Konya Mühendislik Bilimleri Dergisi, c. 9, s. 2, 522-534, 2021
Konya Journal of Engineering Sciences, v. 9, n. 2, 522-534, 2021
ISSN: 2667-8055 (Elektronik)
DOI: 10.36306/konjes.877805

DERİN SİNİR AĞLARI VE YENİDEN ÖRNEKLEME METOTLARI İLE RUTİN KAN TESTLERİNE DAYALI COVID-19 TESPİTİ

[1]Mahmut TOKMAK, [2]Ecir Uğur KÜÇÜKSİLLE

[1]Isparta Uygulamalı Bilimler Üniversitesi, Gelendost Meslek Yüksekokulu, Finans-Bankacılık Ve Sigortacılık Bölümü, Isparta, TÜRKİYE
[2]Süleyman Demirel Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Isparta, TÜRKİYE
[1]mahmuttokmak@isparta.edu.tr, [2]ecirkucuksille@sdu.edu.tr

(Geliş/Received: 02.03.2019; Kabul/Accepted in Revised Form: 26.06.2020)

ÖZ: İlk olarak Aralık 2019'da ortaya çıkan ve dünya çapında bir salgına neden olan Koronavirüs (COVID-19) hastalığı; akut solunum sendromu SARS-CoV-2'nin neden olduğu viral bir hastalık olarak tanımlanmaktadır. COVID-19 hastalığının tespiti için güncel olan rRT-PCR testi kullanılmaktadır. Bu testin uzun geri dönüş süresi, %15-20 civarında yanlış negatif oranları ve pahalı ekipmanları olması nedeniyle rutin kan incelemelerinin değerleri ile tespit yöntemi daha hızlı ve daha ucuz bir alternatif olarak değerlendirilebilmektedir. Bu çalışmada, rutin kan testlerinden Derin Sinir Ağları (DSA) kullanılarak COVID-19 tespit edilmeye çalışılmıştır. Kullanılan veri setinde sınıf dengesizliği olduğu için yeniden örnekleme yöntemleriyle sınıf dengesizliği giderilmiş ve kullanılan algoritmaların performansları değerlendirilmiştir. Yeniden örnekleme yapılırken SMOTE, ADASYN, Geometric SMOTE, Random UnderSampler, Random OverSampler algoritmaları kullanılmıştır. Kurulan model sonunda 0,985 doğruluk değeri ve 0,99 F1-skoru ile en başarılı sonuç, Random OverSampler algoritması ile alınmıştır. Ayrıca yeni girilecek veriler için tahmin yapabilmek amacıyla, PyQt kullanılarak bir uygulama geliştirilmiştir ve kullanılan niteliklerin modele katkıları SHapley Additive Explanations (SHAP) tekniği ile belirlenmiş ve açıklanmıştır.

Anahtar Kelimeler: *Derin Sinir Ağları, Yeniden Örnekleme, COVID-19*

Covid-19 Detection Based on Routine Blood Tests with Deep Neural Networks and Resampling Methods

ABSTRACT: Coronavirus (COVID-19) disease, which first appeared in December 2019 and caused a worldwide outbreak; is described as a viral disease caused by acute respiratory syndrome SARS-CoV-2. The current RRT-PCR test is used to detect COVID-19 disease. Due to long return time of this test, about 15-20% false-negative rates and expensive equipment, the detection method with the values of routine blood analyses can be considered as a faster and cheaper alternative. In this study, COVID-19 was tried to be detected by using Deep Neural Networks (DNN), one of the routine blood tests. Because there is class imbalance in the used data set, class imbalance has been eliminated by resampling methods and the performance of used algorithms has been evaluated. While resampling, SMOTE, ADASYN, Geometric SMOTE, Random UnderSampler, Random OverSampler algorithms were used. As a result of established model, the most successful result was obtained with the Random OverSampler algorithm, with an accuracy of 0.985 and an F1-score of 0.99. In addition, an application has been developed using PyQt to

make predictions for new data to be entered and the contributions of used attributes to the model were determined and explained with the SHapley Additive Explanations (SHAP) technique.

