Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2024;16(1):58-70

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2023-100186

RNA Sekanslama Verileri ile Makine Öğrenimi ve Derin Öğrenme Kullanımı: Metodolojik Bir Çalışma

RNA Sequencing Data with Machine Learning and Deep Learning Usage: A Methodological Study

Ragıp Onur ÖZTORNACI[a]

[a]Koç Üniversitesi Translasyonel Araştırma Merkezi, Biyoistatistik ABD, İstanbul, Türkiye

ÖZET Amaç: Bu çalışmanın amacı, klasik istatistiksel yaklaşımlar yerine RNA sekanslama verilerini analiz etmek için popüler makine öğrenimi ve derin öğrenme yöntemlerini kullanarak farklı bir perspektif sunmaktır. Ayrıca makine öğrenimi ve derin öğrenme konularında bilgi sağlamaktır. **Gereç ve Yöntemler:** Makine öğrenimi ve derin öğrenme yöntemlerini kullanarak, astım ve böbrek transplantasyonuna ait iki farklı ham veri seti (GSE85567 ve GSE129166) "National Center for Biotechnology Information" veri tabanından indirilmiş ve gerekli kalite kontrol ve hizalama prosedürlerinden geçirilmiştir. Hasta-kontrol ayrımını elde etmek için rastgele orman [random forest (RF)], destek vektör makineleri [support vector machines (SVM)] ve derin sinir ağları [deep neural networks (DNN)] modelleri uygulanmıştır. Tüm veri setleri aşırı uyumu önlemek amacıyla %67,5 eğitim, %10 test ve %22,5 doğrulama verisi olarak bölünmüş ve modellerin eğitim aşamalarında 10-katlı çapraz geçerlilik kullanılmıştır. Makine öğrenimi ve derin öğrenme için Python programlama dili ve veri işleme için Unix işletim (AT&T Bell Laboratuvarları, ABD) sistemi kullanılmıştır. **Bulgular:** GSE129166 veri setinde RF modelinin validasyon setinde elde ettiği doğruluk oranı 0,89 olarak hesaplanmıştır. Bu modelin hassasiyeti 0,88 ve duyarlılığı 0,92 olarak belirlenmiştir. SVM modeli validasyon setinde elde ettiği doğruluk oranı 0,88 olarak ölçülmüş, test setinde ise 0,87 olarak belirlenmiştir. GSE85567 veri seti için RF modelinin validasyon setinde doğruluk oranı 0,73 olarak ölçülmüştür. SVM için validasyon setinde doğruluk oranı 0,70 olarak ölçülmüş, DNN için ise 0,75 olarak ölçülmüştür. **Sonuç:** GSE85567 veri seti üzerinde yapılan çalışma, RF ve SVM modellerinin yüksek doğruluk ve performans sergilediğini göstermektedir. DNN modeli ise daha dengeli bir hassasiyet ve duyarlılık oranına sahip olup, önemli bir alternatif olarak gözlemlenmiştir. Üç modelin RNA-sekanslama verileri için hasta-kontrol sınıflaması için uygun olduğu sonucuna varılmıştır.

Anahtar kelimeler: RNA sekanslama verileri; makine öğrenimi; derin öğrenme

ABSTRACT Objective: The aim of this study is to provide a different perspective on the analysis of RNA sequencing data by employing popular machine learning and deep learning methods, rather than classical statistical approaches. Additionally, it aims to provide insights into machine learning and deep learning concepts. **Material and Methods:** Utilizing machine learning and deep learning techniques, two distinct raw datasets pertaining to asthma and kidney transplantation (GSE85567 and GSE129166) were retrieved from the National Center for Biotechnology Information database and subsequently subjected to requisite quality control and alignment procedures. Random forest (RF), support vector machines (SVM), and deep neural networks (DNN) models were implemented to achieve patient-control differentiation. To prevent overfitting, all data sets were divided into 67.5% training, 10% testing, and 22.5% validation data, and 10-fold cross-validation was employed during the training stages of the models. Python programming language was used for both machine learning and deep learning, and Unix operating (AT&T Bell Laboratories, USA) system was utilized for data processing. **Results:** In the GSE129166 data set, the RF model obtained an accuracy rate of 0.89 in the validation set. The precision and recall of this model were determined as 0.88 and 0.92, respectively. The SVM model measured an accuracy rate of 0.88 in the validation set, and 0.87 in the test set. For the GSE85567 data set, the accuracy rate of the RF model in the validation set was measured as 0.73. For SVM, the accuracy rate in the validation set was measured as 0.70, while for DNN, it was measured as 0.75. **Conclusion:** The study conducted on the GSE85567 data set demonstrates that RF and SVM models exhibit high accuracy and performance. The DNN model, on the other hand, has a more balanced precision and recall rate, and is observed to be a significant alternative. Additionally, it is observed that the DNN model shows effective performance on the GSE129166 data set. Particularly, a high accuracy rate and a balanced precision-recall balance were observed in the validation set. It is concluded that all three models are suitable for patient-control classification in RNA-seq data.

Keywords: RNA-sequencing; machine learning; deep learning

KAYNAK GÖSTERMEK İÇİN:
Öztornacı RO. RNA sekanslama verileri ile makine öğrenimi ve derin öğrenme kullanımı: Metodolojik bir çalışma. Turkiye Klinikleri J Foren Sci Leg Med. 2024;16(1):58-70.

Correspondence: Ragıp Onur ÖZTORNACI
Koç Üniversitesi Translasyonel Araştırma Merkezi, Biyoistatistik ABD, İstanbul, Türkiye
E-mail: onur.oztornaci@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

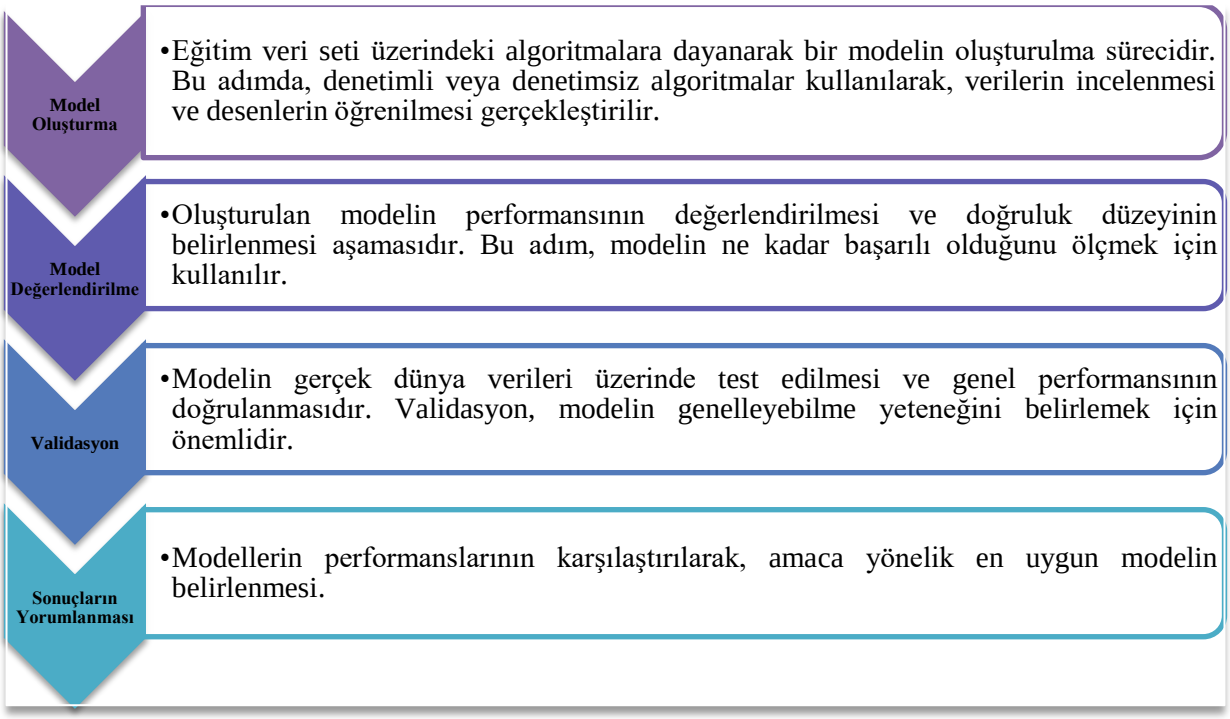
Received: 03 Nov 2023 ***Received in revised form:*** 06 Dec 2023 ***Accepted:*** 11 Dec 2023 ***Available online:*** 05 Jan 2024

