Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 36, Sayı 2 , 2020

Erciyes University
Journal of Institue Of Science and Technology
Volume 36, Issue 2, 2020

Yığınlanmış Özdevinimli Kodlayıcılar ile Göğüs Kanserinin Sınıflandırılması ve Klasik Makine Öğrenme Metotları ile Performans Karşılaştırması

Tayyip Özcan[*1]

[*1] Erciyes Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği, KAYSERİ

(Alınış / Received: 25.04.2020, Kabul / Accepted: 04.08.2020, Online Yayınlanma / Published Online: 17.08.2020)

Anahtar Kelimeler
Göğüs Kanseri,
Derin Öğrenme
Yığınlanmış Özdevinimli Kodlayıcılar,
Veri Ön İşleme,
Makine Öğrenme Metotları

Öz: Göğüs kanseri, her yıl çokça ölüme sebebiyet veren en tehlikeli kanser türleri arasında yer almaktadır. Erken tanı durumları kanser tedavilerinde yapıcı rol oynamaktadır. Bu nedenle araştırmacılar, hastalara ve sağlıklı insanlara ait veriler üzerinde sınıflandırma ve kümeleme yöntemlerini kullanarak deneysel araştırmalar yapmaktadır. Gelişen teknoloji ile makine öğrenme destekli teşhis çalışmalarının yanı sıra derin öğrenme yöntemlerinin kullanımında kayda değer bir artış görülmektedir. Bu çalışmada, bir derin öğrenme metodu olan yığınlanmış özdevinimli kodlayıcılar (stacked autoencoders, SAE) kullanılarak göğüs kanseri sınıflandırılmasında kullanılmak üzere yeni bir model tasarlanmıştır. Tasarlanan SAE ile performans karşılaştırması gerçekleştirmek üzere en yaygın kullanılan makine öğrenme yöntemlerinden destek vektör makineleri, k-en yakın komşuluk, naive bayes ve karar ağaçları metotları bu çalışmada ayrıca kullanılmıştır. Doğruluk oranı metriğinin yanı sıra, eğitim ve test aşamalarındaki geçen süre (zaman karmaşıklığı) deneysel çalışmalarda hesaplanmıştır. Deneysel çalışmalarda, veri ön işleme adımlarından normalizasyon süreci uygulanarak, sınıflandırma başarımına etkisi incelenmiştir. Deneysel sonuçlara göre doğruluk oranı kriteri baz alındığında %79,31 doğruluk oranı ile en başarılı sonuç veri ön işleme destekli SAE ile elde edilmiştir. Zaman karmaşıklığı metriğine göre KNN algoritması eğitim sürecinde en hızlı algoritma olurken SAE algoritması test sürecinde en hızlı olan algoritma olarak tespit edilmiştir.

Classification of Breast Cancer using Stacked Autoencoders and Performance Comparison with Classical Machine Learning Methods

Keywords
Breast Cancer,
Deep Learning,
Stacked Autoencoders,
Data Preprocessing,
Machine Learning
Algorithms

Abstract: Breast cancer is among the most dangerous cancer types that cause many deaths every year. Early diagnosis states play a constructive role in cancer treatments. Therefore, researchers conduct experimental research on the data that belongs to patients and healty peaople using classification and clustering methods. In addition to machine learning assisted diagnostic studies with the developing technology, there is a significant increase in the use of deep learning methods. In this study, a new model is designed to be used in the classification of breast cancer using stacked autoencoders (SAE) which is a deep learning method. Support vector machines, k-nearest neighborhood, naive bayes, and decision trees methods, which are the most commonly used machine learning methods to compare performance with the designed SAE, were also used in this study. In addition to the accuracy rate metric, the elapsed time (time complexity) during the training and testing stages was calculated in experimental studies. In the experimental studies, the effect of the classification performance was examined by applying the normalization process from the data pre-processing steps. According to the experimental results, the most successful result with 79.31% accuracy rate was obtained by data pre-processing aided SAE. According to the time complexity metric, KNN algorithm is the fastest algorithm in the training process, while the SAE algorithm is the fastest in the test process.

*İlgili Yazar, email: tozcan@erciyes.edu.tr

1. Giriş

Kadınlarda en sık görülen kanser tipi olan göğüs kanseri, göğüs dokusundaki süt kanalını meydana getiren ve süt yapıcı hücrelerin kontrolsüz olarak çoğalmasıyla ortaya çıkmaktadır [1]. Göğüs kanserindeki en önemli hususlardan birisi, kanserin kan ve lenf yoluyla diğer organlara yayılmadan teşhisi ve tedavisidir. Dolayısıyla göğüs kanserinde erken teşhis oldukça önemlidir. Bu amaç doğrultusunda araştırmacılar göğüs kanseri üzerine yoğun çalışmalar yapmaktadır.

