

GÖĞÜS RÖNTGENİ GÖRÜNTÜLERİNDEN COVID-19 SINIFLANDIRMASI YAPMAK AMACIYLA DERİN ÖĞRENME MODELLERİNDEN MAKİNE ÖĞRENMESİ ALGORİTMALARINA DERİN ÖZNİTELİK AKTARIMI

ÖZET

Koronavirus hastalığı 2019 (COVID-19), SARS-CoV-2'nin neden olduğu bulaşıcı bir hastalıktır. İlk vakalar Aralık 2019'da Çin'in Wuhan kentinden bildirilmiş olup, 11 Mart 2020'de dünya genelini saran bir pandemi olarak ilan edilmiştir. Hastalık şiddetli bir akut solunum yolu hastalığı olmasına rağmen, çeşitli organları etkiler ve ateş, kuru öksürük, yorgunluk başta olmak üzere, tat veya koku kaybı, ishal, baş ağrısı, ağrı ve sızılar, boğaz ağrısı ve konjonktivit gibi çeşitli semptomlara neden olur. Temmuz 2021 başı itibarıyla 185 milyondan fazla insan COVID-19 hastalığına yakalanmış ve 4 milyondan fazla insan hayatını kaybetmiştir. Bu nedenle en önemli konulardan biri COVID-19'un hızlı ve erken tanısıdır. COVID-19'u teşhis etmenin en temel yöntemi Polimeraz Zincir Reaksiyonu (PCR) testi olsa da farklı teknikler denenmekte ve geliştirilmektedir. COVID-19'un akciğerler üzerinde bıraktığı etki çok büyük olduğu için, hastanın akciğerindeki duruma ve akciğer görüntülerine dayalı tanı yöntemleri üzerinde durulmaktadır. Ancak akciğerleri etkileyen çeşitli hastalıklar da vardır. Bu nedenle, COVID-19'u sınıflandırmak için yüksek başarı oranına sahip bir yöntem bulmak önemli bir zorluk ve problem haline gelmiştir.

Bu tezde, göğüs röntgen görüntüleri aracılığıyla COVID-19 hastalığına sahip hastaları sınıflandırmak için, derin öğrenme modellerinden makine öğrenmesi algoritmalarına derin öznitelik aktarımı yöntemini önermekteyiz. Görüntü verilerine ek olarak, derin özniteliklerden gelen bilgileri desteklemek için, makine öğrenimi sürecinde hastaların demografik bilgilerini de kullanıyoruz. Bölüm 1, problemimiz hakkında giriş, bu çalışmanın amacı, ilgili literatür taraması ve bu tezin yapısını içermektedir. Görüntü verilerimiz olan göğüs röntgenleri ile ilgili temel bilgiler Bölüm 2'de verilmiştir.

Odaklandığımız sorun, COVID-19 hastaları ve diğer insanlar arasındaki ikili sınıflandırmadır. Bu sorunu çözmek için 131 COVID-19 ve 123 COVID-19 olmayan etiketli veri içeren veri kümesi kullandık. Daha sonra veri kümesini, toplam verinin %80'i öğrenme kümesinde olacak şekilde, öğrenme ve test kümelerine böldük. Ayrıca öğrenme kümesi verilerine yatay çevirme, dikey çevirme, 90 derece döndürme, 180 derece döndürme ve 270 derece döndürme ile çoğaltma uyguladık. Böylece öğrenme kümesinde 630 COVID-19 ve 588 COVID-19 olmayan etiketli veri, test setinde 26 COVID-19 ve 25 COVID-19 olmayan etiketli veri elde ettik. Belirtilen çoğaltılma işlemi sadece Evrişimsel Sinir Ağı deneyleri sırasında kullanılmıştır. Bölüm 5'in başlarında, veri seti ve çoğaltma işlemi tekniği detaylandırılmıştır.