Key Words: *Deep Neural Networks, Resampling, COVID-19*

GİRİŞ (INTRODUCTION)

COVID-19 hastalığına neden olan SARS-CoV-2, coronaviridae ailesine ait bulaşıcı bir virüstür. Hastalık, öksürük, ateş, yorgunluk ve nefes darlığı gibi semptomlara neden olmaktadır (Cascella ve diğ., 2020; Mohammad ve Tayarani, 2020; "T.C. Sağlık Bakanlığı", 2020). SARS-CoV-2'nin ilk olarak Aralık 2019'da Çin'in Wuhan kentinde görüldüğü bildirilmiştir. O günden günümüze kadar geçen sürede sürekli olarak tüm dünyaya yayılmıştır. Virüsün yayılımı, insan hayatının her alanında büyük zorluklara neden olmuş, başta insan hayatının sona ermesi olmak üzere; ekonomi, eğitim, sanat, kültür gibi birçok alanda yeni sorunlar ortaya çıkmıştır. Hastalığın tanısı, tedavisi ve ortaya çıkan bu sorunları çözmek için her gün yeni teknikler geliştirilmektedir (Mohammad ve Tayarani, 2020; Shilbayeh ve diğ., 2020).

SARS-CoV-2 enfeksiyonlarını tespit etmek için kullanılan standart test; ters polimeraz zincir reaksiyonu (PCR) veya ters transkriptaz-PCR (RT-PCR) tekniği kullanılarak gerçekleştirilen moleküler testtir. Bununla birlikte, testin yürütülmesi zaman alıcıdır, özel ekipman ve reaktiflerin kullanılmasını, örneklerin toplanması için uzman ve eğitimli personelin katılımını gerektirmektedir (Banerjee ve diğ., 2020; Vogels ve diğ., 2020). Ek olarak, kullanılan standart testin akciğer bilgisayarlı tomografisine (BT) kıyasla %80 doğruluğu ortaya konmuştur (AlJame ve diğ., 2020; Banerjee ve diğ., 2020). Hastaları tespit etmenin bir yolu, profesyonel ekipman gerektiren BT veya X-Ray görüntüleridir. Bu görüntüler değerlendirilerek COVID-19 tespit edilebilmektedir (Jacobi ve diğ., 2020; Maghdid ve diğ., 2020; Mohammad ve Tayarani, 2020).

Hastalık teşhisine, tahminine, önlenmesine, tedavisine, yönetimine ve antiviral ilaç keşfine katkıda bulunarak COVID-19 salgının yayılmasını kontrol altına almak ve mevcut klinik prosedürlere yardımcı olmak amacıyla makine öğrenimi (ML) ve yapay zeka (AI) yaklaşımları büyük ilgi toplamakta ve hızla çalışmalar geliştirilmektedir (AlJame ve diğ., 2020; Banerjee ve diğ., 2020; Cabitza ve diğ., 2021). Bu yaklaşımların; X-Ray, BT görüntülerinin analizi ile tespit yöntemleri ve laboratuvar ortamında yapılan kan örneklerinden elde edilen sonuçların analizi ile tespit çalışmaları olmak üzere üç farklı yöntem üzerinde yoğunlaştığı görülmektedir.

Makine öğrenme yöntemleri ile X-Ray ve BT görüntüleri ile COVID-19 tespit mekanizmaları, X-Ray ve BT cihazları ile çekilen akciğer grafisi verilerinin modellenmesi esasına dayanmaktadır. Burada elde edilen veriler ile yapılan analizlerde başarılı sonuçlar elde edilmektedir. Bununla birlikte görüntü almaya yarayan cihazlardaki radyasyon yayma oranları, cihaz sayısındaki yetersizlik ve yüksek maliyetler bu yöntemle tespit mekanizmalarının dezavantajlı yönü olarak belirtilmektedir (AlJame ve diğ., 2020). X-ray ve BT görüntüleri ile tespit mekanizmalarının çoğunlukla Derin Öğrenme (Deep Learning: DL) yöntemlerini kullandıkları görülmektedir. DL yöntemlerinden Tekrarlayan Sinir Ağları (RNN) ve Evrişimsel Sinir Ağları (CNN) sıklıkla çalışılmıştır. RNN yönteminde LSTM mimarisi, CNN yönteminde ise GoogleNet, AlexNet, VGG16, InceptionResNetV2, ResNet34, ResNet50, DenseNet121, DenseNet169, DenseNet201, VGG19, MobilenetV2, NasNetMobile, ve ResNet15V2 mimarilerinin kullanıldığı görülmüştür. Bu çalışmalarda kullanılan modellerin performans değerlendirme kriterlerine bakıldığında ise; kullanılan modellerle elde edilen doğruluk (Accuracy) değerleri %73-%100 arasında değişmekte, F1-skoru (F1-score) %69,13-%100 arasında değişmektedir (Ahsan ve diğ., 2020; Bogu ve Snyder, 2021; Civit-Masot ve diğ., 2020; Hammoudi ve diğ., 2020; Kamal ve diğ., 2021; Loey ve diğ., 2020; Maghdid ve diğ., 2020; Rajaraman ve diğ., 2020; Singh ve diğ., 2020; Shoeibi ve diğ., 2020; Zheng ve diğ., 2020). Bu çalışmaların yanı sıra, COVID-19 için kullanılan DL yöntemlerini ve yapay zeka yöntemlerini araştırıp inceleyen derleme çalışmaları mevcuttur (Shorten ve diğ., 2020; Syeda ve diğ., 2021).