Bu çalışmada sınıflandırma amacıyla kullanılacak olan Breast Cancer Coimbra (BCC) [2] veri seti üzerine literatürde çeşitli çalışmalar gerçekleştirilmiş bulunmaktadır. Gültepe ve Kartbaev, BCC veri setini kullanarak göğüs kanseri için veri madenciliği çalışmaları gerçekleştirmişlerdir [3]. Yaptıkları çalışmada J48, çok katmanlı algılayıcılar (multilayer perceptron, MLP), k-en yakın komşuluk (K-nearest neighbor, KNN) ve destek vektör makineleri (support vector machines, SVM) algoritmalarını kullanan ekip J48 algoritması ile en başarılı sonucu elde etmiştir. Çalışmada veri setinin bölünmesi (eğitim ve test olarak) hakkında bilgi verilmemiştir. Bir başka çalışmada, Sharma ve Nair özellik ölçeklendirme, çapraz doğrulama ve torbalama tekniği ile çeşitli topluluk makine öğrenme adımlarını uygulamışlardır [4]. BCC veri setini, %90 eğitim %10 test olacak şekilde bölen ekip, karar ağaçları (decision trees, DT) ve KNN ile en başarılı sonucu elde etmiştir. Arunadevi ve Ganeshamoorthi göğüs kanseri sınıflandırmak için kullanılan öznitelikler arasından özellik seçimi için bir çalışma gerçekleştirmiştir [5]. KNN, SVM ve yapay sinir ağları (artificial neural networks, ANN) algoritmalarının sınıflandırma, rastgele orman (random forest, RF) ve genelleştirilmiş doğrusal model (generalized linear model, GLM) tekniklerinin özellik seçici olarak kullanıldığı çalışmada farklı performans metriklerine göre değerlendirmeler gerçekleştirilmiştir. GLM tekniğinin özellik seçici, SVM algoritmasının sınıflandırıcı olarak seçildiği durumda en başarılı sonuçların elde edildiği gözlemlenmiştir. BCC veri seti kullanılarak gerçekleştirilen bir diğer çalışmada, Saritas ve Yasar, ANN ve naive bayes (NB) algoritmaları ile sınıflandırma gerçekleştirmiş ve performans analizleri yapmıştır [6]. Çalışmada, ANN için veri seti, eğitim (75 örnek), doğrulama (12 örnek) ve test (29 örnek) olarak ayrılırken NB için veri seti, eğitim (93 örnek) ve test (23 örnek) olarak ayrılmıştır. Deneysel çalışmalarda her iki yöntem için başarılı sonuçlar elde edilmiştir. KNN, SVM, DT, RF algoritmalarının yer aldığı 8 farklı yöntemin sınıflandırıcı olarak kullanıldığı çalışmada Salod ve Singh [7], göğüs kanseri tespitinde makine öğrenme algoritmalarının performanslarını karşılaştırmışlardır. Özellik seçimi adımının da uygulandığı çalışmada veri seti, %60 eğitim, %30 doğrulama ve %10 test seti olacak şekilde bölünmüştür. Deneysel çalışmalar farklı performans metriklerine göre değerlendirilmiştir.

Makine öğrenme algoritmalarının yanı sıra gelişen teknoloji ile derin öğrenme yöntemleri farklı problemler [8-16] için literatürde sıklıkla kullanılmaya başlanmıştır. Bir derin öğrenme yöntemi olan yığınlanmış özdevinimli kodlayıcılar (stacked autoencoders, SAE), birden fazla gizli katman içeren ve birbiri ardınca çalışan özdevinimli kodlayıcılar (autoencoders, AE) ile oluşmaktadır. Giriş, gizli ve çıkış katmanından meydana gelen SAE, geri yayılım algoritmasını eğitimde kullanan ve giriş verisini çıkış etiketi olarak atayan bir yöntemdir [10]. Özellik çıkarma, gürültü giderme ve sınıflandırma gibi farklı amaçlarla kullanılabilmektedir.

Sınıflandırma yöntemlerinin performanslarını iyileştirmede, modelde kullanılan verinin ön işlemden geçirilmesi çoğunlukla pozitif yönde katkı sağlamaktadır. Girdi olarak kullanılan parametrelerin değerleri arasında büyük fark olması durumunda, girdi değerlerini düzenleyecek ve belirli bir aralığa göre güncelleyecek normalizasyon ön adımına ihtiyaç duyulmaktadır. Min-max, z-score, medyan ve sigmoid en sık kullanılan normalizasyon teknikleri arasında yer almaktadır.

Bu çalışmada, veri ön işleme adımı ile SAE ve softmax sınıflandırıcının sınıflandırma eğitimi aşamasında kullanıldığı bir model tasarlanmıştır. Önerilen modelin performans analizini gerçekleştirebilmek için SVM, KNN, NB ve DT makine öğrenme algoritmaları ayrıca kullanılmıştır. Bütün deneysel çalışmalar BCC veri seti ile gerçekleştirilmiştir. Doğruluk oranı performans metriğinin yanında modellerin eğitimi ve test sürecinde geçen süreye göre zaman karmaşıklığı analizi yapılmıştır. Deneysel sonuçlara göre veri ön işleme adımında girdi değerlerini 0 ile 1 arasına normalize eden min-max normalizasyon tekniğinin kullanıldığı önerilen yöntemle en başarılı sonuç elde edilmiştir.

Bu çalışmanın temel çıktıları şu şekilde sıralanabilir:

- Veri ön işleme ve softmax sınıflandırıcılı SAE kullanılarak bir sınıflandırma modeli tasarlanmıştır.
- Tasarlanan model ile (normalizasyon destekli) %79,31 doğruluk oranı elde edilerek karşılaştırma yapılan diğer algoritmalar geride bırakılmış ve en başarılı sonuç elde dilmiştir.
- Zaman karmaşıklığına göre, eğitim süresi en kısa olan algoritma KNN iken, test süresi en kısa olan yöntem önerilen metottur.

- Normalizasyon veri ön işleme adımı ile kullanılan algoritmalardan SVM, KNN ve önerilen modelin başarı oranları artırılmıştır.

Bu çalışmada kullanılan materyal ve metotlar Bölüm 2 ile detaylı olarak sunulmuştur. Bölüm 3 ile deneysel sonuçlardan bahsedilmiş olup çalışmanın değerlendirilmesi Bölüm 4' te yapılmıştır.

2. Materyal ve Metot

Bu çalışmada BCC veri seti kullanılarak göğüs kanseri sınıflandırmak için bir SAE tabanlı yöntem tasarlanmıştır. Kullanılan yöntemin performans analizi yapılabilmesi için en yaygın kullanılan makine öğrenme algoritmalarından SVM, KNN, NB ve DT diğer sınıflandırma metotları olarak kullanılmıştır. Sınıflandırma problemlerinde başarıyı artırabilecek veri ön işleme adımları çalışmalarda yoğun bir şekilde kullanılmaktadır. Değerleri 0 ile 1 arasında düzenleyen normalizasyon tekniği bu çalışmada veri ön işleme adımı olarak kullanılmıştır.