Kullandığımız derin öğrenme modelleri AlexNet, ResNet-18, ResNet-34, ResNet-50, VGG16 ve VGG19 Evrişimsel Sinir Ağı (ESA) modelleridir. Her ESA modeli için özellikle üç farklı optimizasyon yöntemi uyguladık, ve bunlar SGD momentum, Adam ve ayrıştırılmış ağırlık düşüşüne sahip Adam'dır. Her model için ortak olarak kayıp fonksiyonunun çözebilmek adına çapraz-entropy kayıp fonksiyonunu kullandık. Her

bir görüntü 227 x 227 olarak yeniden boyutlandırıldı, merkezsel olarak kırpıldı, gri tonlamaya dönüştürüldü ve normalleştirildi. Bölüm 3, derin öğrenmeye giriş, ESA'lar hakkında temel bilgiler ve transfer öğreniminin nasıl gerçekleştirileceğinden oluşmaktadır. Bu çalışmada, önceden eğitilmiş modellerin ağırlıklarının mevcut ESA modellerine aktarılması, ve ESA modellerinden çıkarılan derin özniteliklerin makine öğrenimi algoritmalarına aktarılması olarak iki tür aktarımlı öğrenme kullanılmıştır. Ön-eğitilmiş ESA modelleri, resmi olarak ImageNet veri kümesi üzerinde daha önceden eğitilmiş modeller olup, kaydedilen bu ağırlıklarla modeller başlatılmış ve sonrasında yeniden öğrenim gerçekleştirilmiştir. Derin öznitelik aktarımlı öğrenme ise, ESA modelinin özelliklerini tam bağlı model bloğundan çıkartarak makine öğrenmesi algoritmaları gibi başka bir yapay zeka tekniğinde öznitelik matrisi olarak kullanmaktır.

Kullandığımız makine öğrenmesi (MÖ) algoritmaları, Destek Vektör Makineleri (SVM), Lojistik Regresyon (LR), K-En Yakın Komşu (KNN) ve Doğrusal Ayırma Analizi (LDA) olan gözetimli öğrenme algoritmalarıdır. Belirtilen makine öğrenmesi algoritmaları üzerinde L1 normu olarak bilinen Lasso ve L2 normu olarak bilinen Ridge olarak farklı düzenleme tekniklerini uyguladık. 4. Bölüm, makine öğrenmesine giriş, makine öğrenmesi algoritmaları ve düzenleyiciler hakkında temel bilgiler ve çapraz doğrulama tekniğini içermektedir. Başlangıçtaki bölünmüş test kümemize özgü hiper parametrelerin yanı sıra, genelleştirilmiş hiper parametre seçeneklerini elde etmek için öğrenme kümesi üzerinde 10-Katlı çapraz doğrulama gerçekleştirdik. Algoritmalar ve tüm deneyler, demografik bilgilerden oluşan öznitelik kümesine, aktarılmış derin özniteliklerden oluşan öznitelik matrisine ve aktarılan derin öznitelikler ile demografik bilgilerin birleşiminden oluşan öznitelik matrisine ayrı ayrı uygulanmıştır. Demografik bilgi özellik matrisi, bilindiği üzere yaş ve cinsiyet bilgisi olmak üzere iki özellik sütunundan oluşmaktadır. Her örnek için aktarılan derin özniteliklerin uzunluğu bindir. Dolayısıyla, birleşik öznitelik matrisi bin iki sütun içermektedir.

ESA ve MÖ için tüm deneyler, MÖ için veri ön işleme ve hiper parametre seçimi teknikleri dahil olmak üzere, Bölüm 5'te ayrıntılı olarak açıklanmıştır. Her bir öznitelik matrisi ve algoritma için en uygun parametreleri bulmak amacıyla Izgara araması yöntemi kullandık. Deneylerin kaynak kodu ağırlıklı olarak Python programlama dilinde, küçük bir kısmı ise R programlama dilinde yazılmıştır. ESA modelleri Python'da PyTorch kütüphanesi kullanılarak, MÖ algoritmaları ise Python'da Sklearn kütüphanesi kullanılarak uygulanmıştır. Yalnızca düzenlenmiş LDA algoritması, TULIP kütüphanesi kullanılarak R programlama dilinde kodlanmıştır. Daha hızlı ve paralelleştirilmiş işlemler için ESA deneylerimizi GPU üzerinde gerçekleştirdik. Deneylerimiz sırasında kullanabileceğimiz GPU içeren fiziksel bir bilgisayara sahip olma fırsatımız olmadığından, deneyleri Google Colaboratory platformunda gerçekleştirdik. Google Colaboratory, Gmail kullanıcılarına, içermiş olduğu GPU'yu CUDA eklentisi teknolojisi ile kullanma imkanı sağlayan kısmen ücretsiz bir platformdur. ESA sonuçlarını ve en iyi model ağırlıklarını içeren *.pth dosyalarını toplandıktan sonra, MÖ deneyleri lokal olarak CPU üzerinde gerçekleştirilmiştir.