Yeni nesil dizileme teknolojisi ve ribonükleik asit dizileme yöntemleri (RNA-sekanslama) son yıllarda genetik araştırmalarda önemli bir ivme kazanmıştır. RNA-sekanslama, transkriptomun (bir organizmanın belirli bir anda ekspresyon gösteren tüm RNA'ları) sistematik olarak analiz edilmesini sağlayan yüksek çözünürlüklü bir tekniktir. Bu teknik, mRNA moleküllerinin sayısız özelliklerini belirleyebilir, böylece hücresel işlevlerin, organizmalardaki gelişim süreçlerinin ve çevresel streslere verilen yanıtların anlaşılmasında derinlemesine bir perspektif sağlar. RNA-sekanslama teknolojisinin getirdiği olanaklarla, hücresel düzeyde gen ekspresyonunu anlayarak, belirli özelliklerin nedenlerini araştırmak ve hastalığa neden olabilecek genlerin ortaya çıkarılması hedeflenmektedir. Ayrıca RNA-sekanslama verileri ile gen ifadesi profilleri (ekspresyonlar) ve proteinlerin kodlanması sürecinde meydana gelen alternatif uç birleştirme (alternative splicing) süreçleri de belirlenebilmektedir. Bu, yeni genlerin keşfi ve gen düzenleyici bölgelerin haritalanmasının biyoinformatik analizlerle mümkün olacağı anlamına gelmektedir.[1] Bu teknolojiler, genetik araştırmalarda ve hastalıkların moleküler mekanizmalarının anlaşılmasında kritik bir rol oynamaktadır. Geleneksel istatistiksel yaklaşımların ötesine geçerek, popüler makine öğrenimi (MÖ) ve derin öğrenme (DÖ) yöntemlerinin uygulanması, bu veri setlerinden daha kapsamlı ve anlamlı bilgiler elde etme fırsatı sunmaktadır. MÖ yöntemleri, herhangi bir hipoteze dayanmaksızın algoritmaların verilerden karmaşık yapıdaki örüntüleri keşfetmesi amacıyla kullanılmaktadır. Dolayısıyla RNA-sekanslama verileri ile MÖ ve DÖ yöntemlerini kullanmak giderek yaygınlaşmaktadır.[2]

GEREÇ VE YÖNTEMLER

MAKİNE ÖĞRENİMİ TANIMLARI

Makine öğrenimi, bilgisayar sistemlerinin veri analizi, desen tanıma ve örüntü çıkarma gibi temel teknikler kullanarak, verilerden öğrenme ve deneyim kazanma kabiliyetine sahip olduğu önemli bir yapay zekâ alt alanıdır.[3] Bu teknikler, algoritmaların eğitim veri setleri üzerinde çalışarak, verilerdeki yapıları, ilişkileri ve özellikleri belirlemesine ve bu bilgileri modellemeye yardımcı olur. Veri analizi, makine öğrenimi sürecinde veri setlerinin incelenmesini ve anlamlı özelliklerin belirlenmesini sağlar. Veri analizinin temel amacı, verilerdeki yapıları anlamak ve öğrenme sürecini optimize etmektir. Desen tanıma, verilerdeki tekrar eden ve önceden belirlenmiş yapıları tanımak ve bu desenleri belirlemek için kullanılır. Örüntü çıkarma ise verilerdeki gizli özellikleri veya yapılardaki ilişkileri belirlemede önemli bir rol oynar ve bu özelliklerin öğrenme sürecini geliştirmeye yardımcı olur.[4] Makine öğrenimi algoritmaları temel olarak, denetimli öğrenme (sınıflama) ve denetimsiz öğrenme (kümeleme) olarak iki farklı işlev doğrultusunda incelenebilir. Makine öğrenimi aşamaları ise model oluşturma, modelin değerlendirilmesi, validasyon ve yapay zekâya entegrasyonu olarak sıralanabilir.[5] Makine öğrenimi algoritmasının daha iyi ve doğru tahminler yapabilmesi için veriler; eğitim veri seti (training set) ve test veri seti (testing set) olarak 2 gruba ayrılır. Eğitim veri seti, algoritmanın öğrenme sürecinde kullanılan verilerdir ve genellikle verilerin büyük bir kısmını içerir. Test veri seti ise algoritmanın öğrendiklerini değerlendirmek ve tahmin başarısını ölçmek için kullanılan ayrı bir veri kümesidir. Bu ayrılma oranı kesin bir kural olmasa da genellikle eğitim veri seti %67, test veri seti ise %33 olarak belirlenir. Amaç, algoritmayı daha etkili ve yüksek başarı oranıyla kullanabilmek ve gelecekte benzer veri kümeleriyle karşılaşıldığında doğru sonuçlar elde etmek için modelin eğitimini ve performansını değerlendirmektir (Şekil 1).[6,7]



ŞEKİL 1: Makine öğrenimi aşamaları.

DENETİMLİ ÖĞRENME ALGORİTMALARI

Denetimli öğrenme, algoritmanın eğitim veri seti üzerinde etiketlenmiş verileri kullanarak bir ilişki veya deseni öğrenmesini içerir. Bu tür algoritmalar, verilerdeki bağımlı değişkenleri tahmin etmek veya sınıflandırmak için kullanılır. Örnek olarak, e-postaların spam veya spam olmayan olarak sınıflandırılması veya hasta verilerinin hastalık teşhisine göre etiketlenmesi denetimli öğrenme örnekleridir.[6]

RASTGELE ORMAN ALGORİTMASI

Rastgele ormanlar [random forest (RF)] 2001 yılında Breiman tarafından ortaya atılmıştır. Bu yöntem, birçok rastgele karar ağacının birleştirilmesiyle oluşturulur. Her ağaç, aynı dağılıma sahip olacak şekilde bağımsız olarak örneklenir. Temel amacı daha yüksek bir doğruluk değeri (accuracy) elde etmektir. RF oluşturulurken, her ağaç için orijinal veri setinden farklı bootstrap örneklem seçimi yapılır. CART algoritması temel alınarak, dallara ayrılma kriteri olarak bilgi kazancı kullanılır ve uygun sınıflandırma için gini katsayısı kullanılır. Oluşturulan ağaçlar için budama yapılmaz. Bu yöntem, fazla sayıda değişken içeren veri setleri için de kullanılabilir. Ancak bazı programlar, oluşturulan ağaçları geçici bellekte tuttuğu için iyi donanıma sahip yüksek geçici belleği olan bilgisayarlarda kullanılması bir dezavantaj olarak görülebilir.[8]