2.1. Breast Cancer Coimbra veri seti

Rutin kan analizinde elde edilebilen, "Age", "BMI (kg/m2)", "Glucose (mg/dL)", "Insulin (μU/mL)", "HOMA", "Leptin (ng/mL)", "Adiponectin (μg/mL)", "Resistin (ng/mL)" ve "MCP-1(pg/dL)" nicel öznitelikleri içeren veri seti, sağlıklı (1=Healthy controls) ve hasta (2=Patients) olarak iki etiketten oluşmaktadır. Toplamda 116 örnek içeren veri setinde 64 örnek hastalara aitken geriye kalan 52 örnek sağlıklı bireylere aittir. Deneysel çalışmalarda kullanmak üzere veri seti %75 eğitim %25 test olacak şekilde rastgele bölünmüştür. Her bir aşamada kullanılacak olan örnek sayısı Tablo 1 ile gösterildiği gibidir.

Tablo 1. Örnek sayısı dağılımı.

	Eğitim	**Test**	**Toplam**
Sağlıklı	39	13	52
Hasta	48	16	64
Toplam	87	29	116

2.2. Veri ön işleme

Verilerin daha kaliteli hale getirilmesi için bazı ön işlemlerden geçirilmesi gerekmektedir. Verinin yok sayılması, eksik olanların doldurulması gibi adımlar **veri temizleme** ön işlemiyle gerçekleştirilirken, farklı kaynaklardan verilerin bir araya getirilmesi ve kullanıcıya dönüştürülmüş verinin aktarımı **veri entegrasyonu** ön işlemiyle gerçekleştirilmektedir. Girdi parametrelerinin değerleri arasında çok fark olması halinde giriş değerlerini indirgemek için **normalizasyon** ön işlemi kullanılmaktadır. Bu çalışmada değerlerin 0 ile 1 arasına yayıldığı, Denklem 1. ile hesaplanan min-max normalizasyon yöntemi, ön işlem olarak kullanılmıştır.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

2.3. Makine öğrenme algoritmaları

Destek vektör makineleri (support vector machines, SVM), Vapnik ve arkadaşları tarafından önerilmiştir [17]. Regresyon ve sınıflandırma problemlerinde kullanılabilen bu yöntem, bu çalışmada sınıflandırma işlemi için kullanılmıştır. Yapısal risk minimizasyonu ve istatiksel öğrenme teorisine dayanan bu yöntem uygulaması kolay esnek bir algoritmadır. Bir düzlemde bulunan iki grup arasına bir sınır çizgisi çekilerek iki grup ayrılabilir. Fakat bu sınır çizgisinin bulunacağı konum iki grubun üyelerine en uzak olan konumda olmalıdır. SVM algoritması bu sınırın nasıl çizileceği konusunda devreye girer. Overfitting (ezberleme) sorunun bulunmaması ve yüksek doğruluk oranları gibi avantajlarının yanı sıra olasılıksal tahminler yürütememe ve çekirdek fonksiyonlarının pozitif tanımlı sürekli fonksiyonlar olma zorunluluğu dezavantajları arasında sayılabilmektedir [18]. Bu çalışmada, SVM yöntemi için doğrusal çekirdek fonksiyonunu (linear kernel function) kullanılmıştır.

K-en yakın komşuluk (K-nearest neighbor, KNN) algoritması, en basit sınıflandırma algoritmaları arasında yer almaktadır. Algoritmada öncelikle bir **K** değeri belirlenir. Bu değer, verilen bir noktaya en yakın komşuların sayısıdır. Uzaklık fonksiyonları yardımı ile verilen noktanın (yeni veri), mevcut verilere göre tek tek uzaklığı hesaplanır. Özellik değerlerine göre sınıf ataması (etiketleme) gerçekleşir. Hızlı ve basit bir algoritma olması avantajlarının yanında büyük veriler için kullanıldığında büyük bellek gereksinimine ihtiyaç duyma dezavantajı

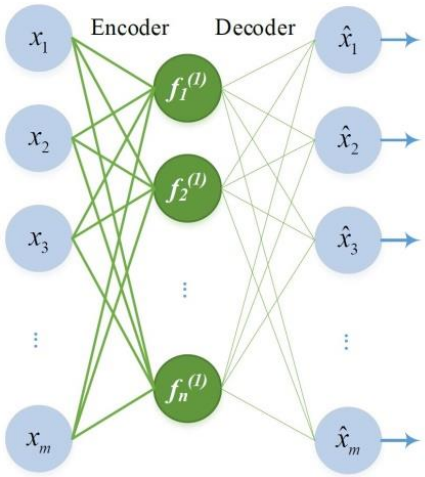
bulunmaktadır [19]. Bu çalışmada, komşuluk sayısı değeri (**K**) 10 olarak belirlenmiş olup, uzaklık metriği olarak 'euclidean' seçilmiştir.

Naive bayes (NB) algoritması, 1812 yılında Thomas Bayes tarafından önerilmiştir. NB' nin temeli Bayes teoremine dayanmaktadır. Algoritmada bir örnek için her durumun olasılığı hesaplanır ve olasılık değeri en yüksek olana göre sınıflandırma işlemi gerçekleştirilir. Az sayıda eğitim verisiyle yüksek başarımlı sonuçların elde edilebileceği bu yöntemle dengesiz veri kümeleri üzerinde de çalışılabilmektedir [20].

Karar ağaçları (decision trees, DT), karmaşık bir işlemin, alt parçalara ayrılıp basit kararlar kümesine dönüştürülmesi esasına dayanmaktadır. Etiketli giriş verileri ile bir veya daha fazla ağaçtan oluşan bir model oluşturulur. Test amacıyla yeni bir veri modele gönderildiğinde, daha önceki eğitimden öğrendiklerine göre yeni verinin sınıfını tahmin eder. Anlaşılması, yorumlaması kolay ve düşük hesaplama karmaşıklığına sahip olma, DT' nin avantajlarından bazılarıdır [21].

