27 Ocak 2021

ÖMER FARUK YILDIZ-2028142006

Karabük Üniversitesi

Biyomedikal Mühendisliğinde Yapay Zeka ve Uygulamaları

Makine öğrenmesi ile Karaciğer fibrozu tespiti

İçindekiler

[**Özet** 2](#_Toc62677701)

[**1.** **Giriş** 2](#_Toc62677702)

[**1.1.** **Veri Setine Genel Bir Bakış** 2](#_Toc62677703)

[**1.2.** **Veri Hazırlık Aşaması** 3](#_Toc62677704)

[**2.** **Çok Katmanlı Algılayıcılar (MLP) ile Model Oluşturma** 4](#_Toc62677705)

[**3.** **Keras ile Model Oluşturma** 5](#_Toc62677706)

[**4.** **Değerlendirme** 6](#_Toc62677707)

[Kaynakça 7](#_Toc62677708)

# **Özet**

Hastalardan alınan biyopsi örnekleriyle karaciğer fibrozunun tespiti invazif bir yöntem olmasından dolayı hastaları oldukça rahatsız etmektedir. Bu yüzden karaciğer fibrozunun derecesini belirlemek amacıyla bir YSA modeli oluşturulmuştur. Oluşturulan bu model sonucunda %27 oranında bir sonuç elde edilmiştir. Geliştirilen bu YSA modelinin biyomedikal alanında kullanılamayacağı sonucuna varılmıştır.

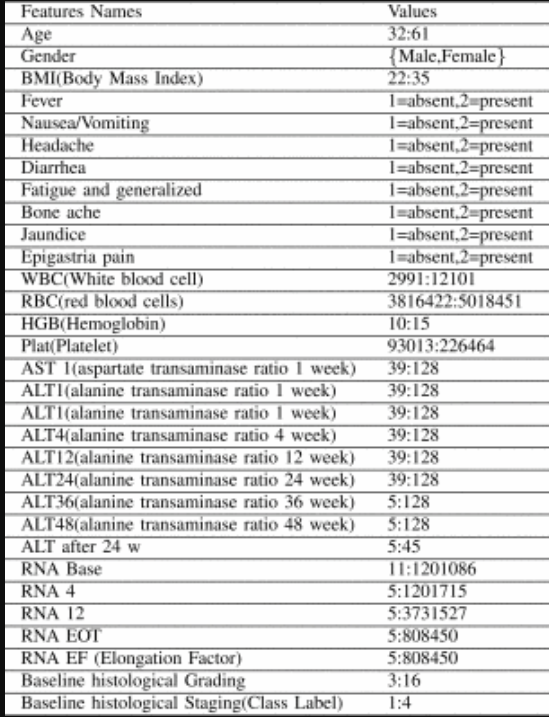
# **Giriş**

Tıp alanında, Karaciğer biyopsisi bazı komplikasyonlarla ilişkili invazif bir prosedürdür. Bu nedenle, karaciğer fibrozunun invazif olmayan teşhisi için farklı biyobelirteçler ve görüntüleme teknikleri geliştirilmiştir. Bununla birlikte, karaciğer fibrozu tanısında bu tür invazif olmayan prosedürlerin karaciğer biyopsisine eşdeğerliği kanıtlanmamıştır. Ayrıca, karaciğer fibrozunun ilerlemesinin riskini ve oranını tahmin edecek hiçbir araç yoktur (Nasr, El-Bahnasy, Hamdy, & Kamal, 2017).  Yayınlanan bu makalede 5 katlı bir crossentropy modeli oluştutulmuş ve veri seti test edilmiştir. Bu veri seti 1741 Mısırlı hastaya ait bilgilerden oluşmaktadır.

Bu çalışmada yayınlanan çalışmaya ait veri seti kullanılarak bir YSA modeli oluşturulmak istenmiştir.

