneysel sonuçlar önerilen yöntemin etkin olduğunu ve erken teşhis için doktorlar tarafından kullanılabileceğini göstermektedir.

Anahtar Kelimeler - Kanser Teşhisi, Meme Kanseri, Makine Öğrenmesi, Temel Bileşen Analizi, İleri Beslemeli Sinir Ağı

[1]*Sorumlu yazar iletişim: ceyupoglu@hho.edu.tr, caneyupoglu@gmail.com (https://orcid.org/0000-0002-6133-8617)
Bilgisayar Mühendisliği Bölümü, Hava Harp Okulu, Milli Savunma Üniversitesi, İstanbul, Türkiye
[2]İletişim: erdem.yavuz@btu.edu.tr, erdemyavuz29@gmail.com (https://orcid.org/0000-0002-3159-2497)
Bilgisayar Mühendisliği Bölümü, Mühendislik ve Doğa Bilimleri Fakültesi, Bursa Teknik Üniversitesi, Bursa, Türkiye

BŞEÜ Fen Bilimleri Dergisi
7(2), 1106-1123, 2020

BSEU Journal of Science
DOI: 10.35193/bseufbd.742456

e-ISSN: 2458-7575 (http://dergipark.gov.tr/bseufbd)

A New Classification Method Based on Machine Learning Techniques for Cancer Diagnosis

ABSTRACT

One of the major causes of human death is cancer. Breast cancer is the main reason for cancer deaths among women. Early diagnosis is the way to reduce deaths due to this cancer type. One of the main objectives of the use of expert systems, artificial intelligence and machine learning techniques in medicine is to assist doctors in early diagnosis of diseases. Among cancer types, the risk of death can be greatly reduced by early diagnosis, especially in breast cancer. In this study, a new cancer diagnosis method based on Principal Component Analysis (PCA) and Feed Forward Neural Network (FFNN) has been proposed. The performance of the proposed method is tested on the Breast Cancer Coimbra Dataset (BCCD) with classification accuracy, precision, recall and F-measure metrics. Besides, the comparative performance analysis of the proposed method with conventional machine learning techniques and studies in the literature is performed. Experimental results show that the proposed method is effective and can be utilized by doctors for early diagnosis.

Keywords - Cancer Diagnosis, Breast Cancer, Machine Learning, Principal Component Analysis, Feed Forward Neural Network

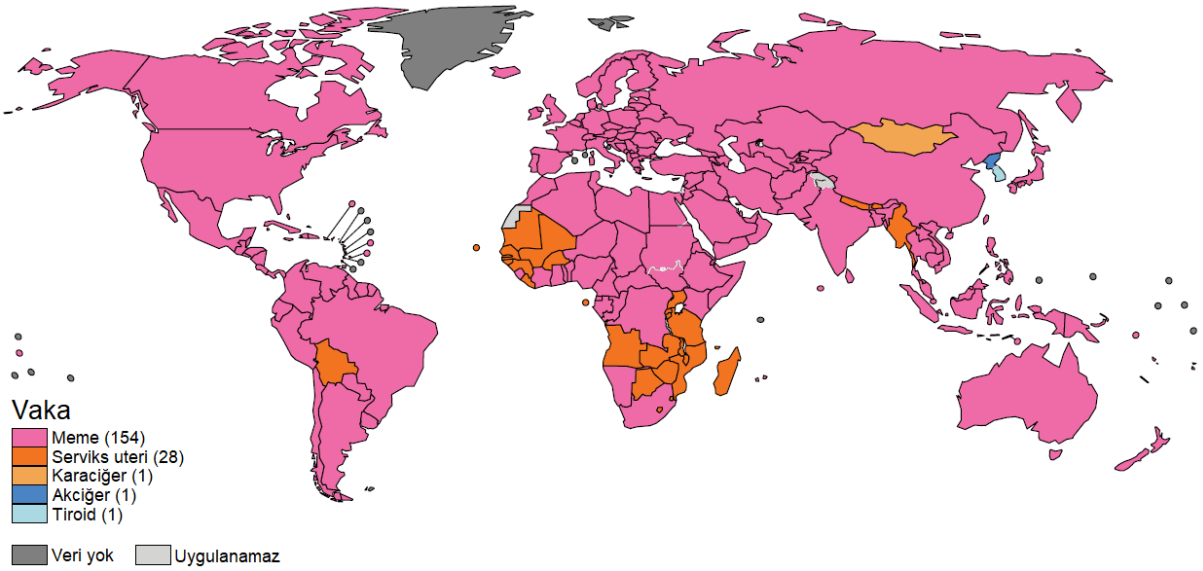
BŞEÜ Fen Bilimleri Dergisi
7(2), 1106-1123, 2020

BSEU Journal of Science
DOI: 10.35193/bseufbd.742456

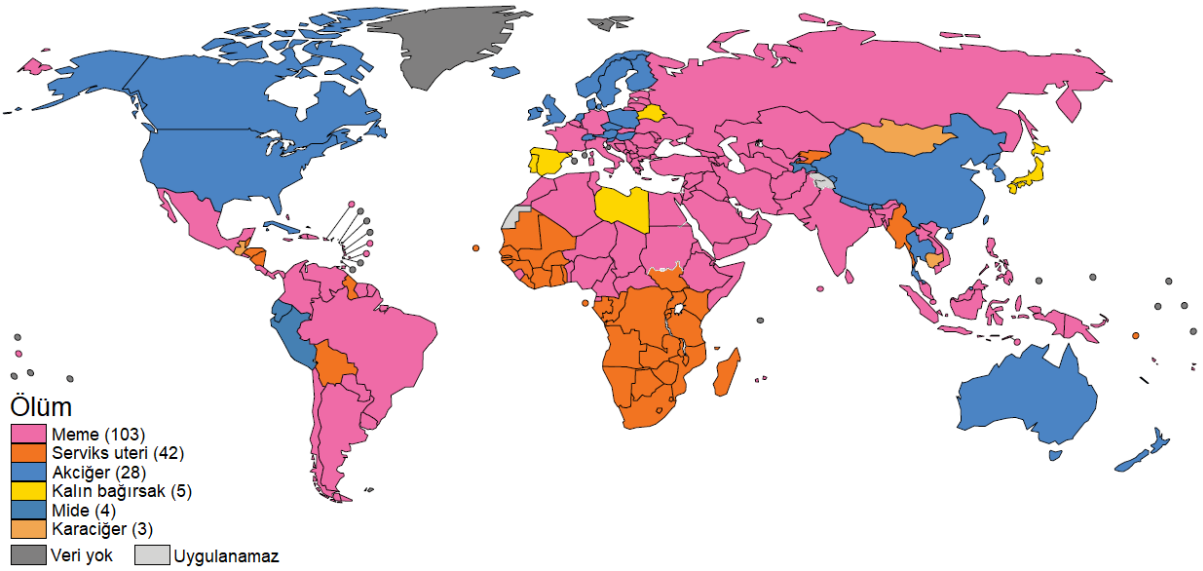
e-ISSN: 2458-7575 (http://dergipark.gov.tr/bseufbd)

I. GİRİŞ

Uluslararası Kanser Araştırmaları Ajansı (International Agency for Research on Cancer-IARC) [1] 2018 yılında dünya çapında yaklaşık 18,1 milyon kişiye kanser hastalığı tanısı konduğunu ve yaklaşık 9,6 milyon kişinin öldüğünü bildirmiştir. Dünya genelinde en sık görülen kanser türü akciğer kanseridir. Tüm vakaların %11,6'sında görülmektedir ve tüm ölümlerin %18,4'ünün sebebidir. En sık görülen ikinci kanser türü ise tüm ölümlerin %11,6'sına neden olan meme kanseridir. Ayrıca kadınlarda görülen kanser ölümlerinin başlıca nedenidir. 2018 yılında yaklaşık 2,1 milyon kadına meme kanseri teşhisi konmuştur. Şekil 1'de 2018 yılında kadınlar arasında dünya çapındaki kanser vaka ve ölüm oranı gösterilmektedir. Görüldüğü üzere meme kanseri 185 ülkenin 154'ünde en sık tanı konulan kanser türüdür. 103 ülkede ise kanser ölümlerinin başlıca sebebidir [2-4].