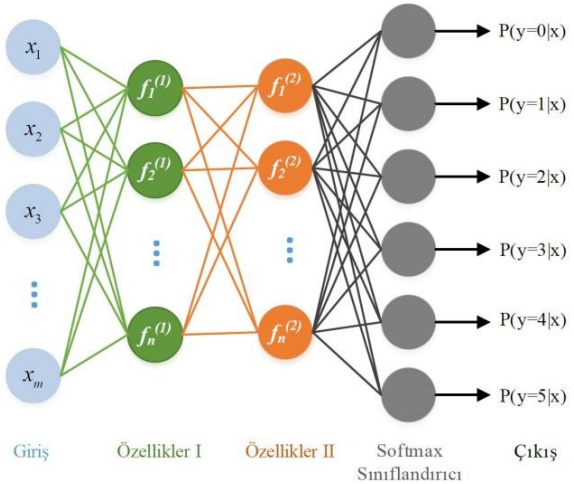
2.4. Yığınlanmış özdevinimli kodlayıcılar

Özdevinimli kodlayıcılar (autoencoders, AE) çıktısında girişin elde edilmeye çalışıldığı bir sinir ağı türüdür. Eğitim sürecinde geri yayılım algoritmasını kullanan bu model, eğitimsiz öğrenme gerçekleştiren ve giriş verisini çıkış etiketi olarak tanımlayan bir derin öğrenme metodudur [10]. Örnek bir AE mimarisi Şekil 1 ile sunulmuştur. Giriş, gizli ve çıkış katmanlarından oluşan AE yapısında kodlayıcı (encoder) ve kod çözücü (decoder) birimler bulunmaktadır.



Şekil 1. Temel bir AE mimarisi [10].

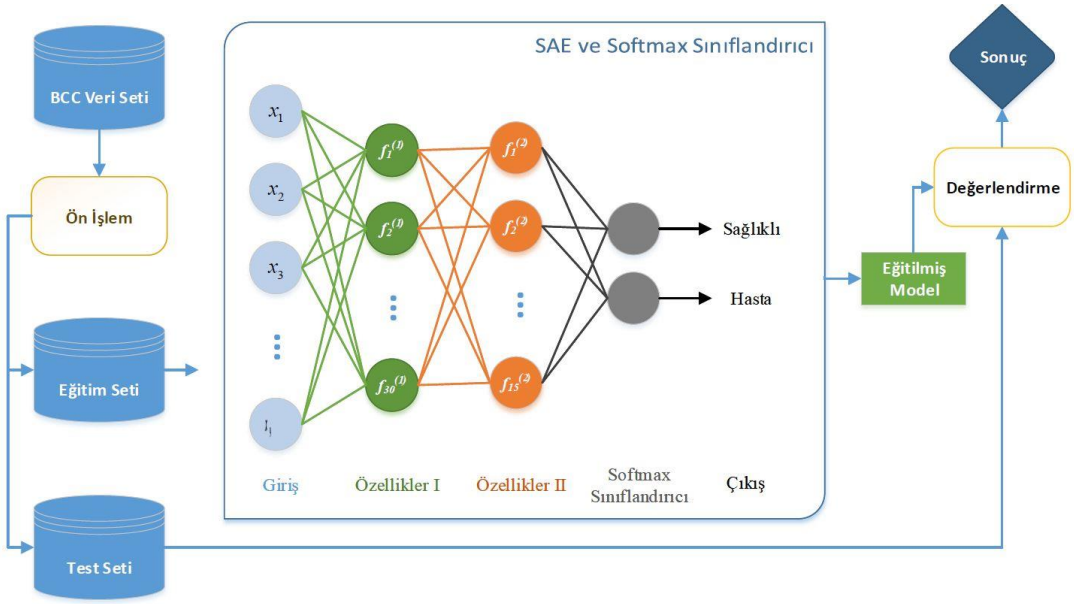
Birden fazla gizli katmanı olan AE mimarisi, yığınlanmış özdevinimli kodlayıcı olarak (stacked autoencoders, SAE) tanımlanmaktadır [22]. Örnek bir SAE mimarisi Şekil 2 ile gösterildiği gibidir.



Şekil 2. Temel bir SAE mimarisi [10].

2.5. Tasarlanan SAE tabanlı sınıflandırıcı

BCC veri setinin sınıflandırılması için SAE tabanlı bir sınıflandırma modeli önerilmiştir. Şekil 3 ile gösterilen modelde, veri setine öncelikle ön işleme adımı uygulanabilmekte daha sonra veri seti eğitim ve test verisi olarak ikiye bölünmektedir. Tablo 2 ile ifade edilen iki AE' den oluşan SAE modeli (AE1 + AE2) ile eğitim seti kullanılarak eğitim işlemi gerçekleştirilir. Daha sonra SAE ve softmax sınıflandırıcıdan oluşan modelin eğitimi ile eğitilmiş model elde edilir. Eğitilmiş model, test seti ile değerlendirme aşamasına girdikten sonra Denklem 2 ile tanımlanan doğruluk oranı metriğine göre sonuç elde edilir.



Şekil 3. Önerilen sınıflandırma modeli.

Tablo 2. Oluşturulan SAE metoduna ait parametreler ve değerleri.

	Parametreler					
	Nöron Sayısı	L2 Düzenlileştirme Katsayısı	Seyreklik Oranı	Encoder Transfer Fonksiyonu	Decoder Transfer Fonksiyonu	Maksimum Epok
AE1	30	0,001	0,05	"logsig"	"purelin"	500
AE2	15	0,001	0,05	"logsig"	"purelin"	500

$$Doğruluk\ oranı = \frac{Doğru\ tahmin\ sayısı}{Test\ seti\ toplam\ örnek\ sayısı} \quad (2)$$

Bir yöntemin hesaplama karmaşıklığı, metodun çalışması için gerekli süreyi ya da bellek alanını ifade etmek üzere kullanılan kavramdır. Çalışmada kullanılan tüm yöntemlere ait hesaplama karmaşıklığı Tablo 3 ile sunulmuştur. Burada N örnek sayısına, M öznitelik sayısına, k komşuluk sayısına, H ağaç yüksekliğine, h gizli nöron sayısına ve K, AE' lerin gizli nöronlar sayısı toplamına karşılık gelmektedir [15].