Bölüm 6'da performans ölçüm tekniklerini ESA ve MÖ süreçleri için deney sonuçlarıyla birlikte açıkladık. En iyi sonuca Adam optimizasyon yöntemi ile ResNet-50 modeli kullanılarak ulaşılmıştır. Bu sonuca ilişkin metrikler, doğruluk,

duyarlılık, özgüllük, kesinlik, F1 skoru ve AUC skoru için sırasıyla %92.16, 0.9216, 0.9215, 0.9216, 0.9216, 0.9215'tir. Hem genelleştirilmiş hem de test verilerimize özel optimum hiper parametreleri elde etmeyi amaçladığımızdan, her ikisi için de sonuçlar rapor edilmiştir. Demografik bilgilerin öznitelik matrisi için hem genelleştirilmiş hem de seçilmiş test kümesi hiper parametreleri için en iyi sonuçlar aynıdır, ve KNN algoritması ile elde edilir. Bu sonuca ilişkin metrikler, %56.86 doğruluk, 0.5686 duyarlılık, 0.5837 özgüllük, 0.6863 kesinlik, 0.4955 F1 skoru ve 0.5745 AUC skorudur. ResNet-50 model ağırlıklarından elde edilen derin öznitelik matrisi için, Ridge düzenleyicili SVM, Ridge düzenleyicili LR, LR ve KNN algoritmaları genelleştirilmiş hiper parametrelere göre aynı sonuçları vermiştir. Bu sonuçlara ilişkin metrikler sırasıyla %92.16, 0.9216, 0.9230, 0.9243, 0.9215, 0.9223 olmak üzere doğruluk, duyarlılık, özgüllük, kesinlik, F1 skoru ve AUC skorudur. Son olarak, demografik bilgiler ve ResNet-50 model ağırlıklarından elde edilerek çıkarılan derin özniteliklerin birleştirilmesiyle oluşan öznitelik matrisi için, Ridge düzenleyicili SVM, Ridge düzenleyicili LR, LR ve KNN algoritmaları da genelleştirilmiş hiper parametrelere göre aynı sonuçları vermiştir. Bu sonuçlara ilişkin metrikler sırasıyla %92.16, 0.9216, 0.9230, 0.9243, 0.9215, 0.9223 olmak üzere doğruluk, duyarlılık, özgüllük, kesinlik, F1 skoru ve AUC skorudur.

En nihayetinde, belirtilen sonuçlara bakılacak olursa, doğrusal ayırma analizi ve Lasso düzenleyicisini kullanarak, düzenleyicinin kullanımının bir gelişme sağlayabilmiş olduğunu gördük. Demografik bilgileri derin özniteliklerle birleştirerek bir iyileştirme elde edemedik. Ancak, sayı olarak daha fazla veri ve doktor raporu, tütün ürünü kullanımı, ilişkili genetik hastalık, solunum testi vb. gibi daha fazla hasta bilgileri kullanıldığı takdirde, görüntü ve görüntü olmayan bilgilerin birleştirilmesi tekniğinin gelişme sağlayacağını ön görüyoruz. ESA test sonuçlarından MÖ test sonuçlarına geçişte, doğruluk, duyarlılık ve F1 skoru açısından bir gelişme göremesek de, özgüllük ve kesinlik metriklerinde bir artış yakaladık. Bölüm 7'de söz ettiğimiz gibi, daha fazla numune içeren bir veri kümesi kullanılır ve bu numuneler, bizim kullandığımız röntgenler için uzman radyologlardan yardım alınabilecek olması gibi, konu uzmanları tarafından incelenerek seçilirse, bu durum çalışmanın daha iyi metrik sonuçlarına ve deneysel aşamalar arasında daha iyi karşılaştırma fırsatlarına sahip olmasını sağlayacaktır.

Anahtar Kelimeler: COVID-19, Göğüs Röntgeni, Veri çoğaltma, İkili sınıflandırma, Demografik bilgi, Derin öğrenme, Evrişimsel Sinir Ağları, Ön-eğitilmiş ESA modelleri, Öğrenme aktarımı, Derin öznitelik çıkarma, Derin öznitelik aktarımı, Makine öğrenmesi, Gözetimli öğrenme, Düzenleme, Lasso, Ridge, Izgara araması.