(a)



(b)

Şekil 1. 2018 yılında kadınlar arasında dünya çapındaki kanser vaka ve ölüm oranı: (a) Vaka ve (b) Ölüm oranı [2-4]

Meme kanseri hastalığı nedeniyle olan ölümlerin sayısını düşürmenin yolu erken teşhistir. Hastalık teşhisi tıp alanında çok karmaşık bir süreçtir ve doğru tanı için çeşitli testler gereklidir. Bilgisayar bilimindeki son gelişmeler sayesinde geliştirilen yapay zeka teknikleri, uzman sistemler ve makine öğrenmesi yöntemleri, hastalıkları erken teşhis etmede doktorlara yardımcı olmak için kullanılmaktadır. Özellikle meme kanseri hastalığında erken teşhis sayesinde hastalara daha fazla tedavi seçenekleri uygulanabilir ve ölüm riski en aza indirilebilir [5,6]. Erken tanının hayati önem taşıdığı meme kanseri hastalığı için rutin kan analizi gibi düşük maliyetli ve kullanımı kolay yöntemlerden yararlanmak, mamografi ve manyetik rezonans görüntülemeden (Magnetic Resonance Imaging-MRI) önce oldukça pratiktir.

2018 yılında rutin kan analizine dayanan ve Meme Kanseri Coimbra Veri Seti (Breast Cancer Coimbra Dataset-BCCD) adı verilen yeni bir veri seti oluşturulmuştur. Bu veri seti rutin kan analizi testleri ile deneklerden toplanan verilerden oluşturulduğu için meme kanseri, makine öğrenmesi yöntemleriyle daha kolay ve düşük maliyetle teşhis edilebilir. BCCD ilk olarak Patricio vd. [7] tarafından yapılan çalışmada 116 gönüllünün 9 klinik özelliği toplanarak oluşturulmuştur. Meme kanseri tanısı için Monte Carlo çapraz geçerleme (Monte Carlo Cross Validation-MCCV) ile birlikte destek vektör makinesi (Support Vector Machine-SVM), lojistik regresyon (Logistic Regression) ve rastgele ormanlar (Random Forests) teknikleri kullanılmıştır. En iyi AUC (Area Under the Curve) değerleri SVM tekniği ile 0,87-0,91 aralığında elde edilmiştir. Bu çalışmanın sonuçları, rutin kan analizi yoluyla toplanan klinik özelliklerin, meme kanseri hastalığının varlığını tespit etmek için kullanılmasında umut verici olduğunu göstermektedir. Aynı veri setini kullanan diğer bir çalışmada Li ve Chen [8], meme kanseri hastalığı tahmini için SVM, karar ağacı (Decision Tree), rastgele ormanlar, yapay sinir ağı (Artificial Neural Network-ANN) ve lojistik regresyon yöntemlerinin performansını incelemiş ve karşılaştırmıştır. Çalışmanın deneysel sonuçları, rastgele ormanlar sınıflandırıcısının performansının 0,785 AUC değeri ve %74,3 doğruluk oranı ile diğer sınıflandırma yöntemlerinden daha iyi olduğunu göstermektedir. Livieris vd. [9] Co-training, Self-training ve Tri-training (CST)-oylama'nın ileri bir modeli olan ve geliştirilmiş CST-oylama (Improved CST-Voting) olarak adlandırılan bir meme kanseri hastalığı tahmin algoritması önermiştir. Bu algoritmanın genel sınıflandırma doğruluğu %77,59'dur. Aslan vd. [10] ise aynı veri setini kullanarak yapmış oldukları çalışmada meme kanseri hastalığını teşhis etmek için ANN, SVM, k-en yakın komşu (k-Nearest Neighbours-k-NN) ve aşırı öğrenme makineleri (Extreme Learning Machine-ELM) tekniklerinin sınıflandırma performansını incelemişlerdir. Karşılaştırmalı deney sonuçları, ELM'nin %80'lik sınıflandırma doğruluğu ile en iyi performansa sahip olduğunu göstermektedir. Bu makalede ise daha önce yapılan çalışmalardan farklı PCA ve FFNN tekniklerini kullanan yeni bir yöntem önerilmiştir. Ayrıca klasik makine öğrenmesi tekniklerinin BCCD veri seti üzerindeki sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü performansı ayrıntılı olarak incelenmiştir.

Bu çalışmanın diğer bölümleri şu şekilde organize edilmiştir: Bölüm 2'de çalışmada kanser teşhisi için kullanılan veri seti, PCA ile FFNN teknikleri ve bu iki tekniği temel alarak önerilen yöntem açıklanmaktadır. Bölüm 3'te önerilen yöntemin klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile karşılaştırması için kullanılan performans metrikleri ve deney sonuçları yer almaktadır. Son olarak Bölüm 4'te ise çalışma sonlandırılmaktadır.