Tablo 3. Yöntemlere ait hesaplama karmaşıklıkları.

Yöntem	**Hesaplama Karmaşıklığı**
SVM	$O(N^2)$
KNN	$O(MLog(k)NLog(N))$
NB	$O(NM)$
DT	$O(H)$
Önerilen	$O(NK + MhN)$

3. Bulgular

Bu çalışmada BCC veri seti üzerinde SVM, KNN, NB ve DT makine öğrenme algoritmalarının yanı sıra SAE tabanlı derin bir sınıflandırıcı model ile göğüs kanseri tespiti yapılmıştır. Verilerin daha düzenli, daha anlamlı hale getirilebilmesi için ön adımlardan geçirilmesi sonucunda problemin çözümünde genellikle daha başarılı sonuçlar elde edilmektedir. Min-max normalizasyon ön işlemi ile deneysel çalışmalarda kullanılan sınıflandırıcıların çoğunluğunda başarı oranı artmıştır. Kullanılan sınıflandırma algoritmalarına ait doğruluk oranları, eğitim ve test sürecinde geçen zaman (zaman karmaşıklığı) Tablo 4 ile sunulmuştur. Doğruluk oranı ve zaman karmaşıklığı kriterlerinin yanı sıra, duyarlılık ve kesinlik metrikleri de performans değerlendirmesinde kullanılmıştır. Denklem 3 ile ifade edilen duyarlılık (recall, sensitivity) değeri pozitif olarak tahmin edilmesi gereken durumların hangi oranda pozitif olarak tahmin edildiğini ifade etmektedir. Burada TP, hasta olarak tahmin edilen ve gerçekte hasta olan kişi sayısını ifade ederken FN değeri, gerçekte hasta olan ama hasta değil olarak tahmin edilen kişi sayısını ifade etmektedir. Denklem 4 ile ifade edilen kesinlik (precision) değeri pozitif olarak tahmin edilen durumların gerçekte ne kadarının pozitif olduğunu ifade etmektedir. Burada FP, hasta olmayanların hasta olarak tahmin edildiği durum sayısını vermektedir.

$$Duyarlılık = \frac{TP}{TP+FN} \tag{3}$$

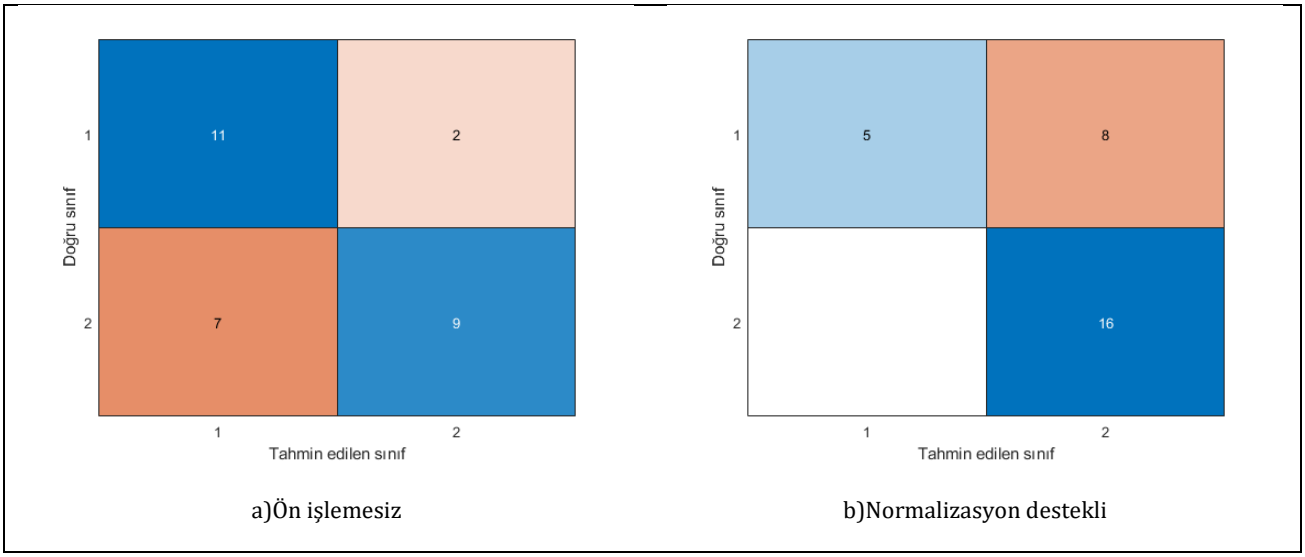
$$Kesinlik = \frac{TP}{TP+FP} \tag{4}$$

Tablo 4 incelendiğinde herhangi bir ön işlem yokken, SVM ve NB algoritmaları makine öğrenme tabanlı algoritmalar içerisinde %68,97 doğruluk oranıyla en başarılı sonuçları verse de ön işlemsiz önerilen modelle %75,86 doğruluk oranı elde edilmiştir. Ön işlem olarak normalizasyon adımı uygulandığında SVM, KNN ve önerilen modelin başarı oranları artarken NB ve DT algoritmalarının başarı oranları düşmüştür. Doğruluk oranı metriğine göre bütün tablo incelendiğinde normalizasyon ön işleminin uygulandığı önerilen yöntemle %79,31 doğruluk oranı ile en başarılı sonuç elde edilmiştir. Bir diğer performans metriği olan zaman karmaşıklığına göre normalizasyon destekli KNN algoritması en kısa eğitim süresine sahipken, normalizasyon destekli önerilen model en kısa test süresine sahiptir. Duyarlılık metriği göz önüne alındığında normalizasyon destekli SVM, KNN ve önerilen model %100,00 değerine ulaşırken, kesinlik metriğine göre ön işlemsiz NB yöntemi ile %88,89 değeri elde edilmiştir.

Tablo 4. Performans değerlendirme tablosu.