Son zamanlarda yapılan klinik çalışmalar, COVID-19 hastalarında kan parametrelerinin önemli değişiklikler gösterdiğini ve bu parametrelerin COVID-19 tanımlanmasını için ilk taramada, rol oynayabileceğini ortaya koymuştur (AlJame ve diğ., 2020; Fan ve diğ., 2020; Formica ve diğ., 2020; Gao ve

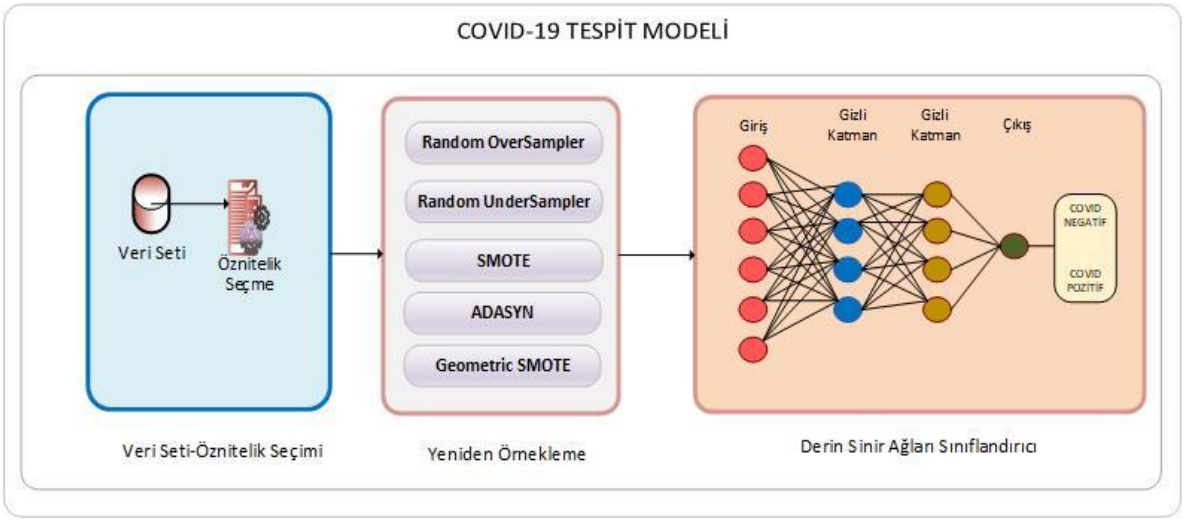
diğ., 2020). Bu bağlamda COVID-19 tespit çalışmalarının yoğunlaştığı üçüncü yöntem olan laboratuvar ortamında elde edilen kan örnekleri ile ilgili yapılan çalışmalar ise; tam kan sayımı ile ilgili verilerin makine öğrenme yöntemleri ile modellenip tespit edilmesi esasına dayanmaktadır.