DESTEK VEKTÖR MAKİNELERİ

El yazısı tanıma, zaman serisi analizi, konuşma tanıma, kanser teşhisi, yapısal protein sınıflandırmasının tahmini gibi birçok alanda başarılı bir şekilde uygulanan destek vektör makineleri [support vector machine (SVM)], makine öğrenimi metotlarından birisidir. İlk olarak 1992 yılında Vapnik ve ark., tarafından sunulmuştur.[9] SVM, diğer sınıflama yöntemleriyle karşılaştırıldığında, eğitim süresinin uzun olmasına rağmen yüksek güvenilirliği ve ezberlemeye karşı olan direnci sayesinde sıkça tercih edilen bir sınıflandırma yöntemidir. SVM'yi iki ana başlık altında toplayabiliriz: Doğrusal SVM (linear support vector machine) ve doğrusal olmayan SVM (nonlinear support vector machine). Verilerin yayılımı doğrusal olduğunda doğrusal SVM kullanılırken, yayılımı doğrusal değilse doğrusal olmayan SVM tercih edilir. Sınıflandırma için iki grup ara-

sında bir sınır çizilerek, bu iki grubu ayırmak mümkündür. Bu sınır, iki grubun da üyelerine en uzak olan yerde çizilmelidir. SVM, bu sınırın nasıl çizileceğini belirler.

MODEL OLUŞTURMA VE MODEL DEĞERLENDİRME

Model oluşturma süreci, genellikle üç aşamadan oluşur: eğitim aşaması, öğrenme aşaması ve modelin geçerliliği aşaması;

1. **Eğitim Aşaması:** Bu aşamada, model için eğitim verileri kullanılarak algoritma üzerindeki parametreler optimize edilir. Eğitim verileri, modelin belirli bir görevi öğrenmesini sağlamak için kullanılır. Ancak bu verilerin modelin performansını gerçek dünya verileri üzerinde değerlendirmek için kullanılması tavsiye edilmez, çünkü model eğitim verilerine aşırı uyum sağlayabilir (overfitting) ve genelleme yeteneği düşebilir.
2. **Öğrenme Aşaması:** Model, eğitim verileri üzerindeki algoritma tarafından öğrenilir ve belirli bir görevi gerçekleştirecek şekilde ayarlanır. Model, girdi verilerini işleyerek ve özelliklerini çıkararak sonuçları üretme yeteneğini kazanır.
3. **Modelin Geçerliliği Aşaması:** Modelin performansını test etmek ve genelleme yeteneğini değerlendirmek için farklı yöntemler kullanılır. Bu aşamada kullanılan veriler, eğitim verilerinden ayrı olan test verileridir. Aşağıda, modelin geçerliliğini değerlendirmek için kullanılan 4 farklı yöntem açıklanmıştır:

- **Holdout (Durdurma):** Veri kümesi rastgele iki parçaya ayrılır: eğitim verileri ve test verileri. Model eğitim verileriyle eğitilir ve ardından test verileri üzerinde değerlendirilir.
- **K-katlı Çapraz Doğrulama (K-fold Cross-Validation):** Veri kümesi k adet alt kümeye bölünür. Ardından, model k defa eğitilir ve her seferinde farklı bir alt küme test verisi olarak kullanılır. Sonuçlar ortalama alınarak modelin performansı değerlendirilir.
- **Birini Dışarıda Bırakma (Leave-One-Out):** Her örnek tek başına test verisi olarak kullanılırken, diğer tüm veriler eğitim verisi olarak kullanılır. Bu işlem tüm veri noktaları için tekrarlanır ve sonuçlar ortalaması alınarak modelin performansı değerlendirilir.
- **Özyükleme (Bootstrapping):** Veri kümesinden rastgele örneklem alınır ve bu örneklem üzerinde model eğitilir. Bu işlem birden fazla kez tekrarlanır ve sonuçlar ortalaması alınarak modelin performansı değerlendirilir.

Doğruluk (accuracy), modelin doğru tahminlerin tüm tahminlere oranını ifade eder. Duyarlılık (sensitivity), gerçek pozitiflerin toplam pozitiflere oranını temsil eder. Özgüllük (specificity), gerçek negatiflerin toplam negatiflere oranını ifade eder. Karışıklık matrisi (confusion matrix), modelin sınıflandırma performansını görselleştirmek için kullanılır ve gerçek sınıf ile tahmin edilen sınıfın karşılaştırmasını sağlar. Bu kavramlar, modelin performansını değerlendirmede önemli bir rol oynar. Bazı veri madenciliği algoritmaları, duyarlılık değeri (sensitivity) yüksekken, diğerlerinde nispeten daha düşük olabilir. Benzer şekilde, bazı algoritmaların doğruluk değerleri (accuracy) eşitken, duyarlılık (sensitivity) ve özgüllük (specificity) değerleri farklılık gösterebilir. Bu nedenle, model seçimi yaparken, belirli bir uygulama veya senaryoya en uygun performans metriklerine dikkat etmek önemlidir.[10]

TABLO 1: Karmaşa matrisi (confusion matrix).

Durum		**Gerçek Durumdaki Sınıf**	
		Pozitif	**Negatif**
Tahmin sınıfı	Pozitif	Gerçek pozitif (a)	Yanlış pozitif (b)
	Negatif	Yanlış negatif (c)	Gerçek negatif (d)

Yukarıdaki tablonun tümüne karmaşa matrisi (confusion matrix) adı verilmektedir (Tablo 1).

$$\text{Duyarlılık (sensitivity)} = \frac{Gerçek\ Pozitif\ (a)}{Gerçek\ Pozitif + Yanlış\ Pozitif\ (a+c)} \quad (1)$$

$$\text{Özgüllük (specificity)} = \frac{Gerçek\ \ Negatif\ (d)}{Gerçek\ Negatif + Yanlış\ Negatif\ (b+d)} \quad (2)$$

$$\text{Doğruluk (accuracy)} = \frac{Gerçek\ pozitif(a) + Gerçek\ Gegatif(d)}{N} \quad (3)$$

$$\text{Negatif bilgi oranı (no information rate)} = \frac{Gerçek\ Negatif(d) + Yanlış\ Pozitif(b)}{N} \quad (4)$$

$$\text{Yakalama oranı (decection rate)} = \frac{Gerçek\ Pozitif\ (a)}{N} \quad (5)$$

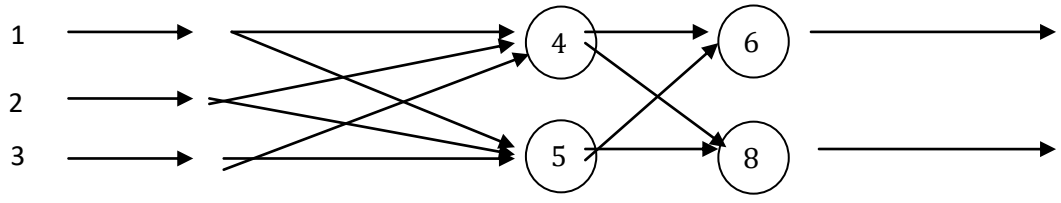
$$\text{Kesinlik} = \frac{Gerçek\ Pozitif\ (a)}{Gerçek\ Pozitif\ (a) + Yanlış\ Pozitif\ (b)} \quad (6)$$

$$\text{F1-Skoru} = 2^{*}\ \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \quad (7)$$

formülleri ile hesaplanmaktadır.[10]