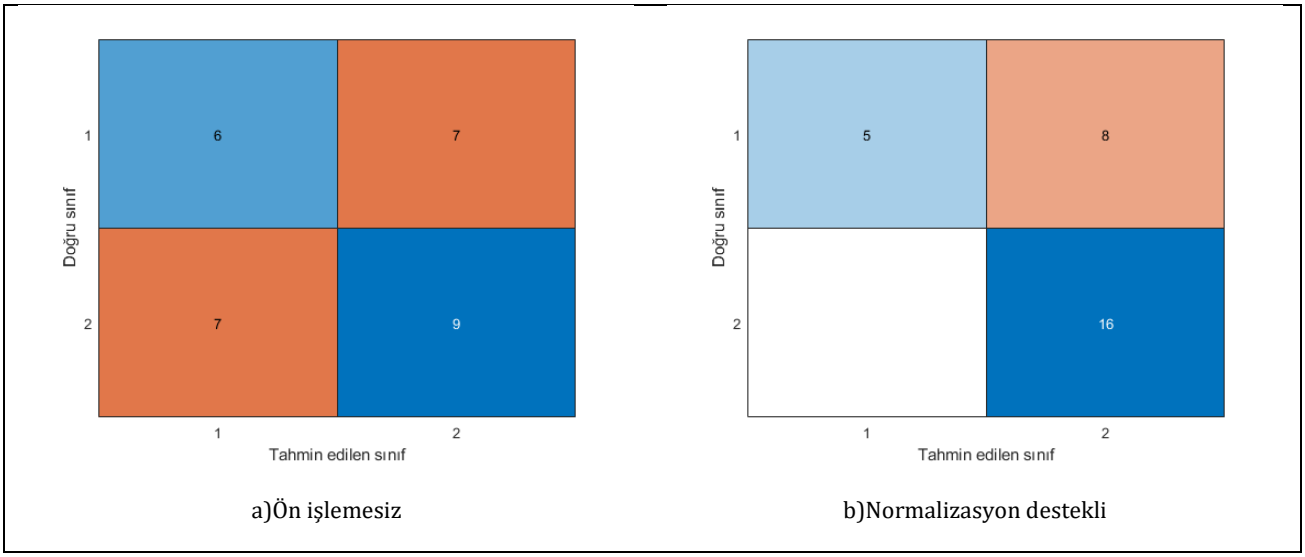
Ön İşleme	Yöntem	Doğruluk Oranı (%)	Duyarlılık (%)	Kesinlik (%)	Eğitim Süresi	Test Süresi
Yok	SVM	68,97	56,25	81,82	1,6315	0,0446
	KNN	51,72	56,25	56,25	0,2608	0,0309
	NB	68,97	50,00	**88,89**	1,8347	0,5359
	DT	62,07	81,25	61,90	1,8705	0,0392
	Önerilen model	75,86	75,00	80,00	1,7143	0,0105
Normalizasyon	SVM	72,41	**100,00**	66,67	0,5528	0,0511
	KNN	72,41	**100,00**	66,67	**0,1626**	0,0198
	NB	65,52	68,75	68,75	0,3085	0,0301
	DT	51,72	75,00	54,55	0,2599	0,0164
	Önerilen model	**79,31**	**100,00**	72,73	1,4843	**0,0086**

Sınıflandırma problemlerinde performans gösteriminin bir diğer versiyonu test seti üzerinde doğru ve tahmin edilen sınıflarla oluşturulan karmaşıklık matrisleridir. BCC veri seti üzerinde SVM, KNN, NB ve DT klasik makine öğrenmelerine ait karmaşıklık matrisleri Şekil 4 – 7 ile sırasıyla verilmiştir.



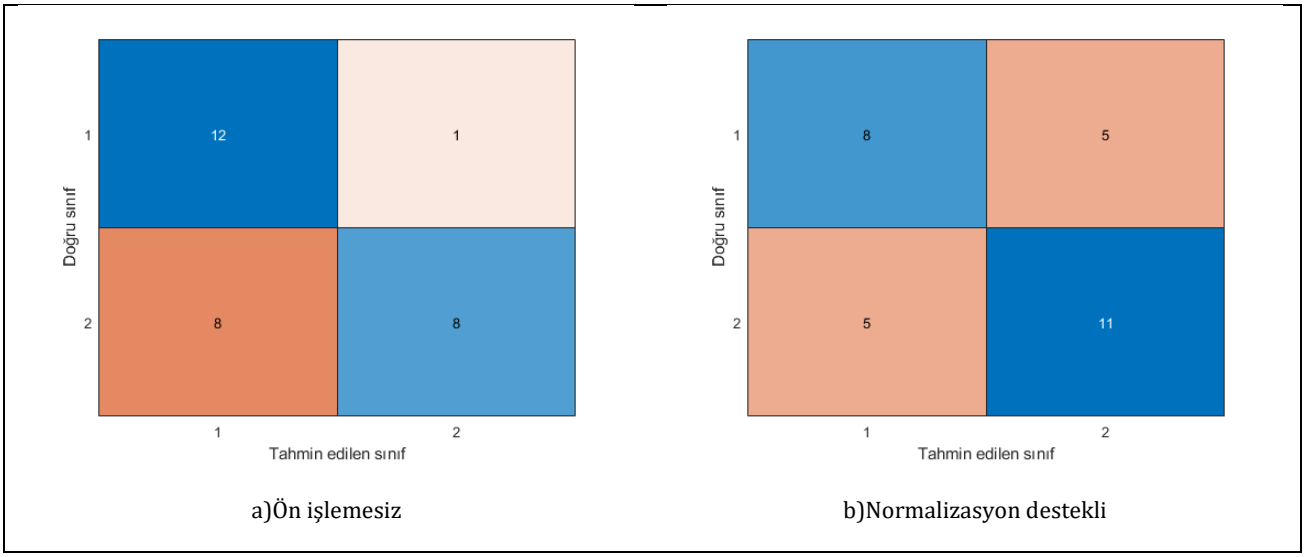
a)Ön işlemesiz b)Normalizasyon destekli

Şekil 4. SVM için veri ön işlemesiz ve normalizasyon destekli koşmalara ait karmaşıklık matrisleri.