Bu çalışmalarla ilgili literatür taraması yapıldığında, çoğu araştırmacının, Brezilya'nın Sao Paulo kentindeki Israelita Albert Einstein Hastanesi'nden toplanan laboratuvar test sonuçları veri setini kullandığı görülmüştür (AlJame ve diğ., 2020; "Kaggle, Einstein Data4u", 2020). Bu veri setinde bulunan tam kan sayımı değerleri Navie Bayes (NB), Bayes Ağları (Bayesian Networks: BN), Yapay Sinir Ağları (NN:Neural Network), Rastgele Orman (RF: Random Forest), Ekstra Ağaçlar (Extra Trees: ET), Lojistik Regresyon (Logistic Regression: LR), Destek Vektör Makineleri (Support Vector Machines: SVM), Gradient Boosted Ağaçları (Gradient Boosted Trees: GBT), XGBoost, Gradient Boosting (GB), Kollektif Öğrenme (Ensemble Learning: EL), GLMNET (Lassoelastic-Net Regularized Generalized Linear), Çok Katmanlı Algılayıcı (Multi Layer Perceptron: MLP) gibi ML algoritmaları ile eğitilerek sonuçlar alınmıştır(AlJame ve diğ., 2020; Avila ve diğ., 2020; Banerjee ve diğ., 2020; Czako ve diğ., 2020; de Moraes Batista ve diğ., 2020; Dlotko ve Rudkin, 2020; Mohammad ve Tayarani, 2020; Schwab ve diğ., 2020; Soares, 2020; Yavaş ve diğ., 2020).

Bu çalışmada, literatürde incelediğimiz diğer çalışmalarda yöntem olarak çalışılmamış, DSA yöntemi kullanılarak rutin kan testlerinden COVID-19 tespit edilmeye çalışılmıştır. Aynı zamanda kullanılan veri setindeki sınıf dağılımı dengesizliğini gidermek için literatürdeki çalışmalar, SMOTE algoritması üzerinde odaklanmışlardır (AlJame ve diğ., 2020; Czako ve diğ., 2020; de Freitas Barbosa ve diğ., 2021; Soares, 2020; Yavaş ve diğ., 2020). Bu çalışmada ise yeniden örnekleme algoritmalarının performanslarını ortaya koymak amacıyla SMOTE, ADASYN, Geometric SMOTE, Random UnderSampler, Random OverSampler algoritmaları kullanılmış ve algoritmaların DSA modeli ile performans ölçütleri ortaya konmuştur

MATERYAL VE YÖNTEM (MATERIAL AND METHOD)

Bu bölümde, sınıflandırma için kullanılan modelin ayrıntılı açıklamasını yapılmıştır. İlk olarak, kullanılan veri kümesini ve seçilen özellikleri açıklanmıştır. Veri hazırlama süreci içinde, veri kümesi yeniden örnekleme metotları SMOTE, ADASYN, Geometric SMOTE, Random UnderSampler, Random OverSampler ile dengesiz sınıf özellikleri dengeli hale getirilme işlemi açıklanmıştır. Son olarak, DSA ile sınıflandırma ayrıntıları açıklanmıştır. Çalışmada kullanılan model Şekil 1'de gösterilmiştir.



Şekil 1. COVID-19 tespit modeli
Figure 1. *COVID-19 detection model*

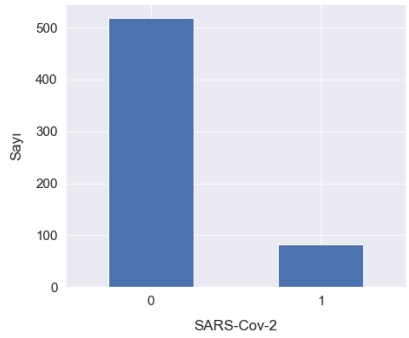
Veri Seti ve Öznitelik Seçimi (Dataset and Feature Selection)

Çalışmada Kaggle platformu tarafından paylaşılan Einstein Data4u anonim veri seti kullanılmıştır ("Kaggle, Einstein Data4u", 2020). Kaggle tarafından paylaşılan veri seti; Birezilya'nın São Paulo kentinde bulunan Albert Einstein Hastanesine başvuran 5644 hastadan elde edilmiştir. Veri seti 28 Mart 2020-3 Nisan 2020 tarihleri arasında toplanan verilerden oluşmakta olup; kan testleri, idrar testleri, SARS-CoV-2 testi, rt-PCR testi, influenza A virüslerinin varlığı dahil olmak üzere 111 laboratuvar test sonucu niteliğine sahiptir. Klinik veriler, ortalama sıfıra ve birim standart sapmaya sahip olacak şekilde paylaşan platform tarafından normalize edilmiştir. Veri setinde hastaneye başvuran 5644 kişiden 558'i SARS-Cov2 pozitif, 5086 tanesi ise SARS-Cov2 negatif olarak tespit edilmiştir.