DÖ YÖNTEMLERİ VE KÜTÜPHANELERİ

Makine öğrenimin alt dalı olan DÖ, yapay sinir ağları (YSA) gibi doğadan ilham alınarak geliştirilen algoritmalar topluluğudur. Diğer bir ifade ile DÖ temel olarak makine öğreniminde de kullanılan YSA algoritmalarının bir bütünü hâlindedir ve çok sayıda gizli katmandan oluşmaktadır. Makine öğreniminden farklı olarak özellik seçimine (feature selection) dolayısıyla da veri ön işleme sürecine gerek duymaz.[11] Konu bu perspektiften incelenmek isterse, makine öğrenimi ile DÖ yöntemlerini doğru sınıflama oranı gibi performans ölçütleri ele alınarak aradaki farkların kavranması açısından bilgi edinilebilir. Makine öğrenimi için özellik seçimi esnasında araştırmacı tarafından kullanılan korelasyon, ki-kare testi gibi veri ön işleme yöntemleri yerine, DÖ algoritmaları için gizli katmanlar (hidden layer), "batch" ve "epoch" sayıları gibi DÖ'ye ait kavramların kullanılabilir ve bu DÖ kavramları, DÖ algoritmalarının kendi kendine öğrenme sürecine etki ederek, algoritmaların performanslarını değiştirdiği söylenebilir.[12] Gizli katmanların sayısının artırılması yanı sıra doğru şekilde ağırlıklandırma ile optimize edilerek de doğrusal olmayan sınıflama problemlerin çözüm gücü artırılabilir. Gizli katmanlar ve YSA'ya kısaca göz atmak istersek; bir YSA temel olarak 3 katmandan oluşmaktadır. Bu katmanlar, giriş katmanları, gizli katmanlar ve çıkış katmanları olarak adlandırılırlar. Bu çerçevede, YSA incelenecek olursa; 1, 2 ve 3 numaralı katmanlar giriş katmanları, 6 ve 7 numaralı katmanlar çıkış katmanları, 4 ve 5 numaralı katmanlar ise ağ içerisinde kalan ve iletimi sağlayan gizli katmanlardır (Şekil 2).[11,12]

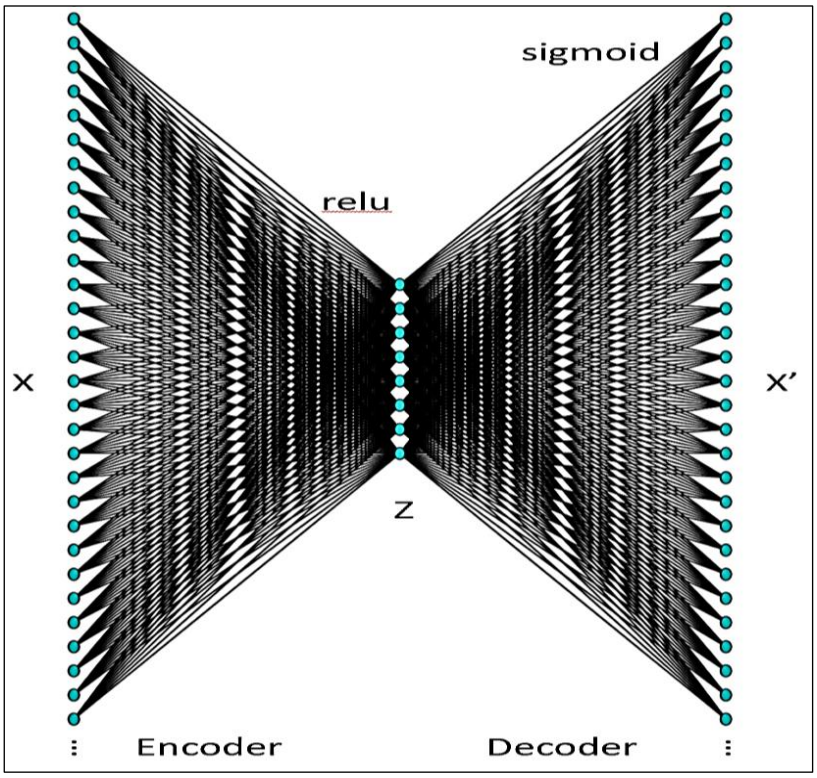


ŞEKİL 2: Yapay sinir ağları.

Birçok farklı alanda kullanılan DÖ yöntemleri için farklı kütüphaneler mevcuttur. Genellikle yüksek hesaplama gücü gerektiren DÖ algoritmaları için yüksek donanım gerektiren bilgisayarlar ve iyi optimize edilmiş kütüphanelere ihtiyaç vardır. Kullanılan kütüphanelerin önemli olanlarını listelemek istersek; Microsoft Cognitive Tool Kit (CNTK) (Microsoft, ABD), Tensorflow, Keras, Caffe, Torch, ve MXNet olarak hem Python (Python Software Foundation, Amsterdam) programlama dilinde hem de R programlama dilinde kolayca erişilebilen kütüphaneler olarak karşımıza çıkmaktadırlar. Bu kütüphanelerden yalnızca CNTK kütüphanesinin Python programlama dilinde kullanımı mevcuttur.[13]

DÖ KAVRAMLARI

YSA'dan farklı olarak DÖ için "encoder", "decoder", "deep autoencoder", "batch", "epoc", "tensor", ağırlıklar, aktivasyon fonksiyonları gibi farklı kavramlar vardır. Bu kavramları görsel olarak aşağıdaki şekilde incelememiz mümkündür.



ŞEKİL 3: Derin öğrenme kavramları.

Encoder ve Decoder: Yukarıdaki şekil dikkatli incelenirse, bir DÖ modeli için girdiler kodlayıcı bir fonksiyon oluşturularak gizli katmanlara iletilir ve oradan da çıktı (output) elde edebilmek için şifreleme işleminin çözümünü gerektiren bir fonksiyon olarak "Decoder" kavramına başvurulur ve çıktı (output) elde edilir (Şekil 3).[12-14] Autoencoder (otomatik kodlayıcı) denetimsiz DÖ (unsupervised deep learning) modellerinde de kullanılırlar ve amaç veri boyutunu azaltmaktır. Boyut azaltılırken hem encoder hem de decoder aynı işlem içerisinde otomatik olarak yapılmaktadır. Matematiksel ve istatistiksel olarak kullanılmak istenirse boyut indirgeme problemlerinde kullanılabilirler.[14]

Batch: DÖ modelinde aynı anda eğitim verisindeki tüm özellikleri dâhil etmek yerine verileri parçalar hâlinde dâhil etme işlemine batch adı verilir.[14,15]

Epoch: Bir DÖ modelindeki eğitim setinde yer alan elemanlarının tümünün birden kullanılarak (hidden layerdan) sinir ağından geçip bir kez ileri ve geri aktarılmasıdır.[16]

Tensör: Matematiksel olarak bir tensör çok boyutlu verilerin geometrik olarak temsilidir.[15,16]

Ağırlıklar: Ağırlıklar, iki nöron arasındaki sinyali veya bağlantının gücünü kontrol eder. Başka bir deyişle, bir ağırlıklar ile girdinin çıktı üzerinde ne kadar etkisi olacağına karar verir.[16,17]

Aktivasyon Fonksiyonu: Ağırlıklandırılmış verilerin toplanıp gizli katmanlardan geçmesi için gerekli olan işleme aktivasyon fonksiyonu adı verilir ve çözülmek istenen problemin türüne göre doğrusal veya doğrusal olmayan birçok aktivasyon fonksiyonu mevcuttur.[16,17]