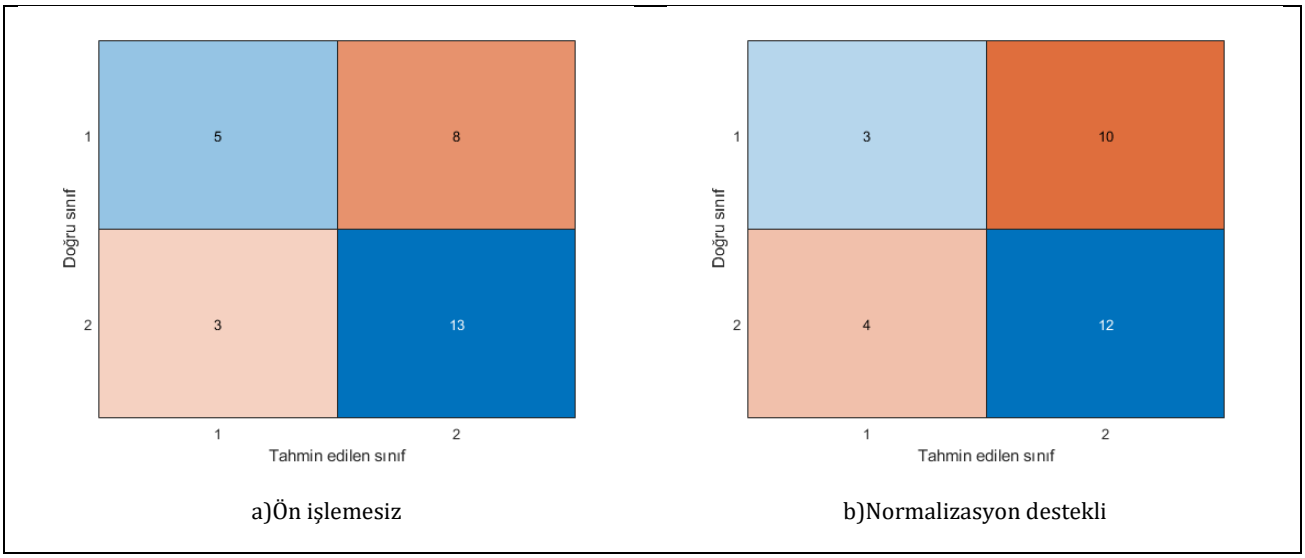
a)Ön işlemesiz b)Normalizasyon destekli

Şekil 5. KNN için veri ön işlemesiz ve normalizasyon destekli koşmalara ait karmaşıklık matrisleri.



a)Ön işlemesiz b)Normalizasyon destekli

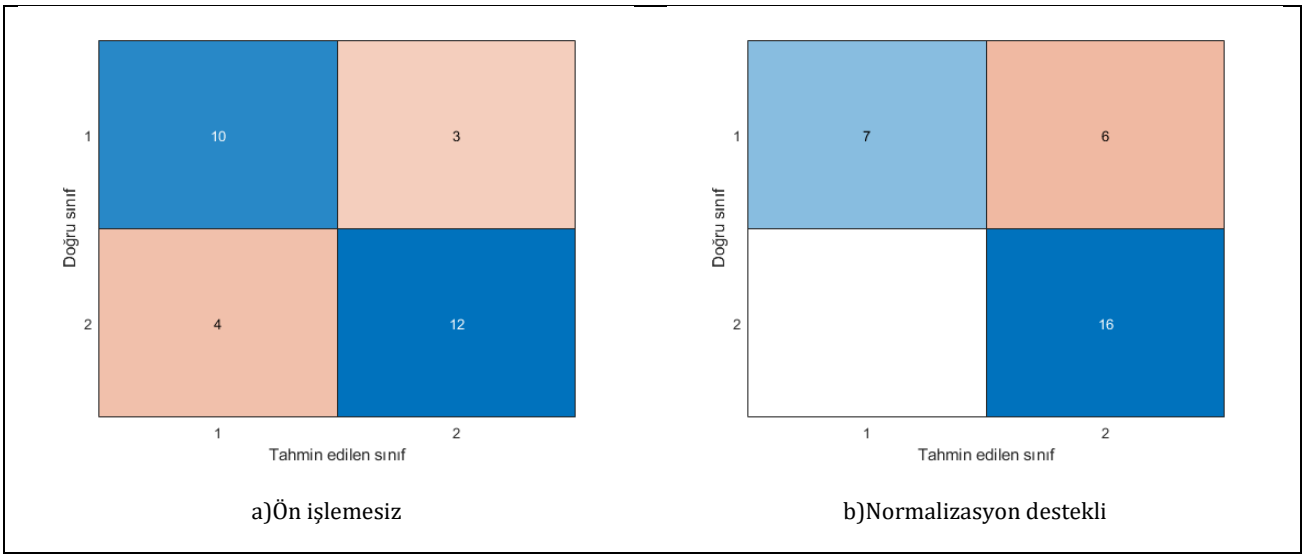
Şekil 6. NB için veri ön işlemesiz ve normalizasyon destekli koşmalara ait karmaşıklık matrisleri.



a)Ön işlemesiz b)Normalizasyon destekli

Şekil 7. DT için veri ön işlemesiz ve normalizasyon destekli koşmalara ait karmaşıklık matrisleri.

Derin öğrenme tabanlı önerilen modelle elde edilen sonuçlara ait karmaşıklık matrisleri Şekil 8 ile sunulmuştur.



a)Ön işlemesiz b)Normalizasyon destekli

Şekil 8. Tasarlanan model için veri ön işlemesiz ve normalizasyon destekli koşmalara ait karmaşıklık matrisleri.

Literatürde yer alan diğer çalışmalar ile önerilen yöntem için deneysel kurulum ve sonuçlara ait bilgiler Tablo 5 ile özetlenmiştir. Bu veri seti için çeşitli sınıflandırma yöntemlerinin çeşitli veri seti ayrımı (eğitim, doğrulama, test) teknikleri ile analiz edildiği görülmektedir.

Tablo 5. BCC veri seti üzerine literatürde yer alan çalışmalara ait özet bilgiler.