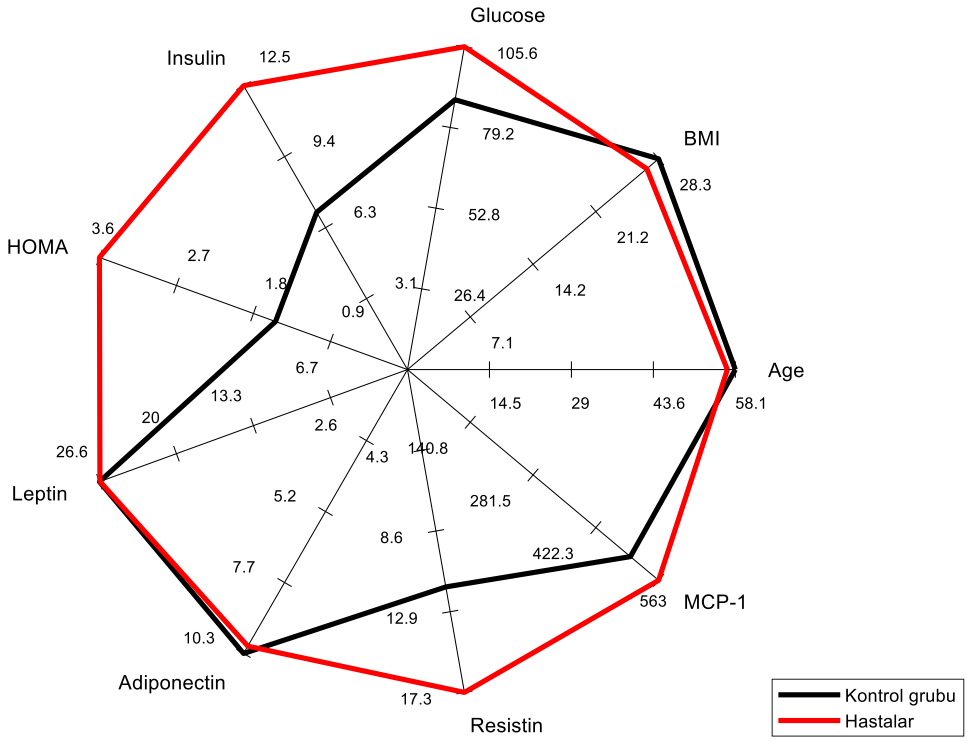
II. MATERYAL VE METOT

A. *Veri Seti Açıklaması*

Bu çalışmada hastalık teşhisi için Kaliforniya Üniversitesi - Irvine Makine Öğrenmesi Deposunda (Machine Learning Repository) açık erişimli olarak sunulan ve rutin kan analizi ile elde edilmiş olan BCCD veri seti kullanılmıştır [7,11]. Bu veri seti, 64'ü meme kanseri hastası ve 52'si sağlıklı (kontrol grubu bireyleri) olan 116 örneği içermektedir. Veri setinde Age (yaş), BMI, Glucose (glikoz), Insulin (insülin), HOMA, Leptin, Adiponectin (adiponektin), Resistin, MCP-1 ve hastaları veya sağlıklı kontrol grubunu gösteren bir ikili sınıf özelliği olan toplam dokuz nicel öznitelik vardır. Bu öznitelikleri içeren veri seti Tablo 1'de özetlenmiştir. Değişkenlerin numaralandırılması V1-V9 olacak şekilde azalan önem derecesini vurgulamaktadır. Diğer bir deyişle Glucose en yüksek önem derecesine sahip, MCP-1 ise en düşük öneme sahip özniteliktir [7]. Ayrıca hasta ve sağlıklı kontrol gruplarının klinik özelliklerinin ortalama profilleri Şekil 2'de daha iyi anlaşılması için grafiksel olarak sunulmuştur. Burada her radyal çizgi bir özniteliğe (hasta grubu için kırmızı çizgi ve sağlıklı kontrol grubu için siyah çizgi) karşı düşmektedir.

Tablo 1. BCCD'nin özeti

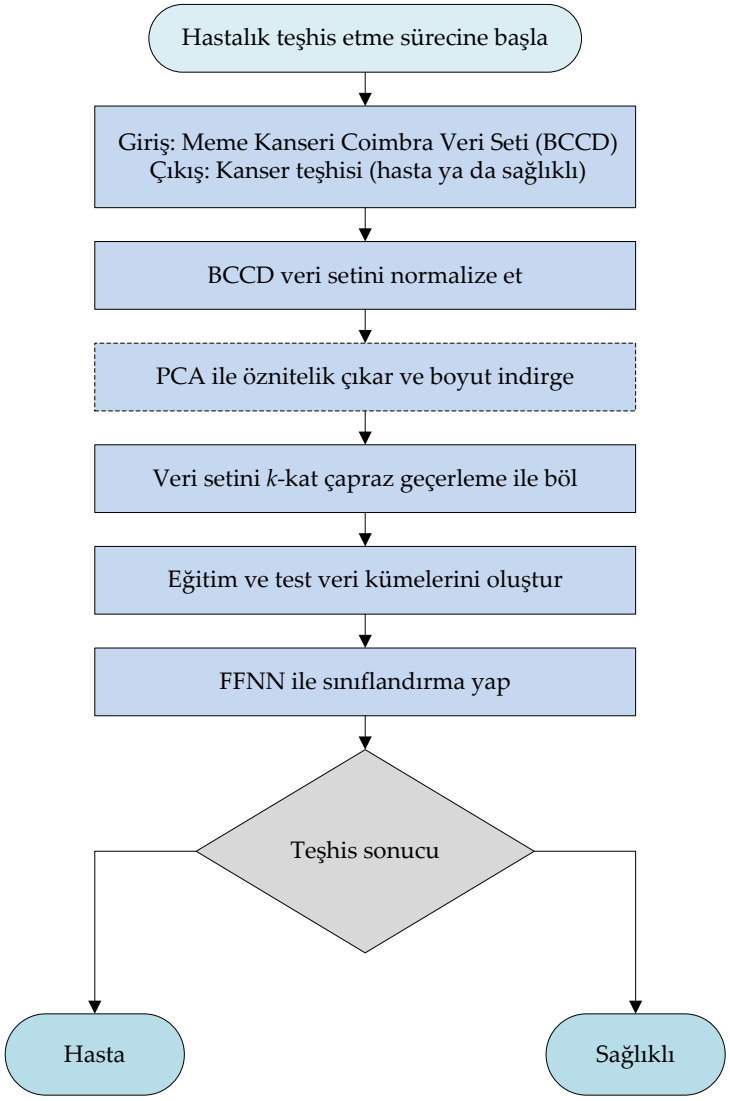
Öznitelik #	Değişken	Öznitelik adı	Birim
1	V3	Age	years
2	V4	BMI	kg/m^2
3	V1	Glucose	mg/dL
4	V7	Insulin	µU/mL
5	V5	HOMA	-
6	V6	Leptin	ng/mL
7	V8	Adiponectin	µg/mL
8	V2	Resistin	ng/mL
9	V9	MCP-1	pg/dL



Şekil 2. Hasta ve sağlıklı kontrol gruplarının klinik özelliklerinin ortalama profilleri

B. Önerilen Metot

Bu çalışmada PCA ve FFNN tekniklerine dayanan yeni bir kanser teşhisi yöntemi önerilmiştir. Önerilen yöntemin akış diyagramı Şekil 3'te gösterilmektedir. Bu yöntemde ilk olarak BCDD veri seti algoritmaya giriş olarak verilir ve min-max normalizasyonu yapılır. Sonrasında PCA tekniği kullanılarak veri setinin öznitelikleri çıkarılır ve boyut indirgeme yapılmış olur. Ardından k-kat çapraz geçerleme metodu ile veri seti bölünerek eğitim ve test kümeleri oluşturulur. Son olarak FFNN tekniği ile sınıflandırma yapılarak teşhis sonucu (hasta/sağlıklı) elde edilir. Bu yöntemin temelini oluşturan PCA ve FFNN teknikleri sonraki bölümlerde ayrıntılı olarak anlatılmaktadır.



Şekil 3. Kanser teşhisi için önerilen yöntemin akış diyagramı

C. Temel Bileşen Analizi

PCA, şu ana kadar çok çeşitli alanlarda yaygın olarak kullanılmış en bilinen boyut indirgeme tekniği türlerinden biridir ve pek çok uygulamada öznitelik çıkarma ve boyut indirgeme amacıyla yoğun olarak kullanılmaktadır. PCA, çok sayıda ilişkili değişkeni ortogonal dönüşümler kullanarak daha küçük bir ilişkisiz kümeye dönüştüren çok değişkenli bir tekniktir [12, 13]. Bu teknik, her eksene asıl bileşen adı verilen ve orjinal verilerdeki değişkenlerin doğrusal bir birleşimi olarak adlandırılan ortogonal yeni bir uzay oluşturur. Nicel olarak titiz bir hesaplama ile elde edilen tüm temel bileşenler birbirine dik olduğundan bu yeni alanda bilgi fazlalığı yoktur [14]. PCA, birinci eksenin o eksen boyunca varyansı maksimize etmek için noktaların en büyük varyansı yönünde yerleştirildiği özel bir koordinat sistemi oluşturur [14]. Birincisine dik olan ikinci eksen, ilgili eksen boyunca olan varyansı maksimize eden ikinci temel bileşene karşılık gelir. Benzer bir biçimde diğer tüm temel bileşenler, kalan varyanstaki paylarını en üst düzeye çıkaracak şekilde oluşturulur. PCA tarafından dönüştürülen bu yeni temsil uzayında, temel bileşenler mümkün olan en büyük varyanstan en düşük olana kadar sıralanır, yani birinci bileşen her zaman en büyük varyansı açıklar ve onu takip edenler öncekinden daha küçük bir varyans değerini temsil eder [12].