Keras kütüphanesi içerisinde birçok farklı model vardır. Bunların başlıca kullanılanlarını aşağıdaki gibi sıralamak mümkündür;

- Sequential (Sıralı) Model
- Functional (Fonkskiyonel) API
- Model Subclassing (Alt Sınıflara Bölünen Model)

Sequential (sıralı) model, her katmanın tam olarak bir giriş tensörüne ve bir çıkış tensörüne sahip olduğu düz bir katman yığını için uygundur, dolayısıyla girdi değişkenlerini birebir olarak ağırlıklandırmak mümkündür.[17]

Buraya kadar anlatılan DÖ bilgilerini matematiksel olarak ifade etmek istersek:

$$\Phi(X, W^1, \ldots\ldots W^K) = \Psi_K(\Psi_{K-1}(\Psi_2(X * W^1) * W^2) \ldots W^{K-1}) * W^K \quad (8)$$

Burada (8), Φ simgesi NxC boyutunda bir matris, C= d_K çıktıların (output) ağdaki boyutları (sınıflandırılmak istenen sınıf sayısına eşit olmak üzere), W=$\{W^K\}_{k=1}^K$ ağdaki X girdilerinin ağırlıklarıdır. Ağırlıkları eğitim aşamasında optimizasyonu için kullanılan formülasyon ise

$$\min_{W=\{W^K\}_{k=1}^K} \ell(Y, \Phi(X, W^1, \ldots\ldots W^K)) + \lambda\, \theta(W^1, \ldots\ldots W^K) \quad (9)$$

Şeklinde (9) tanımlanır. Burada (28), $\ell(Y, \Phi)$ terimi loss (zarar/kayıp) fonksiyonu olarak tanımlanır. Φ ise aşırı uyum sorunun ortadan kaldırmak için kullanılan bir regülasyon terimidir ve Φ(W)=$\sum_{k=1}^K ||W^k||_F^2$ ile hesaplanır ve loss fonksiyonu $\ell(Y, \Phi) = ||Y - \Phi||_F^2$ ile hesaplanır ve

$$\min_w \ell(Y, \sum_i h_i(X) w_i) + \lambda\, ||w|| \quad (10)$$

Denkleminin (10) sonucuna göre sınıflandırma yapılır, burada $h_i(X)$ olası bir gizli birim aktivasyonunu temsil eder ve X eğitim veri setidir ve $\lambda > 0$ denge parametresidir.[18]

DÖ ALGORİTMALARI

Makine öğrenimi gibi DÖ de hem sınıflama (gözetimli öğrenme) hem de kümeleme (gözetimsiz öğrenme) problemlerinde kullanılabilir. Sınıflama problemleri için en sık kullanılan DÖ algoritmaları; derin sinir ağları [deep neural network (DNN)], evrişimli sinir ağları [convolutional neural network (CNN)] ve yinelenen sinir ağları [recurrent neural network (RNN)] olarak karşımıza çıkmaktadırlar. Kümeleme problemleri için kullanılan algoritmalar ise özdüzenleyici haritalar [self organizing maps (SOM)], Boltzman Makineleri [boltzman machines (BM)], otokodlayıcı [autoencoder (AE)] olarak karşımıza çıkmaktadırlar.[16-18]

DNN

DNN, temel olarak çok katmanlı algılayıcı [multilayer perceptron (MLP)] yapısına benzer şekilde ancak birden çok katmana sahip bir yöntemdir.[19] MLP, YSA yapısının tek bir gizli katmana sahip olan ve ileri beslemeli (feed-forward) yapıda bilgi akışına göre sınıflama yapan bir makine öğrenimi modelidir. Dolayısıyla bir derin sinir ağı yapısı da birden çok katman içeren ve bilgi akışının girdi katmanından çıktı katmanına doğru olduğu (feed-forward) bir yöntemdir. MLP'lerin, YSA yapısındaki gizli katman sayısının az olmasından ötürü ve DÖ'deki epoch ve batch yapıları gibi yapılara sahip olmadığından ötürü sınıflama başarısı DÖ'ye göre daha düşük olduğu söylenebilir.[20]

CNN

Görüntü işleme, ses tanıma ve otonom sürüş gibi birçok alanda kullanılan CNN yöntemi genel olarak 5 katmandan oluşmaktadır ve ileri beslemeli (feed-forward) bir derin öğrenim yöntemidir. Bu katmanlar sırasıyla girdi katmanı, evrişim katmanı, havuzlama katmanı, tam bağlaşımlı katman ve çıktı katmanıdır.[21] CCN'yi diğer metotlardan ayıran kısmı evrişim katmanına sahip olmasıdır. Matematiksel olarak integral işlemi ile hesaplanan ve fonksiyonel analizde de kullanılan evrişim katmanı, görüntü işleme ve sinyal algılama gibi alanlarda kullanılmaktadır. Evrişim (convolution) fonksiyonu yardımıyla, DÖ yöntemine yeni bir mimari ekleyerek kullanan bu katman, doğru sınıflama oranı gibi ölçütlere etki etmektedir.[22]

RNN

Metin madenciliği özellikle de doğal dil işlemede (natural language process) oldukça sık kullanılan bir DÖ yöntemidir.[23] Diğer yöntemlerden farklı olarak, işlenecek olan veri kümesinde tekerrür eden veriler varsa, örneğin metin madenciliğinde, aynı cümle içerisinde aynı kelime birden fazla kez tekrar ediliyorsa, doğru sınıflama yapabilmek adına bilginin bir önceki katmandan bir sonraki katmana aktarımı birbirine bağımlıdır ve bu yapıdaki tekrarlardan dolayı, katmanlarda da ileri beslemeli (feed-forward) yapısı yerine tekrarlayan bir yapı söz konusudur.[24]

SOM

Denetimsiz öğrenme algoritması olan SOM yöntemi ile DÖ yöntemleri kullanılarak, çok boyutlu verilerde boyut indirgeme işlemleri yapılabilir.[25] İki katmanlı ileri beslemeli (feed-forward) bir yöntemdir. Daha çok verilerinin birbirlerine benzer niteliklerini ortaya koymak adına kümeleme problemlerine grafiksel bir çözüm getirir.[26]

BM

Yapısında bulunan YSA'nın eğitim aşamasında en çok olabilirlik (maximum likelihood) yöntemini kullanarak kümeleme problemlerini çözmeyi sağlayan DÖ algoritmasıdır.[27] BM, simetrik olarak eşleşen ikili birimlerden oluşan ağ yapısına sahiptir. Hesaplamalar yapılırken birden fazla basamaklı ağ yapısı kullanılırsa buna derin inanç ağı adı verilir.[28]

AE

AE, girişi ağı ile çıkış ağı arasındaki süreci yeniden yapılandırarak, boyut indirgeme işlemini yapan kümeleme problemleri için kullanılan bir DÖ metodudur. Temel yapısı gizli katmanlardan oluşur. Girdi katmanı ile gizli katman arasındaki sürece kodlayıcı adı verilir.[29]

DÖ'NÜN BİYOİNFORMATİK ALANDA KULLANIMI

DÖ'nün biyoinformatik alanında kullanımını özetlemek istersek, aşağıdaki tabloda olduğu gibi protein yapılarının tahmin edilmesi, gen ekspresyonlarının regülasyonları gibi birçok alanda kullanılmaktadır (Tablo 2).[22]

TABLO 2: Derin öğrenme algoritmaları ve kullanım alanları.