Bu veri setinden özniteliklerin seçimi; kan sayımı ilgili mevcut klinik çalışmalar (Fan ve diğ., 2020; Formica ve diğ., 2020; Gao ve diğ., 2020) ve literatürde bu veri seti ile yapılmış çalışmalar dikkate alınarak yapılmıştır (AlJame ve diğ., 2020; Avila ve diğ., 2020; Banerjee ve diğ., 2020; Czako ve diğ., 2020; de Moraes Batista ve diğ., 2020; Dlotko ve Rudkin, 2020; Schwab ve diğ., 2020; Yavaş ve diğ., 2020). Aynı zamanda, 111 nitelikten %90'ın üzerinde eksik veri içeren nitelikler elimine edildiğinde 39 nitelik kalmaktadır. Kalan nitelikler içinde bir tanesi hastaya ait id numarası, üç tanesi ise hastanın alındığı servis ile ilgilidir. Seçilen öznitelikler Çizelge 1 de gösterilmiştir.

Tüm veri setinden Çizelge 1'deki öznitelikler seçildikten sonra satırlar düzeyinde %70'in üstünde doluluk ihtiva eden 602 veri seçilmiştir. Eksik olan veriler ortalama yerine koyma metodu ile tamamlanmıştır.

Şekil 2'de gösterildiği gibi SARS-Cov-2 niteliği 519 (%86,3) negatif, 83 (%13,7) pozitif değer içermektedir. Veri seti içinde "positive" ve "negative" şeklinde yer alan değerler; "negative" için 0, "positive" için 1 olacak şekilde etiketlenmiştir. "Negative" nitelik sayısının "positive" nitelik sayısından çok fazla olmasından dolayı hedef nitelikte dengesiz bir dağılım görülmektedir. Bu dengesizliğin giderilmesi için yeniden örnekleme (Resampling) yöntemleri kullanılmıştır.



Şekil 2. SARS-Cov-2 pozitif ve negatif sayıları

***Figure 2**. SARS-Cov-2 positive and negative counts*

Çizelge 1. Veri seti öznitelikleri ve istatistikleri
Table 1. *Dataset feature and statistics*

Öznitelikler	Veri Sayısı	Ortalama	Standart Sapma	En Küçük Değer	En Büyük Değer	Eksik Veri Oranı %
Creatinine	423	0,004102216	0,998792901	-2,389998674	5,053571701	29,73
Proteina C reativa mg/dL	502	0,003367328	1,004205154	-0,535362244	8,02667141	16,61
SARS-Cov-2 exam result	602	0,137873754	0,345054141	0	1	0
Hematocrit	602	-0,001267365	1,001178115	-4,501419544	2,662703753	0
Hemoglobin	602	-0,000899624	1,001418474	-4,345602989	2,671867847	0
Platelets	602	-3,535003563	1,000831594	-2,5524261	9,53203392	0
Mean platelet volume	599	7,43814225	1,000835769	-2,457574606	3,713052034	0,5
Red blood Cells	602	8,424446978	1,000831604	-3,970608234	3,645706177	0
Lymphocytes	602	-7,866736362	1,000831601	-1,865069628	3,764099598	0
Mean corpuscular hemoglobin concentration (MCHC)	602	1,014863142	1,000831599	-5,431808472	3,331070662	0
Leukocytes	602	6,215832763	1,000831604	-2,020302534	4,522041798	0
Basophils	602	-6,633739537	1,000831604	-1,140143752	11,07821941	0
Mean corpuscular hemoglobin (MCH)	602	-3,453009945	1,000831602	-5,937603951	4,098546028	0
Eosinophils	602	7,206147097	1,0008316	-0,835507691	8,350875854	0
Mean corpuscular volume (MCV)	602	-4,155369497	1,000831604	-5,101581097	3,410979986	0
Monocytes	601	-3,220113667	1,000832984	-2,163721323	4,533397198	0,17
Neutrophils	513	5,908361247	1,000976086	-3,339774609	2,535929203	14,78
Urea	396	0,000921818	1,002359765	-1,630410194	11,24656868	34,22
Red blood cell distribution width (RDW)	602	1,020432509	1,000831602	-1,598094344	6,982183933	0
Potassium	370	-0,001845544	1,002074329	-2,283079386	3,401634932	38,54
Sodium	368	0,001171902	1,000792239	-5,246945858	4,096930027	38,87