Temel bileşenleri hesaplamak için gereken bağıntılar buradan sonra verilmiştir. t=1,2, …, n için $\{x(t)\}$, uygun gelen örnekleri ve sıfır ortalamalı özellikleri içeren rastgele bir veri kümesi olsun. Bu durumda kovaryans matrisi $\boldsymbol{R}$ aşağıdaki gibi hesaplanır:

$$\boldsymbol{R} = \frac{1}{n-1} \sum_{t=1}^{n} \left[\boldsymbol{x}(t)\, \boldsymbol{x}(t)^{T} \right] \tag{1}$$

Aşağıdaki denklem, orijinal verilerdeki değişkenlerin doğrusal kombinasyonlarını, yani $\boldsymbol{x}(t)$'den $\boldsymbol{y}(t)$'ye lineer dönüşümü hesaplamak için kullanılabilir.

$$\boldsymbol{y}(t)=\boldsymbol{M}^{T}\boldsymbol{x}(t) \tag{2}$$

Burada $\boldsymbol{M}$, n × n boyutunda bir ortogonal matris ifade eder ve bu matrisin i. sütunu esasen i. özvektöre eşittir. Bu noktada, özdeğer problemi başlangıçta aşağıdaki denklem ile çözülecek şekilde ayarlanır:

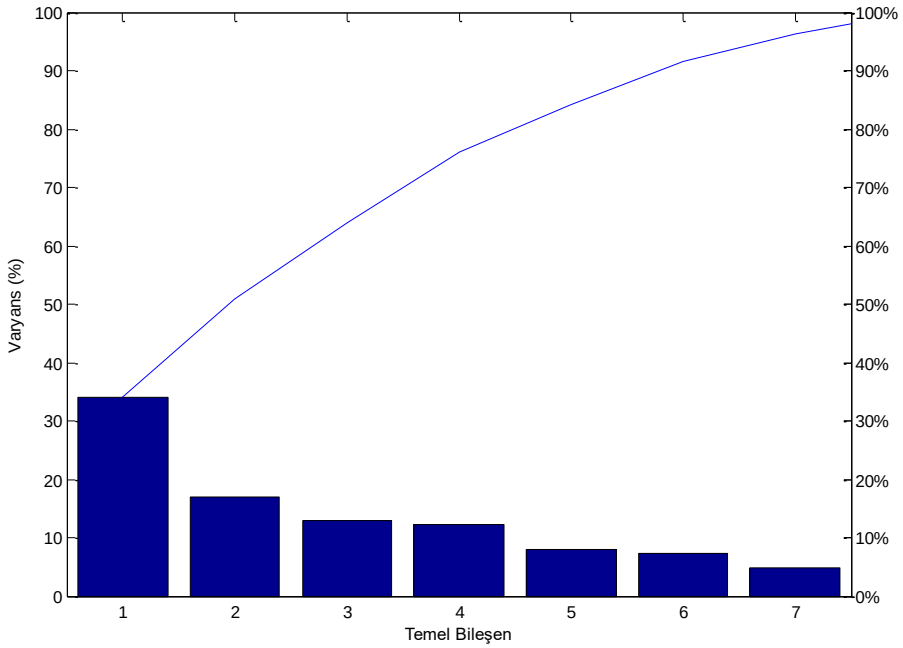
$$\lambda_i \boldsymbol{q}_i = \boldsymbol{R}\, \boldsymbol{q}_i \tag{3}$$

Burada $\boldsymbol{q}_i$, özvektöre karşılık gelir ve λ_i ise $\boldsymbol{R}$ kovaryans matrisinin bir özdeğerini ifade eder ($\lambda_1 > \lambda_2 > \cdots > \lambda_n$). Denklem 3'e dayanarak temel bileşenler şu şekilde hesaplanır:

$$y_i(t) = \boldsymbol{q}_i^{T}\boldsymbol{x}(t)\,,\ i=1,\ldots,n \tag{4}$$

Burada $y_i(t)$, i. temel bileşeni sembolize eder. Bu konu ile ilgili daha fazla bilgi için ilgili çalışmalara [12, 13] bakılabilir.

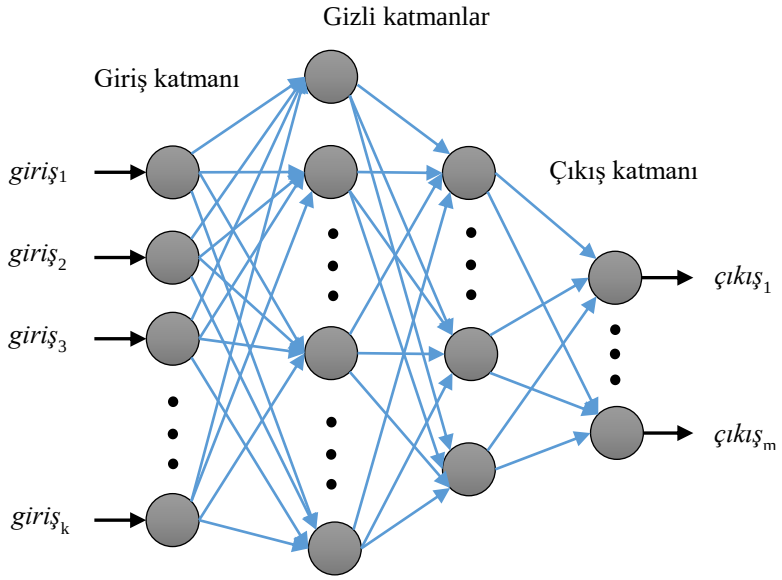
Şekil 4, yukarıda bahsedilen meme kanseri veri kümesi için hesaplanan her bir temel bileşen tarafından belirtilen varyans yüzdesini göstermektedir. Şekildeki grafik, toplam varyansın yaklaşık %95'ini açıklayan sadece ilk yedi bileşeni (toplam dokuz yerine) göstermektedir.



Şekil 4. Temel bileşenler tarafından belirtilen varyans yüzdeleri

D. İleri Beslemeli Sinir Ağı

FFNN genel yapısı Şekil 5'te gösterilmektedir. Mimariden görüldüğü üzere FFNN, iki veya daha fazla işlem katmanının birbiri ardına eklenmesiyle oluşmaktadır. İlk katman giriş katmanıdır ve ağ girişleri için bir kapı görevi yapar. Bu sinir ağı türünde birden fazla gizli katman olabilir. Bu katmanlar çözülecek problemle ilgili ana ve karmaşık ilişkileri barındıran nöronlardan oluşmaktadır. Her bir katmandaki işlem birimleri arasında ağırlıklar vardır. Bu ağırlıklar katmanlarda üretilen sonuçların bir sonraki katmana iletilmesini ve problemle ilgili dağıtık bir depolama mimarisinin olmasını sağlar. Son olarak çıkış katmanı, nihai yaklaşım ya da izdüşüm sonucunu üretir. Çok katmanlı FFNN ağlarında her katmanda bazı aktivasyon fonksiyonları vardır. Tüm katmanlar göz önünde bulundurulduğunda bileşik bir fonksiyon uygulaması olabilir. Dolayısıyla FFNN'nin sağladığı nihai yaklaşım sonucu bileşik bir fonksiyonla karakterize edilebilir [15-17]. Bu ağın çalışması ile igili daha ayrıntılı bilgi için [18, 19] kaynaklarından yararlanılabilir.



Şekil 5. FFNN'nin genel yapısı

E. Klasik Makine Öğrenmesi Teknikleri

Bu çalışmada önerilen yöntemin performansı klasik makine öğrenmesi teknikleriyle kıyaslanmıştır. Bu yöntemler şunlardır: Bayesian lojistik regresyon (Bayesian logistic regression) [20], Naive Bayes [21], lojistik regresyon [22], radyal tabanlı fonksiyon ağı (Radial Basis Function Network-RBFN) [23], stokastik dereceli azalma (Stochastic Gradient Descent-SGD) [24], SVM [25], voted perceptron [26], *k*-NN [27], K* (K-yıldız) [28], AdaBoostM1 [29], OneR [30], decision stump [31], Hoeffding ağacı [32], C4.5 (J48) karar ağacı [33], lojistik model ağaçları (Logistic Model Trees-LMT) [34], Naive Bayes ile karar ağacı (Decision Tree with Naive Bayes) [35] ve rastgele ormanlar [36].