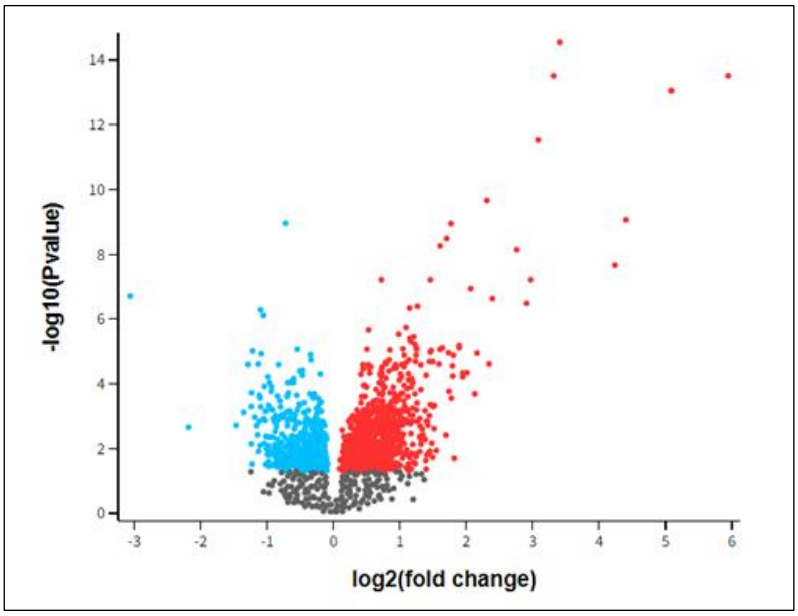
Kullanım alanı algoritma	Omik alanda	Biyomedikal görüntüleme	Biyomedikal sinyal işleme
Derin sinir ağları	Protein yapıları	Anomali sınıflaması	Beyin kod çözme
	Gen ekspresyonları	Segmentasyon	
	Protein sınıflaması	Tanımlama	Anomali sınıflaması
	Anomali sınıflaması	Beyin kod çözme	
Evrişimli sinir ağları	Gen ekspresyonları	Anomali sınıflaması	Beyin kod çözme
		Segmentasyon	
		Tanımlama	Anomali sınıflaması
Yinelenen sinir ağları	Protein yapıları ayırt etme		Beyin kod çözme
	Gen ekspresyonları		Anomali sınıflaması
	Protein sınıflama		

Günümüzde özellikle görüntü işleme süreçlerinde ve genomik alanda sıklıkla DÖ algoritmalarına başvurulmaktadır. En sık kullanılan algoritmaları ve kullanım alanlarını yukarıdaki tablo ile özetlemek mümkündür (Tablo 2).[22]

BULGULAR

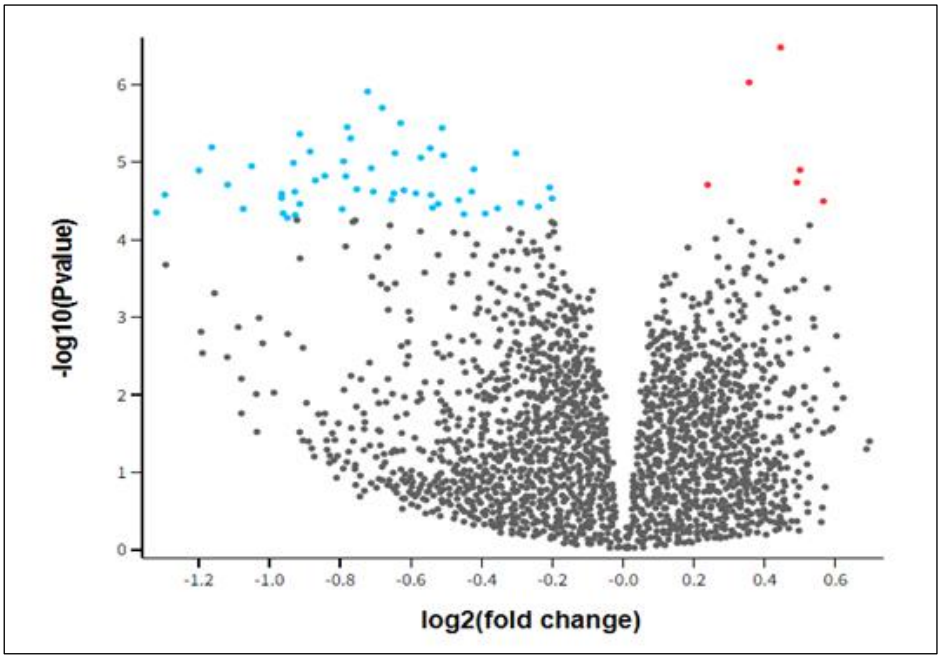
RNA SEKANSLAMA VERİLERİ BULGULARI

Genellikle RNA-sekanslama analizleri sonuçlarını göstermek için Volcano Plot grafikleri kullanılır. Bu grafiklerde yatay eksen (x-ekseni) değişim büyüklüğünü temsil eder, genellikle log2 (fold değişimi) olarak ölçülür. Bu, belirli bir genin ifade seviyesinin iki koşul veya grup arasında ne kadar değiştiğini gösterir. X-ekseninin sağ tarafındaki pozitif değerler yukarı regülasyonu (bir koşulun diğerine göre daha yüksek ifadesini) gösterirken, sol tarafındaki negatif değerler aşağı regülasyonu (bir koşulun diğerine göre daha düşük ifadesini) gösterir. Y-ekseni ise istatistiksel önemi temsil eder, genellikle p değerinin negatif logaritması (taban 10) olarak ölçülür. Bu değer, gözlemlenen fold değişiminin rastgele şansa bağlı olma olasılığını gösterir. Y-ekseninin daha üstünde (grafikteki en üst kısım), daha düşük p değerlerine sahip ve istatistiksel olarak daha anlamlı kabul edilen genler bulunur.[30]



ŞEKİL 4: Volcano Plot GSE85567 gen ekspresyon sonuçları.

GSE85567 erişim numaralı çalışmada; astımlı (74) ve astımlı olmayanların (41), toplam 115 kişinin epitel hücrelerinden genom çapında RNASeq verileri üretilmiştir. Bu çalışma sonucunda, rs2517955 ile rs12936321 polimorfizmlerinin astım ile ilişkili olabilecekleri ortaya konulmuştur ve ORMDL3 genin istatistiksel olarak anlamlı derece astımla ilişkili olabileceği öne sürülmüştür (Şekil 4).[31]



ŞEKİL 5: Volcano Plot GSE129166 gen ekspresyon sonuçları.

GSE129166 erişim numaralı çalışmada; böbrek nakli yapılmış kişilerden alınan periferik kan örnekleri kontrol grubu 46 kişi olarak belirlenmiş ve böbrek allogreft biyopsisi yapılmış kişilerden alınan örnekler ise hasta grubu olarak belirlenmiştir (166) ve toplam 212 kişinin RNASeq verileri üretilmiştir. Bu çalışma sonucunda, antikor aracılı noninvaziv teşhisi kullanabilmek için; *CXCL10, FCGR1A, FCGR1B, GBP1, GBP4, IL15, KLRC1, TIMP1* genlerinin kullanabileceği ortaya atılmıştır (Şekil 5).[32]

MAKİNE ÖĞRENİMİ BULGULARI

GSE85567 astım verisi için, hasta ve kontrol ayrımını en iyi yapabilen modeli belirlemek adına uygulanan 3 farklı model arasında validasyon veri setlerinde en yüksek doğruluk oranına RF modeli 0,89 ile sahipken, test setleri arasında SVM 0,87 başarı değerine ulaşmıştır. DNN modeli ise dengeli sonuçlara sahip olmakla birlikte validasyon veri setinde 0,88 doğruluk değerine sahiptir. RF, SVM ve DNN modelleri arasında, her biri kendi avantajları ve sınırlamalarıyla benzer düzeyde etkili olmuşlardır (Tablo 3).