## **Veri Setine Genel Bir Bakış**

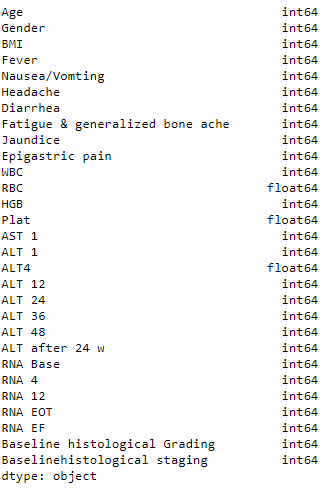
Karaciğer fibrozuna ait veri setinde toplam 1385 veri bulunmaktadır. Bu veriler bu hastalığa ait 29 farklı kategoriye ait bilgileri içermektedir. Veri setinde kullanılan sütunlara ait bilgiler Şekil-1’de gösterilmektedir. Veri setinde tahmin edilmek istenen hedef sütun “Baseline histological Staging” adlı sütundur. "Temel histolojik evreleme", {F0, F1, F2, F3, F4} değerlerine sahip sınıf etiketleridir. Bu etiketler, Karaciğer Fibrozunun farklı prognoz seviyelerini şu şekilde temsil eder: Fibroz Yok (FO), Portal Fibroz (F1), Birkaç Septa (F2), Çok Septa (F3), Siroz (F4). Bu şekilde tahmin edilecek olan veriler 5 farklı kategoriden oluştuğu görülmektedir.



Şekil 1 Veri setine ait sütun bilgileri

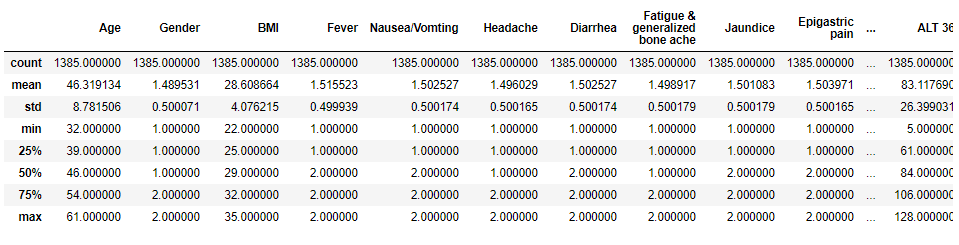
## **Veri Hazırlık Aşaması**

YSA uygulamalarında model oluşturulmadan önce veri seti üzerinde bazı ön işlemlerin yapılması gerekmektedir. Bunlardan birisi obje türünde bir sütun varsa sayısal verilere dönüştürülmeli veya veri setinden kaldırılmalıdır. Kullanılan veri setine bakıldığında bütün verilerin int veya float olarak tutulduğu görülmektedir (Şekil-2).



Şekil 2 Veri setine ait veri tipleri

Bakılması gereken diğer bir önemli özellik ise veriler arasında aşırı verilerin olup olmadığıdır. Veri setine ait sütunlardaki değerlere ait özellikleri bakıldığında uç verilerin olmadığı görülmektedir (Şekil-3).

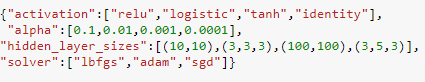


Şekil 3 Veri setine ait bilgiler tablosu

# **Çok Katmanlı Algılayıcılar (MLP) ile Model Oluşturma**

Makine Öğrenmesi yöntemlerinde model oluşturmak için iki seçenek bulunmaktadır. Birinci seçenek gizli katman kullanmadan model oluşturmak, ikinci seçenek ise modele ara katmanlar ekleyerek modeli geliştirmek. Yapılan bu çalışmada sisteme gizli katmanlar eklenerek modelin başarısı daha yükseğe çıkarılmaya çalışılmıştır.

Bu veri setine ait YSA modeli oluşturulmadan önce verilerin normal bir dağılıma sahip olması amacıyla normalizasyon işlemi uygulanmıştır. Daha sonra hangi parametrelerin bu veri setinde daha iyi sonuç elde edeceğini öğrenmek için Grid Search kullanılmıştır. Grid Search yapısını kullanabilmek amacıyla bir sözlük yapısı oluşturulmuş ve parametrelere ait seçenekler sisteme verilmiştir. Bu parametreler Şekil-4’te gösterilmektedir.



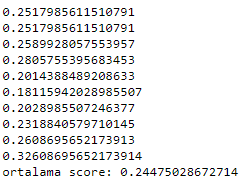
Şekil 4 Parametrelere ait seçenekler

Grid Search yapısının sonucunda elde edilen en iyi parametreler Şekil-5’te gösterilmektedir.