III. BULGULAR VE TARTIŞMA

Bu çalışmada önerilen yöntem ve karşılaştırma için kullanılan sınıflandırma teknikleri, Intel Core i7 8565U işlemci (1.80 GHz) ve 8 GB RAM ile Windows 10 Pro 64-bit işletim sisteminde çalışan Weka 3.8.3'te gerçeklenmiştir. Bu bölümde, karşılaştırma için kullanılan performans metriklerine, önerilen yöntemin klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile kıyaslanmasına yer verilmektedir.

A. Karşılaştırma için Kullanılan Performans Metrikleri

Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin performansını değerlendirmek için kesinlik, duyarlılık, F-ölçütü ve sınıflandırma doğruluğu metrikleri kullanılmıştır. Doğru sınıflandırılmış test seti örneklerinin yüzdesi doğruluk olarak tanımlanır. Kesinlik ve duyarlılık metrikleri ise diğer performans ölçümleridir. F-skoru veya F_1 skoru olarak da adlandırılan F-ölçütü, prensipte kesinlik ve bütünlük metriklerinin harmonik ortalamasını temsil eden bir metriktir [37-40].

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$\text{Kesinlik} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{F–ölçütü} = \frac{2 \times \text{Kesinlik} \times \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \tag{8}$$

Burada *TP* (True Positives), *TN* (True Negatives), *FP* (False Positives) ve *FN* (False Negatives) dört olası neticeyi sembolize eder ve ilgili durumla sonuçlanan örnek sayılarını tutar. Bu konu hakkında daha fazla ayrıntı için okuyucular ilgili çalışmadan [41] faydalanabilir.

B. Klasik Makine Öğrenmesi Teknikleri ile Karşılaştırma

Önerilen yöntemin ve literatürdeki klasik makine öğrenmesi tekniklerinin performansı kesinlik, duyarlılık, F-ölçütü ve sınıflandırma doğruluğu metrikleri açısından incelenmiştir. Deneyler farklı öznitelik kombinasyonları üzerinde yapılarak önerilen yöntemin ve var olan tekniklerin başarısının performans metrikleri açısından nasıl etkilendiği araştırılmıştır. Test için kullanılan öznitelik kombinasyonları şu şekildedir: V1-V2, V1-V3, V1-V4, V1-V5, V1-V6 ve V1-V9. Önerilen yöntemin ve karşılaştırma için kullanılan sınıflandırma metotlarının parametreleri deneysel olarak belirlenmiştir. Bu bağlamda önerilen yöntemdeki FFNN'nin gizli katmanındaki nöron sayısı farklı öznitelik kombinasyonları için 10-300 aralığında olacak şekilde grid arama yaklaşımıyla optimize edilmiştir. Tablo 2-7'de önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin farklı öznitelik kombinasyonları için 10-kat çapraz geçerlemedeki performans metrik sonuçları verilmektedir. Tablo 2, V1-V2 durumu için olan performans sonuçlarını göstermektedir. Tablodan görüldüğü üzere normalizasyon adımından sonra PCA'nın uygulanması RBFN, SGD + SVM, voted perceptron, K*, OneR, decision stump, Hoeffding ağacı, C4.5 (J48) karar ağacı, LMT, Naive Bayes ile karar ağacı ve rastgele ormanlar tekniklerinin tüm metriklerdeki performansını artırmıştır. Lojistik regresyon, SGD + lojistik regresyon ve *k*-NN (normalizasyon için k=26, normalizasyon + PCA için k=15) tekniklerinde sınıflandırma doğruluğunun değişmemesine, diğer metriklerde ise aynı ya da çok yakın sonuç vermesine sebep olmuştur. Bayesian lojistik regresyon, Naive Bayes, SVM ve AdaBoostM1 + LMT yöntemlerinde genel olarak tüm metriklerde performans düşüşüne neden olmuştur. V1-V2 kombinasyonu için sadece normalizasyonun uygulandığı durumda önerilen yöntem (gizli katmandaki nöron sayısı=200) %76,7241'lik sınıflandırma doğruluğu ile diğer tüm tekniklerden daha iyi performans göstermiştir. Ayrıca kesinlik, duyarlılık ve F-ölçütü metriklerinde de en iyi performans sonuçları önerilen yönteme aittir. Normalizasyon adımından sonra PCA'nın uygulandığı durumda ise önerilen yöntemin (gizli katmandaki nöron sayısı=100) sınıflandırma doğruluğu OneR, decision stump ve C4.5 (J48) karar ağacı ile aynıdır. Diğer metriklerde ise bu tekniklerle aynı ya da çok yakın performans göstermiştir.

Tablo 2. Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin V1-V2 kombinasyonu için performans metrik sonuçları

Sınıflandırma Yöntemleri	Normalizasyon				Normalizasyon + PCA			
	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)
Bayesian lojistik regresyon [20]	0,747	0,733	0,722	73,2759	0,712	0,707	0,708	70,6897
Naive Bayes [21]	0,707	0,664	0,657	66,3793	0,718	0,655	0,643	65,5172
Lojistik regresyon [22]	0,693	0,690	0,690	68,9655	0,693	0,690	0,690	68,9655
RBFN [23]	0,741	0,741	0,740	74,1379	0,750	0,750	0,748	75,0000
SGD + SVM [24]	0,688	0,681	0,682	68,1034	0,718	0,707	0,707	70,6897
SGD + Lojistik regresyon [24]	0,707	0,707	0,707	70,6897	0,707	0,707	0,707	70,6897
SVM [25]	0,657	0,629	0,587	62,9310	0,606	0,560	0,425	56,0345
Voted perceptron [26]	0,561	0,569	0,523	56,8966	0,718	0,716	0,716	71,5517
k-NN [27]	0,741	0,741	0,741	74,1379	0,741	0,741	0,739	74,1379
K* [28]	0,680	0,681	0,680	68,1034	0,697	0,698	0,697	69,8276
AdaBoostM1 + LMT [29]	0,735	0,733	0,733	73,2759	0,689	0,690	0,686	68,9655
OneR [30]	0,629	0,629	0,629	62,9310	0,759	**0,759**	**0,759**	**75,8621**
Decision stump [31]	0,673	0,664	0,664	66,3793	**0,762**	**0,759**	0,755	**75,8621**
Hoeffding ağacı [32]	0,695	0,647	0,638	64,6552	0,718	0,655	0,643	65,5172
C4.5 (J48) karar ağacı [33]	0,654	0,655	0,647	65,5172	**0,762**	**0,759**	0,755	**75,8621**
LMT [34]	0,698	0,698	0,698	69,8276	0,733	0,733	0,733	73,2759
Naive Bayes ile karar ağacı [35]	0,681	0,681	0,675	68,1034	0,754	0,750	0,745	75,0000
Rastgele ormanlar [36]	0,662	0,664	0,660	66,3793	0,672	0,672	0,672	67,2414
Önerilen yöntem	**0,770**	**0,767**	**0,764**	**76,7241**	**0,762**	**0,759**	0,755	**75,8621**

V1-V3 kombinasyonu için olan performans sonuçları Tablo 3'te gösterilmektedir. Görüldüğü üzere sadece normalizasyonun uygulandığı durumda önerilen yöntem (gizli katmandaki nöron sayısı=200) RBFN ile birlikte %81,8966'lik doğruluk oranıyla en iyi sonuca sahiptir. Diğer metriklerde ise RBFN ile neredeyse aynı performansı göstermiştir. RBFN'nin dışındaki klasik makine öğrenmesi tekniklerinden ise tüm metriklerde daha iyi performans göstermiştir. PCA'nın uygulandığı durumda önerilen yöntemin (gizli katmandaki nöron sayısı=150) performansı tüm metriklerde *k*-NN (normalizasyon için *k*=10, normalizasyon + PCA için *k*=9) ve Naive Bayes ile karar ağacı tekniklerine çok yakındır.