GSE129166 böbrek transplantasyonu verisi için hasta ve kontrol ayrımını validasyon veri setinde en iyi yapan model 0,75 doğruluk değeri ile DNN modeli olarak görülmüştür. SVM modelinin 0,95'lik özgüllük değeri tüm veri setleri ve modeller arasında en yüksek değere sahiptir. RF modeli için test setinde düşük performans gösterdiği söylenebilirken, validasyon seti için 0,73 doğruluk ve 0,73 kesinlik değerleri ile yeterli başarıyı sağladığı söylenebilir (Tablo 3).

TABLO 3: Makine öğrenimi ve derin öğrenme sınıflama performans sonuçları.

RNA sekanslama verisi	Gen ekspresyon sayısı	Toplam örnek genişliği	Model	Veri seti	Doğruluk	Kesinlik	Duyarlılık	F1 Skor	Özgüllük
GSE85567	60661	115	Random Forest	Validasyon	0,89	0,88	0,92	0,93	0,92
				Test	0,85	0,86	0,88	0,92	0,87
			SVM	Validasyon	0,88	0,87	0,93	0,93	0,89
				Test	0,87	0,85	0,92	0,92	0,91
			DNN	Validasyon	0,88	0,82	0,88	0,85	0,85
				Test	0,86	0,78	0,84	0,81	0,83
GSE129166	54676	212	Random Forest	Validasyon	0,73	0,73	0,84	0,78	0,89
				Test	0,67	0,67	0,87	0,75	0,88
			SVM	Validasyon	0,70	0,73	0,91	0,81	0,95
				Test	0,65	0,67	0,93	0,78	0,92
			DNN	Validasyon	0,75	0,67	0,92	0,78	0,88
				Test	0,73	0,73	0,89	0,80	0,82

SVM: Destek vektör makineleri; DNN: Derin sinir ağları.

Tüm modeller veri setleri bakımından kıyaslanınca, GSE85567 astım verisinin GSE129166 böbrek transplantasyonu verisine göre daha başarılı sonuçlar göstermiş olduğu gözlemlenmektedir. Bu durumun nedeni olarak veri yapılarının karmaşıklığının yanı sıra gen ekspresyon sayılarındaki farklılıklar da dikkate alınmalıdır (Tablo 3).

TARTIŞMA

RNA sekanslama verileri ile makine öğrenimi kullanımı giderek yaygınlaşmaktadır.[33] Daha önce yapılan benzer çalışmalarda, RF ile birlikte SVM modeli kullanılarak, servikal kanser, akciğer kanseri ve Alzheimer hastalıkları hasta-kontrol sınıflandırma performansları incelenmiş ve bu modellerin kullanımını önerilmiştir.[34] Benzer mantıkla yapılan bir diğer simülasyon çalışmasında da hem makine öğrenimi hem de DÖ kullanılarak modellerin performansları kıyaslanmıştır. Renal kanser ve akciğer kanseri verileri ile simülasyon sonuçları kıyaslanmıştır. Elde edilen sonuçlarda hem SVM hem de DNN modellerinin yüksek oradan hasta-kontrol sınıflandırmasında başarı elde ettikleri görülmüştür.[35] Bu çalışmada kullanılan astım ve böbrek transplantasyonu veri setleri için 3 modelde başarılı sonuçlar elde edilmiş olması literatür ile uyumlu olmakla birlikte, farklı türde veri setlerinde makine öğrenimini kullanmak açısından da yenilik ortaya koymuştur. Bu durum, gelecekte benzer çalışmalarda bu modellerin ve yöntemlerin daha yaygın olarak kullanılmasının önünü açabilir.

GSE85567 veri seti için RF modelinde, ağaç sayısı 20 olarak kullanıldığında, validasyon veri seti için doğruluk 0,88, kesinlik 0,88 duyarlılık 0,91, f1 skor değeri 0,90 ve özgüllük değeri 0,91 olarak bulunmuştur. Ağaç sayısı 50 olarak belirlendiğinde ise doğruluk 0,89, kesinlik 0,87, duyarlılık 0,90, f1 skor değeri 0,92 ve özgüllük değeri 0,91 olarak bulunmuştur. Ağaç sayısı 100 ve üzeri durumlar da ise doğruluk 0,89, kesinlik 0,88, duyarlılık 0,92, f1 skor değeri 0,93 ve özgüllük değeri 0,92 olarak bulunmuştur ve en uygun ağaç sayısı 100 olarak belirlenmiştir. GSE129166 veri seti için RF modelinde, ağaç sayısı 20 olarak validasyon veri seti için doğruluk 0,70, kesinlik 0,71 duyarlılık 0,82, f1 skor değeri 0,77 ve özgüllük değeri 0,89 olarak bulunmuştur. Ağaç sayısı 50 olarak belirlendiğinde ise doğruluk 0,72, kesinlik 0,73, duyarlılık 0,83, f1 skor değeri 0,78 ve özgüllük değeri 0,88 olarak bulunmuştur. Ağaç sayısı 100 ve üzeri durumlar da ise doğruluk 0,73, kesinlik 0,73, duyarlılık 0,84, f1 skor değeri 0,78 ve özgüllük değeri 0,89 olarak bulunmuştur ve en uygun ağaç sayısı 100 olarak belirlenmiştir. Her iki veri seti için SVM modelinde, RNA sekanslama verisi yapısı gereği veriler arasındaki ilişkilerin doğrusal olmadığı düşünüldüğü için radyal çekirdek kullanılmıştır. Radyal çekirdek için aylak değişken olan "C" parametresindeki değişikliklerin, makine öğrenimi sonuçlarını etkilemediği görülmüştür (Tablo 4). Tüm algoritmalar için en uygun parametreler; Scikit-learn kütüphanesini kullanarak, GridSearchCV fonksiyonu yardımıyla bulunmuştur.

TABLO 4: En uygun parametre tablosu.

Model	En uygun parametre
RF	n_estimators=100, max_features='sqrt', max_depth=20, min_samples_split=2, min_samples_leaf=1
SVM	C=1, kernel='rbf', gamma='scale', degree=3
DNN	layers=(128, 64, 32), learning_rate=0.001, dropout_rate=0.3, batch_size=64, epochs=20

RF: Rastgele orman; SVM: Destek vektör makineleri; DNN: Derin sinir ağları.

SONUÇ

RNA-sekanslama verileri ile hasta-kontrol ayrımı için her 3 model de amaca uygun şekilde kullanılabilir. Tüm modellerin performans metrikleri birbirlerine yakın ve dengelidir. Çalışmamızın literatüre katkısı ise RNA sekanslama verileri üzerinde makine öğrenimi yöntemlerinin etkinliğini ve farklı türde veri setlerinde bu modellerin başarıyla uygulanabileceğini göstermesidir. Bu tür çalışmalar, tıbbi teşhis ve tedavi alanlarında önemli bir rol oynayabilir ve gelecekte bu alandaki araştırmalara öncülük edebilir. Ayrıca makine öğrenimi optimizasyon parametreleri, RNA-sekanslama verileri ve RF algoritması için 100 ve üzeri ağaç sayısı kullanımı önerilirken, SVM modelinin radyal çekirdek ile kullanımı önerilmektedir.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Bu çalışma tamamen yazarın kendi eseri olup başka hiçbir yazar katkısı alınmamıştır.