Şekil 5 En iyi sonuç veren parametreler

Grid Search ile sistemin en iyi sonuç verdiği parametreler elde edildikten sonra modelin eğitim sırasında ezber işlemini yapması veya öğrenememesi gibi sorunları ortadan kaldırmak amacıyla K-Fold yapısı kullanılmıştır. Bu yapı ile sistem 10 parçaya bölünerek her parçada sistem eğitilmiştir ve böylece ortalama bir skor elde edilmiştir. Her bir parçaya ait skor ve ortalama skor Şekil-6’da gösterilmektedir.

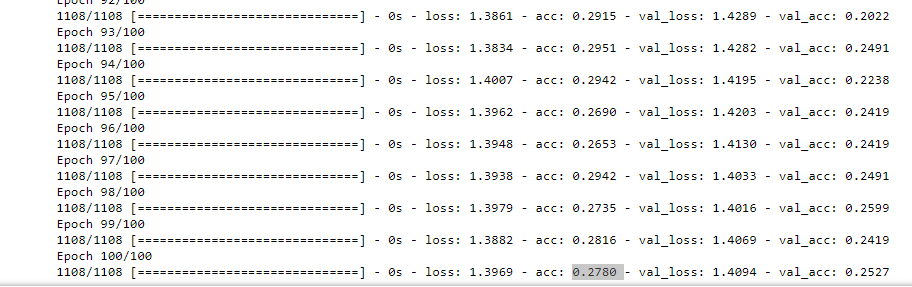


Şekil 6 MLP'ye ait skorlar

# **Keras ile Model Oluşturma**

Çok katmanlı algılayıcılar ile çalışılırken belirlenen gizli katmanlara erişilememekte ve her gizli katmana ait farklı özellikler seçilememektedir. Bu problemi kaldırmak amacıyla model oluşturmak için Keras kütüphanesi kullanılmaktadır. Kullanılan bu veri setiyle Keras aracılığıyla bir model oluşturulmuştur. Bu modelde 5 farklı katman belirlenmiştir. İlk katmandaki nöron sayısı (girdi sayışı + çıktı sayısı) /2 formülü ile belirlenmiştir. Böylece katmanlardaki nöron sayıları sırasıyla 16, 32, 64, 128 ve 256 olarak belirlenmiştir.

Modelin kaç defa eğitileceği (epoch) CPU ile çalışılırken 100 olarak belirlenmiştir. Optimizer olarak “adam”, loss fonksiyonu olarak ise “spars categorical crossentropi” kullanılmıştır. Model eğitilmiş ve 100. eğitimde %27 accuricy oranı elde edilmiştir (Şekil-7).



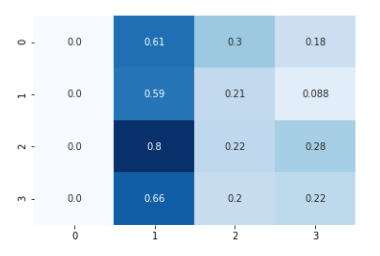
Şekil 7 Keras ile uygulama sonuçlrı

Google üzerinden sunulan COLAB ile yapılan Keras uygulamasında epoch değeri 200 verilerek bu değer %34’e kadar çıkarılmıştır.

# **Değerlendirme**

Karaciğer fibrozunu tespit etmek için oluşturulan bu modelde accuracy oranı %23 olarak elde edilmiştir. GPU desteği ile yapılan Keras uygulamasında ise 200 eğitim sonucu oluşan başarı %34 olarak elde edilmiştir.

Yapılan çalışmaya ait confusion matrix Şekil-8 de gösterilmektedir.



Şekil 8 Confusion Matrix

Bu sonuçlara bakılacak olursa, veri setindeki bilgilerin yeterli olmaması veya veri ön hazırlık evresinde verilerdeki eksikliğin fark edilmemesi TP ve TN değerlerinin düşük çıkmasına neden olduğu görülmektedir. Yapılan çalışmada en yüksek doğru tahmin oranına f1 değeri sahip olmuştur.

Veri setinden elde edilen f1 skoruna bakacak olursak bu değer %19 olarak karşımıza çıkmaktadır. Bu oran, modelin birçok veriyi yanlış tahmin ettiğini göstermektedir.

# Kaynakça

Kodlama, N. (2021, 01 27). *kaggle*. Kaggle: https://www.kaggle.com/ninjacoding/breast-cancer-wisconsin-benign-or-malignant adresinden alındı

Nasr, M., El-Bahnasy, K., Hamdy, M., & Kamal, S. M. (2017). A novel model based on non invasive methods for prediction of liver fibrosis. *Ieee Xplore*.