Tablo 3. Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin V1-V3 kombinasyonu için performans metrik sonuçları

Sınıflandırma Yöntemleri	Normalizasyon				Normalizasyon + PCA			
	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)
Bayesian lojistik regresyon [20]	0,675	0,664	0,645	66,3793	0,751	0,750	0,750	75,0000
Naive Bayes [21]	0,688	0,655	0,651	65,5172	0,624	0,595	0,589	59,4828
Lojistik regresyon [22]	0,751	0,750	0,750	75,0000	0,751	0,750	0,750	75,0000
RBFN [23]	0,820	**0,819**	**0,818**	**81,8966**	0,724	0,724	0,721	72,4138
SGD + SVM [24]	0,716	0,716	0,716	71,5517	0,723	0,724	0,724	72,4138
SGD + Lojistik regresyon [24]	0,723	0,724	0,724	72,4138	0,732	0,733	0,732	73,2759
SVM [25]	0,748	0,672	0,627	67,2414	0,709	0,647	0,595	64,6552
Voted perceptron [26]	0,662	0,664	0,660	66,3793	0,712	0,707	0,708	70,6897
k-NN [27]	0,793	0,793	0,793	79,3103	**0,828**	**0,828**	**0,827**	**82,7586**
K* [28]	0,758	0,759	0,757	75,8621	0,742	0,741	0,738	74,1379
AdaBoostM1 + LMT [29]	0,732	0,733	0,732	73,2759	0,751	0,750	0,750	75,0000
OneR [30]	0,672	0,672	0,672	67,2414	0,596	0,595	0,595	59,4828
Decision stump [31]	0,673	0,664	0,664	66,3793	0,636	0,629	0,605	62,9310
Hoeffding ağacı [32]	0,694	0,664	0,660	66,3793	0,624	0,595	0,589	59,4828
C4.5 (J48) karar ağacı [33]	0,724	0,724	0,722	72,4138	0,762	0,759	0,755	75,8621
LMT [34]	0,732	0,733	0,732	73,2759	0,742	0,741	0,738	74,1379
Naive Bayes ile karar ağacı [35]	0,747	0,741	0,736	74,1379	0,821	0,810	0,806	81,0345
Rastgele ormanlar [36]	0,767	0,767	0,767	76,7241	0,784	0,784	0,784	78,4483
Önerilen yöntem	**0,821**	**0,819**	0,817	**81,8966**	0,806	0,802	0,799	80,1724

Tablo 4, V1-V4 kombinasyonu için olan performans sonuçlarını göstermektedir. Tablodan görüldüğü üzere normalizasyonun tek başına uygulandığı durumda en iyi performans sonuçları kesinlik için 0,810, duyarlılık için 0,810, F-ölçütü için 0,809 ve sınıflandırma doğruluğu için %81,0345 olmak üzere önerilen yönteme (gizli katmandaki nöron sayısı=30) aittir. Normalizasyon adımından sonra PCA'nın uygulandığı durumda ise önerilen yöntem (gizli katmandaki nöron sayısı=70) tüm metriklerde *k*-NN (normalizasyon için *k*=5, normalizasyon + PCA için *k*=3) ve RBFN tekniklerine çok yakın performans göstermiştir. Diğer tekniklerle kıyaslandığında ise tüm metriklerde daha iyi performansa sahiptir.

Tablo 4. Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin V1-V4 kombinasyonu için performans metrik sonuçları

Sınıflandırma Yöntemleri	Normalizasyon				Normalizasyon + PCA			
	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)
Bayesian lojistik regresyon [20]	0,708	0,707	0,702	70,6897	0,733	0,733	0,733	73,2759
Naive Bayes [21]	0,713	0,681	0,678	68,1034	0,691	0,681	0,682	68,1034
Lojistik regresyon [22]	0,743	0,741	0,742	74,1379	0,743	0,741	0,742	74,1379
RBFN [23]	0,793	0,793	0,792	79,3103	0,819	0,819	0,818	81,8966
SGD + SVM [24]	0,743	0,741	0,742	74,1379	0,695	0,690	0,690	68,9655
SGD + Lojistik regresyon [24]	0,698	0,698	0,698	69,8276	0,716	0,716	0,716	71,5517
SVM [25]	0,671	0,672	0,668	67,2414	0,672	0,672	0,672	67,2414
Voted perceptron [26]	0,646	0,647	0,638	64,6552	0,759	0,759	0,759	75,8621
k-NN [27]	0,801	0,802	0,801	80,1724	**0,829**	**0,828**	**0,828**	**82,7586**
K* [28]	0,697	0,698	0,697	69,8276	0,735	0,724	0,714	72,4138
AdaBoostM1 + LMT [29]	0,759	0,759	0,759	75,8621	0,732	0,733	0,731	73,2759
OneR [30]	0,672	0,672	0,672	67,2414	0,526	0,517	0,518	51,7241
Decision stump [31]	0,673	0,664	0,664	66,3793	0,526	0,483	0,437	48,2759
Hoeffding ağacı [32]	0,707	0,672	0,668	67,2414	0,691	0,681	0,682	68,1034
C4.5 (J48) karar ağacı [33]	0,690	0,690	0,690	68,9655	0,697	0,698	0,695	69,8276
LMT [34]	0,724	0,724	0,724	72,4138	0,674	0,672	0,673	67,2414
Naive Bayes ile karar ağacı [35]	0,751	0,750	0,747	75,0000	0,616	0,612	0,613	61,2069
Rastgele ormanlar [36]	0,767	0,767	0,766	76,7241	0,758	0,759	0,758	75,8621
Önerilen yöntem	**0,810**	**0,810**	**0,809**	**81,0345**	0,802	0,802	0,800	80,1724

V1-V5 kombinasyonu için olan performans sonuçları Tablo 5'te gösterilmektedir. Sadece normalizasyonun uygulandığı durumda önerilen yöntemin (gizli katmandaki nöron sayısı=200) performansı tüm metriklerde *k*-NN (normalizasyon için *k*=8, normalizasyon + PCA için *k*=5) tekniğine yakındır. *k*-NN'nin dışındaki klasik makine öğrenmesi tekniklerinden ise tüm metriklerde daha iyi performans göstermiştir. PCA'nın uygulandığı durumda önerilen yöntem (gizli katmandaki nöron sayısı=250) RBFN ile birlikte %80,1724'lük sınıflandırma doğruluğu ile en iyi sonuca sahiptir. Diğer metriklerde ise RBFN ile yaklaşık olarak aynı performansı göstermiştir.