Yeniden Örnekleme (Resampling)

Dengesiz veri setlerinde sınıf dağılımını değiştirerek sınıf dağılımını dengeli hale getirmek için bazı ön işleme adımları kullanılmaktadır. Sınıf dağılımı dengeli hale getirilirken Yeniden Örnekleme (Resampling) teknikleri kullanılmaktadır. Yeniden örnekleme, veri setini gerçek veri seti kullanılarak yeniden yapılandırma işlemi olarak tanımlanmaktadır. Yeniden örnekleme yöntemleri üç başlık altına toplanabilmektedir (Ankara ve Sahïnturk, 2019; Galar ve diğ., 2011; Haixiang ve diğ., 2017):

Rastgele Veri Çıkarma (Random Undersampling): Sayıca çok olan sınıf örneklerinin rastgele ortadan kaldırılması ile dengesizliği gidermeyi amaçlayan yeniden örnekleme yöntemidir (Ankara ve Sahïnturk,

2019).

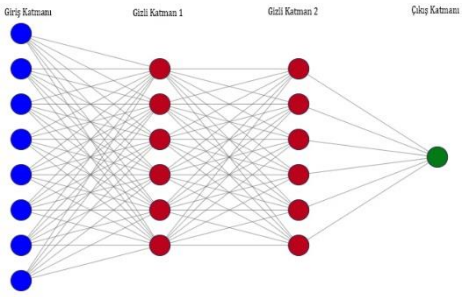
Rastgele Aşırı Örnekleme (Random Oversampling): Azınlık sınıf etiketi örneklerini sayısı fazla olan sınıf etiketine yaklaştırmak için rastgele çoğaltılması ile dengesizliği gidermeyi amaçlayan yeniden örnekleme yöntemidir (Ankara ve Sahinturk, 2019).

Sentetik Veri Üretme (Synthetic samples): Verilerin yetersiz olması halinde orijinal veri seti ele alınarak yapay veri üretilerek dengesizliği gidermeyi amaçlayan yeniden örnekleme yöntemidir. Bu yöntemde kullanılan yaygın algoritmalar SMOTE, ADASYN ve Geometric SMOTE gibi algoritmalardır (Ankara ve Sahinturk, 2019).

Çalışmada sınıf dağılımını dengelemek için; Python imbalanced-learn kütüphanesi kullanılmıştır. Bu kütüphane kullanılarak; Random OverSampler, Random UnderSampler, SMOTE, Geometric SMOTE, ADASYN algoritmaları ile yeniden örnekleme yapılmıştır.

Derin Sinir Ağları (Deep Neural Networks)

DSA, giriş katmanı, çıkış katmanı ve birden çok gizli katmanı olan bir sinir ağı olarak ifade edilmektedir. DSA'nın her katmanında ele alınan problemle ilgili çıkarılan öznitelikler öğrenilmekte ve söz konusu katmanda öğrenilen öznitelikler, kendisinden sonra gelen katman için girdi değerlerini oluşturmaktadır. Bu sayede birinci katmandan başlayarak nihai katmana doğru özniteliklerin öğrenilerek gidildiği bir ağ yapısı oluşturulmuş olmaktadır (Tokmak ve Küçüksille, 2019). Bir giriş katmanı, 2 adet gizli katman ve bir çıkış katmanına sahip bir DSA Şekil 3'te gösterilmiştir.