KAYNAKLAR

1. Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: From raw data to effective interpretation. Front Genet. 2023;14:997383. [Crossref] [PubMed] [PMC]
2. Bao S, Li K, Yan C, Zhang Z, Qu J, Zhou M. Deep learning-based advances and applications for single-cell RNA-sequencing data analysis. Brief Bioinform. 2022;23(1):bbab473. [Crossref] [PubMed]
3. Sandeep SR, Ahamad S, Saxena D, Srivastava K, Jaiswal S, Bora A. To understand the relationship between machine learning and artificial intelligence in large and diversified business organisations. Materials Today: Proceedings. 2022;56(4):2082-6. [Crossref]
4. Öztornacı RO, Coşgun E, Taşdelen B. Genom-boyu ilişki çalışmalarında, makine öğrenimi ve derin öğrenme yöntemlerinin farklı örnek genişliklerinde performanslarının değerlendirilmesi [Evaluation of machine learning methods and deep learning method performance in different sample size in genome association studies]. Turkiye Klinikleri Journal of Biostatistics. 2020;12(2):204-10. [Crossref]
5. González García C, Núñez Valdéz ER, García Díaz V, Pelayo García-Bustelo BC, Cueva Lovelle JM. A review of artificial intelligence in the internet of things. International Journal of Interactive Multimedia and Artificial Intelligence. 2019;5(4):1. [Crossref]
6. Alpaydin E. Introduction To Machine Learning. 4th ed. Cambridge: MIT Press; 2020.
7. Kiranmai B, Damodaram A. A review on evaluation measures for data mining tasks", International Journal of Engineering and Computer Science. 2014;3(7):7217-20. [Link]
8. Breiman L. Random forests. Machine Learning. 2001;45:5-32. [Crossref]
9. Zhao B, Zhou H, Li X, Han D. Water saturation estimation using support vector machine. Society of Exploration Geophysicists. 2006;1693-7. [Crossref]

10. Korkmaz S. Küçük ilaç moleküllerinin derin sinir ağları kullanılarak sınıflandırılması [Small drug molecule classification using deep neural networks]. Turkiye Klinikleri J Biostat. 2019;11(2):93-101. [Crossref]
11. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep Learning. Vol. 1. Cambridge: MIT Press; 2016.
12. Köse T, Özgür S, Coşgun E, Keskinoğlu A, Keskinoğlu P. Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study. Biomed Res Int. 2020;2020:1895076. [Crossref] [PubMed] [PMC]
13. Seide F, Agarwal A. CNTK: Microsoft's Open-Source Deep-Learning Toolkit. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p.2135. [Crossref]
14. Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks. 2015;61:85-117. [Crossref] [PubMed]
15. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A System for Large-Scale Machine Learning. 12th (USENIX) Symposium on Operating Systems Design and Implementation (OSDI' 16). November 2-6, 2016; Savannah, GA, USA: Usenix; 2016. p.265-83. [Link]
16. Brownlee J. Deep Learning With Python: Develop Deep Learning Models on Theano and Tensorflow Using Keras. 1st ed. Machine Learning Mastery; 2016.
17. Gulli A, Pal S. Deep Learning With Keras. 1st ed. Birmingham: Packt Publishing Ltd; 2017.
18. Vidal R, Bruna J, Giryes R, Soatto S. Mathematics of deep learning. Arxiv. 2017;1712.04741. [Link]
19. Şeker A, Diri B, Balık HH. Derin öğrenme yöntemleri ve uygulamaları hakkında bir inceleme [A review about deep learning methods and applications]. Gazi Mühendislik Bilimleri Dergisi. 2017;3(3):47-64 [Link]
20. Doğan F, Türkoğlu İ. (2018). Derin öğrenme algoritmalarının yaprak sınıflandırma başarımlarının karşılaştırılması [The comparison of leaf classification performance of deep learning algorithms]. Sakarya University Journal of Computer and Information Sciences. 2018;1(1):10-21. [Link]
21. Kurt F. Evrişimli sinir ağlarında hiper parametrelerin etkisinin incelenmesi [Yüksek lisans tezi]. Ankara: Hacettepe Üniversitesi; 2018. Erişim tarihi: 29.02.2024 [Link]
22. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851-69. [PubMed]
23. Küçük D, Arıcı N. Doğal dil işlemede derin öğrenme uygulamaları üzerine bir literatür çalışması [A literature study on deep learning applications in natural language processing]. Uluslararası Yönetim Bilişim Sistemleri ve Bilgisayar Bilimleri Dergisi. 2018;2(2):76-86. [Link]
24. Güreşen E. Dynamic market value forecasting using artificial neural networks [PhD thesis]. İstanbul: İstanbul Technical University; 2008. Erişim tarihi: 29.02.2024 [Link]
25. Van Hulle MM. Self-organizing maps. In: Rozenberg G, Bäck T, Kok JN, eds. Handbook of Natural Computing. 1st ed. Berlin, Heidelberg: Springer; 2012. p.585-622. [Crossref]
26. Özçalıcı M. Özdüzenleyici haritalar yöntemi ile bankacılık sektörü piyasa bölümlendirilmesi [Market segmentation with self-organizing maps in banking indus-try]. BDDK Bankacılık ve Finansal Piyasalar Dergisi. 2017;11(2):9-30. [Link]
27. Pekmezci M. Kısıtlanmış Boltzmann makinesi ile zaman serilerinin tahmini [Yüksek lisans tezi]. İstanbul: Maltepe Üniversitesi; 2012. Erişim tarihi: 29.02.2024 [Link]
28. Aminanto E, Kim K. Deep learning in intrusion detection system: an overview. 2016 International Research Conference on Engineering and Technology (2016 IRCET). Higher Education Forum. 2016. [Link]
29. Binbusayyis A, Vaiyapuri T. Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM. Applied Intelligence. 2021;51(10):7094-108. [Crossref]
30. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform. 2019;20(6):2044-54. [Crossref] [PubMed] [PMC]
31. Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, Naureckas ET, et al. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. JCI Insight. 2016;1(20):e90151. [Crossref] [PubMed] [PMC]
32. Van Loon E, Gazut S, Yazdani S, Lerut E, de Loor H, Coemans M, et al. Development and validation of a peripheral blood mRNA assay for the assessment of antibody-mediated kidney allograft rejection: A multicentre, prospective study. EBioMedicine. 2019;46:463-72. [Crossref] [PubMed] [PMC]
33. Goksuluk D, Zararsiz G, Korkmaz S, Eldem V, Zararsiz GE, Ozcetin E, et al. MLSeq: Machine learning interface for RNA-sequencing data. Comput Methods Programs Biomed. 2019;175:223-31. [Crossref] [PubMed]
34. Zararsız G, Goksuluk D, Korkmaz S, Eldem V, Zararsiz GE, Duru IP, et al. A comprehensive simulation study on classification of RNA-Seq data. PLoS One. 2017;12(8):e0182507. [Crossref] [PubMed] [PMC]
35. Kasikci M, Coşgun E, Karabulut E. Classification performance comparison of deep learning and classical data mining methods on RNA-seq data set. International Journal of Data Mining and Bioinformatics. 2021;26(3-4):188-201. [Crossref]