Tablo 5. Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin V1-V5 kombinasyonu için performans metrik sonuçları

Sınıflandırma Yöntemleri	Normalizasyon				Normalizasyon + PCA			
	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)
Bayesian lojistik regresyon [20]	0,715	0,716	0,714	71,5517	0,762	0,759	0,759	75,8621
Naive Bayes [21]	0,678	0,612	0,595	61,2069	0,682	0,655	0,652	65,5172
Lojistik regresyon [22]	0,769	0,767	0,768	76,7241	0,769	0,767	0,768	76,7241
RBFN [23]	0,786	0,784	0,785	78,4483	**0,802**	**0,802**	**0,802**	**80,1724**
SGD + SVM [24]	0,743	0,741	0,742	74,1379	0,688	0,681	0,682	68,1034
SGD + Lojistik regresyon [24]	0,741	0,741	0,741	74,1379	0,682	0,681	0,681	68,1034
SVM [25]	0,680	0,681	0,678	68,1034	0,679	0,681	0,679	68,1034
Voted perceptron [26]	0,712	0,707	0,699	70,6897	0,718	0,716	0,716	71,5517
k-NN [27]	**0,829**	**0,828**	**0,828**	**82,7586**	0,793	0,793	0,792	79,3103
K* [28]	0,647	0,647	0,647	64,6552	0,768	0,767	0,767	76,7241
AdaBoostM1 + LMT [29]	0,707	0,707	0,707	70,6897	0,708	0,707	0,707	70,6897
OneR [30]	0,672	0,672	0,672	67,2414	0,609	0,612	0,600	61,2069
Decision stump [31]	0,673	0,664	0,664	66,3793	0,515	0,509	0,510	50,8621
Hoeffding ağacı [32]	0,678	0,612	0,595	61,2069	0,671	0,647	0,644	64,6552
C4.5 (J48) karar ağacı [33]	0,690	0,690	0,690	68,9655	0,607	0,603	0,604	60,3448
LMT [34]	0,716	0,716	0,716	71,5517	0,757	0,750	0,751	75,0000
Naive Bayes ile karar ağacı [35]	0,715	0,716	0,713	71,5517	0,636	0,638	0,636	63,7931
Rastgele ormanlar [36]	0,784	0,784	0,784	78,4483	0,767	0,767	0,766	76,7241
Önerilen yöntem	0,812	0,810	0,809	81,0345	0,801	**0,802**	0,801	**80,1724**

Tablo 6, V1-V6 kombinasyonu için olan performans sonuçlarını göstermektedir. Normalizasyonun tek başına uygulandığı durumda önerilen yöntem (gizli katmandaki nöron sayısı=280) *k*-NN (normalizasyon için *k*=10, normalizasyon + PCA için *k*=28) ile birlikte %81,0345'lik doğruluk oranı ile en iyi sonuca sahiptir. Diğer metriklerde ise *k*-NN ile neredeyse aynı performansı göstermiştir. Normalizasyon adımından sonra PCA'nın uygulandığı durumda ise en iyi performans sonuçları sınıflandırma doğruluğu için %83,6207, kesinlik, duyarlılık ve F-ölçütü için 0,836 olmak üzere önerilen yönteme (gizli katmandaki nöron sayısı=200) aittir.

Tablo 6. Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin V1-V6 kombinasyonu için performans metrik sonuçları

Sınıflandırma Yöntemleri	Normalizasyon				Normalizasyon + PCA			
	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)
Bayesian lojistik regresyon [20]	0,706	0,707	0,705	70,6897	0,764	0,759	0,759	75,8621
Naive Bayes [21]	0,671	0,603	0,584	60,3448	0,651	0,647	0,647	64,6552
Lojistik regresyon [22]	0,771	0,767	0,768	76,7241	0,762	0,759	0,759	75,8621
RBFN [23]	0,777	0,776	0,776	77,5862	0,813	0,810	0,811	81,0345
SGD + SVM [24]	0,737	0,733	0,733	73,2759	0,708	0,707	0,707	70,6897
SGD + Lojistik regresyon [24]	0,752	0,750	0,751	75,0000	0,691	0,690	0,690	68,9655
SVM [25]	0,644	0,647	0,644	64,6552	0,637	0,638	0,637	63,7931
Voted perceptron [26]	0,629	0,629	0,629	62,9310	0,698	0,690	0,690	68,9655
k-NN [27]	0,810	**0,810**	**0,810**	**81,0345**	0,750	0,750	0,750	75,0000
K* [28]	0,680	0,672	0,673	67,2414	0,723	0,724	0,723	72,4138
AdaBoostM1 + LMT [29]	0,680	0,681	0,680	68,1034	0,708	0,707	0,707	70,6897
OneR [30]	0,672	0,672	0,672	67,2414	0,489	0,491	0,490	49,1379
Decision stump [31]	0,673	0,664	0,664	66,3793	0,456	0,457	0,456	45,6897
Hoeffding ağacı [32]	0,684	0,621	0,605	62,0690	0,651	0,647	0,647	64,6552
C4.5 (J48) karar ağacı [33]	0,691	0,690	0,690	68,9655	0,672	0,672	0,672	67,2414
LMT [34]	0,733	0,733	0,733	73,2759	0,771	0,767	0,768	76,7241
Naive Bayes ile karar ağacı [35]	0,715	0,716	0,713	71,5517	0,680	0,681	0,680	68,1034
Rastgele ormanlar [36]	0,741	0,741	0,740	74,1379	0,759	0,759	0,757	75,8621
Önerilen yöntem	**0,812**	**0,810**	0,809	**81,0345**	**0,836**	**0,836**	**0,836**	**83,6207**

Tüm özniteliklerin kullanıldığı V1-V9 kombinasyonu için olan performans sonuçları Tablo 7'de gösterilmektedir. Tablodan görüldüğü üzere sadece normalizasyonun uygulandığı durumda önerilen yöntem (gizli katmandaki nöron sayısı=150) AdaBoostM1 + LMT ve *k*-NN (normalizasyon için *k*=7, normalizasyon + PCA için *k*=5) teknikleri ile birlikte %76,7241'lik doğruluk oranıyla en iyi sonuca sahiptir. Diğer metriklerde ise bu iki teknik ile yaklaşık olarak aynı performansı göstermiştir. PCA'nın uygulandığı durumda ise önerilen yöntem (gizli katmandaki nöron sayısı=150) tüm metriklerde Bayesian lojistik regresyon tekniğine çok yakın performans göstermiştir.

Tablo 7. Önerilen yöntemin ve klasik makine öğrenmesi tekniklerinin V1-V9 kombinasyonu için performans metrik sonuçları

Sınıflandırma Yöntemleri	Normalizasyon				Normalizasyon + PCA			
	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)	Kesinlik	Duyarlılık	F-ölçütü	Doğruluk (%)
Bayesian lojistik regresyon [20]	0,662	0,664	0,662	66,3793	**0,786**	**0,784**	**0,785**	**78,4483**
Naive Bayes [21]	0,662	0,603	0,587	60,3448	0,627	0,612	0,612	61,2069
Lojistik regresyon [22]	0,733	0,733	0,733	73,2759	0,768	0,767	0,767	76,7241
RBFN [23]	0,744	0,741	0,742	74,1379	0,725	0,724	0,725	72,4138
SGD + SVM [24]	0,725	0,716	0,716	71,5517	0,736	0,724	0,725	72,4138
SGD + Lojistik regresyon [24]	0,732	0,733	0,732	73,2759	0,718	0,716	0,716	71,5517
SVM [25]	0,676	0,672	0,673	67,2414	0,662	0,664	0,662	66,3793
Voted perceptron [26]	0,663	0,664	0,657	66,3793	0,661	0,655	0,656	65,5172
k-NN [27]	**0,769**	**0,767**	**0,768**	**76,7241**	0,768	0,741	0,740	74,1379
K* [28]	0,609	0,595	0,594	59,4828	0,656	0,647	0,647	64,6552
AdaBoostM1 + LMT [29]	0,771	**0,767**	**0,768**	**76,7241**	0,718	0,716	0,716	71,5517
OneR [30]	0,653	0,655	0,652	65,5172	0,519	0,517	0,518	51,7241
Decision stump [31]	0,673	0,664	0,664	66,3793	0,465	0,457	0,458	45,6897
Hoeffding ağacı [32]	0,662	0,603	0,587	60,3448	0,617	0,603	0,603	60,3448
C4.5 (J48) karar ağacı [33]	0,691	0,690	0,690	68,9655	0,649	0,638	0,638	63,7931
LMT [34]	0,693	0,690	0,690	68,9655	0,777	0,776	0,776	77,5862
Naive Bayes ile karar ağacı [35]	0,680	0,681	0,678	68,1034	0,647	0,647	0,647	64,6552
Rastgele ormanlar [36]	0,732	0,733	0,731	73,2759	0,749	0,750	0,749	75,0000
Önerilen yöntem	0,767	**0,767**	0,766	**76,7241**	0,776	0,776	0,776	77,5862

C. Literatürdeki Çalışmalar ile Karşılaştırma

BCCD veri seti çok yeni olduğundan onu kullanan az sayıda çalışma vardır. Bu bölümde, literatürde yer alan çalışmalarda elde edilen performans sonuçları önerilen yönteminkiyle karşılaştırılmaktadır. Önerilen yöntemin performansının AUC ve sınıflandırma doğruluğu açısından mevcut çalışmalarla kıyaslanmasına Tablo 8'de yer verilmektedir. Tabloda görüldüğü üzere var olan çalışmaların sınıflandırma doğruluğu başarısı yaklaşık %74 ile %80 arasındadır. Bu çalışmada önerilen yöntem ise %83,62'lik başarı oranı ile literatürdeki çalışmalardan daha iyi performans göstermiştir.