Şekil 3. Derin sinir ağı yapısı
Figure 3. *Deep neural network structure*

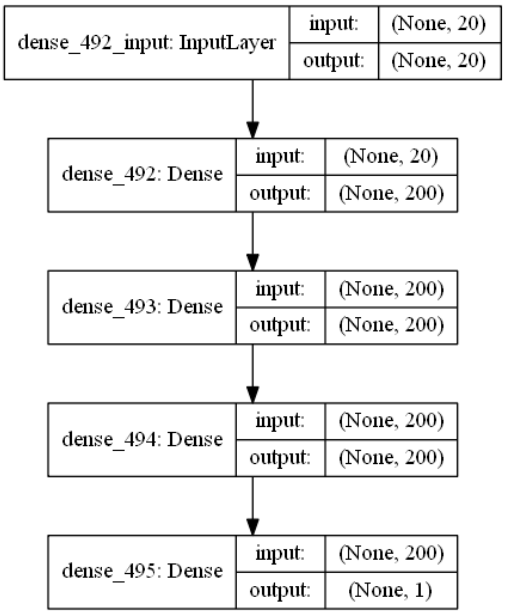
Şekil 3'teki DSA yapısında dairelerle temsil edilen nöronlar, ağırlık değeri, bias ve aktivasyon fonksiyonuna sahiptirler. Nöronlar giriş değeri ve ağırlıkların çarpılmasından sonra, bias değeri eklenerek bir çıkış değeri üretirler bu çıkış değerini kontrol etmek için yani nöronun aktif olup olamayacağına karar vermek için aktivasyon fonksiyonu kullanılmaktadır.

DSA; doğal dil işleme, görüntü işleme, ses tanıma, zaman serisi analizi, zararlı yazılım tespiti gibi birçok alanda kullanılmaktadır ((Barros ve diğ., 2017; Cui ve diğ., 2018; Kolosnjaji ve diğ., 2016; Mezgec ve diğ., 2019; Tokmak ve Küçüksille, 2019; Zeyer ve diğ., 2017)

Bu çalışmada veri seti ön işleme adımlarından sonra elde edilen öznitelikler kurulan bir DSA ile modellenmiştir. DSA modeli Python kullanılarak oluşturulmuş ve sklearn, numpy, keras pandas kütüphaneleri kullanılmıştır. Modelleme sırasında verilerin rastgele %80'i eğitim, %20'si test için ayrılmıştır. Kurulan DSA modeli bir adet giriş katmanı, 3 adet gizli katman ve 1 adet çıkış katmanı içermektedir. Gizli katmanlar 200 adet düğümden oluşmaktadır. Giriş katmanı ve gizli katmanlarda aktivasyon fonksiyonu olarak Rectifier Lienar Units (RELU) kullanılmıştır. Çıkış katmanında kullanılan aktivasyon fonksiyonu ise sigmoid'dir. Modeli eğitmeden önce öğrenme sürecinin yapılandırması işlemi compile fonksiyonu yardımı ile gerçekleştirilmiştir. Bu fonksiyona öğrenme hızını ayarlayan fonksiyonu belirlemek için gerekli olan optimizer parametresi olarak adam, model ikili bir sınıflandırma içerdiği için loss parametresi binary_crossentropy ve modelin performansını değerlendirmek için kullanılacak fonksiyonu belirlemek için verilen metrics parametresine accuracy değerleri verilmiştir. Son olarak model

fit fonksiyonu kullanılarak eğitilmiştir. fit fonksiyonuna modeli eğitmek için kaç defa çalışacağını belirleyen epoch parametresi 30 ve verileri kaçar kaçar alacağını belirleyen batch_size parametresi 10 olarak verilmiştir. Kurulan model Şekil 4'te gösterilmiştir.

Kurulan modelin başarımının değerlendirilebilmesi için kullanılan; doğru sınıflandırılan sınıf örneklerinin oranı olan doğruluk değeri Eşitlik 1'de, pozitif olarak tahmin edilen örneklerin gerçekte ne kadarının pozitif olduğunu ifade eden kesinlik (precision) değeri Eşitlik 2'de, gerçek pozitif değerlerin ne kadarının doğru olduğunu ifade eden duyarlılık (recall) değeri Eşitlik 3'te, kesinlik ve duyarlılık değerlerinin harmonik ortalaması olan F1-skoru (F1 score) Eşitlik 4'te, ROC AUC eğrisi çizilirken kullanılan TPR değeri Eşitlik 5'te ve FPR değeri Eşitlik 6'da gösterilmiştir.