Çalışma	Metot	Eğitim-Test Ayrımı	Doğruluk (%)
Gültepe ve Kartbaev [3]	J48	Bilinmiyor	76,92
Sharma ve Nair [4]	DT, KNN	%90 Eğitim - %10 Test	100,00
Arunadevi ve Ganeshamoorthi [5]	GLM, SVM	Bilinmiyor	91,30
Saritas ve Yasar [6]	ANN	75 Eğitim - 12 Doğrulama - 29 Test	86,95
Saritas ve Yasar [6]	NB	93 Eğitim Verisi - 23 Test Verisi	83,54
Tasarlanan Yöntem	SAE	%75 Eğitim - %25 Test	79,31

4. Tartışma ve Sonuç

Göğüs kanseri, kadınlarda en sık görülen kanser türüdür. Hastalığın erken teşhisi, tedavisi için oldukça önem arz etmektedir. Bu çalışmada, bir derin öğrenme metodu olan SAE tabanlı sınıflandırıcı ile kan değerlerinden oluşan BCC veri seti başarılı bir şekilde sınıflandırılmıştır. Metodun başarısını diğer yöntemlerle karşılaştırmak için en yaygın kullanılan makine öğrenme algoritmalarından SVM, KNN, NB ve DT algoritmaları BCC veri setini sınıflandırmada kullanılmıştır. Metotların performansını artırmak için veri ön işleme adımı literatürde sıklıkla tercih edilen bir yöntemdir. Bu çalışmada sınıflandırma metotlarının performansına etkisini gözlemlemek için min-max normalizasyon ön işleme tekniği tüm sınıflandırıcılar için kullanılmıştır. Bu ön işlemin SAE, SVM ve KNN metotlarının doğruluk oranını artırdığı gözlemlenmiştir. Deneysel sonuçlar incelendiğinde tasarlanan SAE tabanlı sınıflandırıcı modelin %79,31 doğruluk oranı ile en başarılı yöntem olduğu görülmüştür. Bu çalışmada ayrıca, yöntemlerin zamansal karmaşıklıkları incelenmiş olup, KNN algoritmasının eğitim aşamasında, tasarlanan yöntemin test aşamasında diğer yöntemlere göre daha hızlı çalıştığı saptanmıştır.

Gelecek çalışmalarda, farklı veri ön işleme adımları kullanılarak yöntemlerin performansına etkisi analiz edilebilir. Ayrıca önerilen yöntem, farklı makine öğrenme algoritmaları ile karşılaştırılabilir veya yöntemlerin performansını ölçmek için farklı veri seti kullanılabilir. Son olarak, önerilen yöntemde SAE yerine farklı bir derin öğrenme metodu kullanılarak çalışmalar gerçekleştirilebilir.

Kaynakça

[1] Bicer , M. B., Aydin, E. A., Akdagli, A. 2014. Meme kanseri görüntülemesinde mikrodalganın yeri. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 30(4), 257-263.

[2] 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).

[3] Gültepe, Y., Kartbaev, T. 2019. A Study of Data Mining Methods for Breast Cancer Prediction. Proceedings Book, 303.

[4] Sharma, R. K., Nair, A. R. 2019. Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models. 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 100-104.

[5] Arunadevi, J., Ganeshamoorthi, K. 2019. Feature Selection Facilitated Classification For Breast Cancer Prediction. International Conference on Computing Methodologies and Communication (ICCMC), 560-563.

[6] Saritas, M. M., Yasar, A. 2019. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), 88-91.

[7] Salod, Z., Singh, Y. 2019. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. Journal of Public Health Research, 8(3), 112-118.

[8] Ozcan, T., Basturk, A. 2020. Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition. Neural Computing and Applications, 31(12), 8955-8970.

[9] Ozcan, T., Basturk, A. 2019. Lip reading using convolutional neural networks with and without pre-trained models. Balkan Journal of Electrical and Computer Engineering, 7(2), 195-201.

[10] Ozcan, T., Basturk, A. 2020. Human action recognition with deep learning and structural optimization using a hybrid heuristic algorithm. Cluster Computing, 1-14.

[11] Ozcan, T., Basturk, A. 2020. Performance Improvement of Pretrained Convolutional Neural Networks for Action Recognition. The Computer Journal, 1-13.

[12] Özcan, T., Baştürk, A. 2019. Static image-based emotion recognition using convolutional neural network. 27th Signal Processing and Communications Applications Conference, 24-26 Nisan, Sivas, 1-4.

[13] Kilic, E., Ozturk, S. 2019. A subclass supported convolutional neural network for object detection and localization in remote-sensing images. International journal of remote sensing, 40(11), 4193-4212.

[14] Gul, E., Ozturk, S. 2019. A novel hash function based fragile watermarking method for image integrity. Multimedia Tools and Applications, 78(13), 17701-17718.

[15] Adem, K., Kiliçarslan, S., Cömert, O. 2019. Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. Expert Systems with Applications, 115, 557-564.

[16] Kilicarslan, S., Adem, K., Celik, M. 2020. Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network. Medical Hypotheses, 137, 109577.

[17] Boser, B. E., Guyon, I. M., Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory; Pittsburgh, Pennsylvania, USA.

[18] Tas, O. 2016. Destek Vektör Makineleri. https://www.slideshare.net/oguzhantas/destek-vektr-makineleri-support-vector-machine (Erişim Tarihi: 13.04.2020).

[19] Ulgen, E. K. 2017. K-En Yakın Komşuluk. https://medium.com/@k.ulgen90/makine-öğrenimi-bölüm-2-6d6d120a18e1 (Erişim Tarihi: 15.04.2020).

[20] Hatipoglu, E. 2018. Naive Bayes. https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4 (Erişim Tarihi: 16.04.2020).

[21] Turgut, S. 2017. Makine öğrenmesi yöntemleri kullanarak kanser teşhisi. İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 55s, İstanbul.

[22] Ozcan, T. 2020. Derin öğrenme ile insan edimlerinin tanınması. Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 204s, Kayseri.