Tablo 8. Önerilen yöntemin performansının literatürdeki çalışmalarla kıyaslanması

Çalışma	Metotlar	AUC	Doğruluk (%)
Patricio vd. [7]	Lojistik regresyon – Rastgele ormanlar – SVM	0,87-0,91	-
Li ve Chen [8]	Rastgele ormanlar	0,785	74,30
Livieris vd. [9]	Geliştirilmiş CST-oylama	-	79,69
Aslan vd. [10]	ELM	-	80,00
Önerilen yöntem	PCA + FFNN		**83,62**

IV. SONUÇLAR VE ÖNERİLER

Kanser, insan ölümlerinin en büyük sebeplerinden birisidir. Meme kanseri ise kadınlar arasındaki kanser ölümlerinin başlıca nedenidir. Erken teşhis ile bu kanser türü sebebiyle yaşanan ölüm oranları düşürülebilir. Hastalıklara erken tanı koymada doktorlara yardımcı olmak amacıyla uzman sistemler, yapay zeka ve makine öğrenmesi teknikleri kullanılmaktadır. Erken teşhis, özellikle meme kanseri hastalığında ölüm riskini azaltmada çok büyük rol oynamaktadır. Bu çalışmanın ana amacı doktorlara hastalıkları erken teşhis etmede yardımcı olacak bir kanser teşhisi yöntemi önermektir. PCA ve FFNN teknikleri temel alınarak önerilen yöntemin performansı; BCCD veri seti üzerinde sınıflandırma doğruluğu, kesinlik, duyarlılık ve F-ölçütü metrikleri kullanılarak incelenerek, klasik makine öğrenmesi teknikleri ve literatürdeki çalışmalar ile ayrıntılı olarak karşılaştırılmıştır. Çalışmanın deneysel sonuçları önerilen yöntemin etkili bir kanser erken teşhis aracı olarak kullanılabileceğini göstermektedir.

KAYNAKLAR

[1] International Agency for Research on Cancer. (2020). https://www.iarc.fr/, (25.05.2020).

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394-424.

[3] World Health Organization. (2020). https://www.who.int/, (25.05.2020).

[4] New Global Cancer Data: GLOBOCAN 2018. (2020). https://www.uicc.org/new-global-cancer-data-globocan-2018, (25.05.2020).

[5] Eyupoglu, C. (2018). Breast cancer classification using k-nearest neighbors algorithm. *The Online Journal of Science and Technology*, *8*(3), 29-34.

[6] Jeleń, Ł., Krzyżak, A., Fevens, T., & Jeleń, M. (2016). Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies. *Computers in Biology and Medicine*, *79*, 80-91.

[7] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1), 29.

[8] Li, Y., & Chen, Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. *Applied and Computational Mathematics*, *7*(4), 212-216.

[9] Livieris, I., Pintelas, E., Kanavos, A., & Pintelas, P. (2018). An improved self-labeled algorithm for cancer prediction. *Advances in Experimental Medicine and Biology*.

[10] Aslan, M. F., Celik, Y., Sabanci, K., & Durdu, A. (2018). Breast cancer diagnosis by different machine learning methods using blood analysis data. *International Journal of Intelligent Systems and Applications in Engineering*, *6*(4), 289-293.

[11] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). *Breast Cancer Coimbra Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra, (25.05.2020).

[12] Salo, F., Nassif, A. B., & Essex, A. (2019). Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Computer Networks*, *148*, 164-175.

[13] Jackson, J. E. (2005). *A user's guide to principal components*. John Wiley & Sons.

[14] MathWorks. (2018). *Statistics and Machine Learning Toolbox*. The MathWorks Inc.

[15] Yavuz, E., & Eyüpoğlu, C. Meme Kanseri Teşhisi İçin Yeni Bir Skor Füzyon Yaklaşımı. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, *7*(3), 1045-1060.

[16] Yavuz, E., Eyupoglu, C., Sanver, U., & Yazici, R. (2017). An ensemble of neural networks for breast cancer diagnosis. *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 538-543, 5-8 October, Antalya, Turkey.

[17] Yavuz, E., Kasapbaşı, M. C., Eyüpoğlu, C., & Yazıcı, R. (2018). An epileptic seizure detection system based on cepstral analysis and generalized regression neural network. *Biocybernetics and Biomedical Engineering*, *38*(2), 201-216.

[18] Du, K. L., & Swamy, M. N. S. (2006). *Neural Networks in a Softcomputing Framework*. Springer Science & Business Media.

[19] Schalkoff, R. J. (1997). *Artificial Neural Networks*. McGraw-Hill.

[20] Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, *49*(3), 291-304.

[21] John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *10th Conference on Uncertainty in Artificial Intelligence (UAI'95)*, pp. 338-345, 18-20 August, Montréal, Qué, Canada.

[22] Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(1), 191-201.

[23] Frank, E. (2014). *Fully supervised training of Gaussian radial basis function networks in WEKA*. Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[24] Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400-407.

[25] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, *13*(3), 637-649.

[26] Freund, Y., & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, *37*(3), 277-296.

[27] Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37-66.

[28] Cleary, J. G., & Trigg, L. E. (1995). K*: An instance-based learner using an entropic distance measure. *12th International Conference on Machine Learning*, pp. 108-114, 9-12 July, Tahoe City, California.

[29] Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *13th International Conference on Machine Learning*, pp: 148-156, 3-6 July, Bari, Italy.

[30] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*(1), 63-90.

[31] Iba, W., & Langley, P. (1992). Induction of one-level decision trees. *9th International Conference on Machine Learning*, pp. 233-240, 1-3 July, Aberdeen, Scotland.

[32] Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 97-106, 26-29 August, San Francisco, California.

[33] Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

[34] Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*(1-2), 161-205.

[35] Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. *2nd International Conference on Knoledge Discovery and Data Mining*, pp. 202-207, 2-4 August, Portland, Oregon.

[36] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

[37] Eyüpoğlu, C. (2018). *Büyük veride etkin gizlilik koruması için yazılım tasarımı*. Doktora Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, İstanbul.

[38] Yavuz, E., & Eyupoglu, C. (2019). A cepstrum analysis-based classification method for hand movement surface EMG signals. *Medical & Biological Engineering & Computing*, *57*(10), 2179-2201.

[39] Eyupoglu, C., Aydin, M. A., Zaim, A. H., & Sertbas, A. (2018). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, *20*(5), 373.

[40] Yavuz, E., & Eyupoglu, C. (2020). An effective approach for breast cancer diagnosis based on routine blood analysis features. *Medical & Biological Engineering & Computing*. https://doi.org/10.1007/s11517-020-02187-9

[41] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437.