Şekil 4. Kurulan DSA modeli
Figure 4. *Deep Neural Networks structure*

$$Doğruluk = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

$$Kesinlik = \frac{TP}{TP + FP} \quad (2)$$

$$Duyarlılık = \frac{TP}{TP + FN} \quad (3)$$

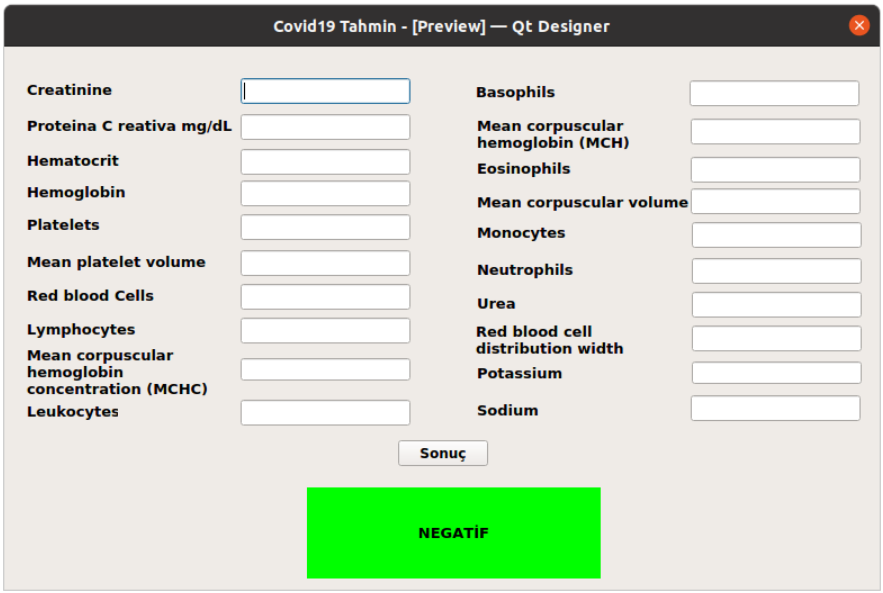
$$F1 - Skor = \frac{2xDuyarlılıkxKesinlik}{Duyarlılık + Kesinlik} \quad (4)$$

$$TPR = \frac{TP}{FN + TP} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} \quad (6)$$

Sağlık alanında ML modelleri ile alınabilecek klinik kararlar, hastaların yaşamları açısından önemlidir. Bu nedenle, bu uygulamalarda, hem doğru hem de yorumlanabilir tahmin modellerine sahip olmak önem arz etmektedir (AlJame ve diğ., 2020). Bu bağlamda modeli yorumlamak amacıyla, tahmin

edilen çıktının belirlenmesinde her bir özelliğin önemini değerlendirmek için SHAP tekniği (Lundberg ve Lee, 2017) kullanılmıştır. SHAP tekniğinde her özelliğin tahmine katkısını açıklayan Shapley değerlerine dayalı olarak model yorumlanabilmektedir (Rodríguez-Pérez ve Bajorath, 2020). Ayrıca kurulan modeli kullanarak yeni girilecek veriler için tahmin yapabilmek amacıyla, Şekil 5'teki uygulama PyQt kullanılarak geliştirilmiştir.



Şekil 5. COVID-19 tahmin uygulaması
Figure 5. *COVID-19 prediction application*

BULGULAR (RESULTS)

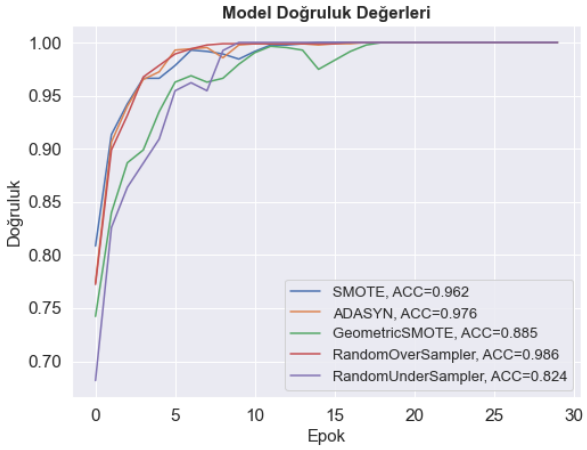
COVID-19 hastalığının neden olduğu dünya çapında, halk sağlığını tehdit eden salgın; farklı araştırma gruplarını ve araştırmacıları COVID-19'un teşhisini olabildiğince otomatikleştirmek amacıyla ML uygulamaları geliştirmeye motive etmiş durumdadır.

Bu çalışmada, COVID-19 hastalığının tespiti amacıyla, veri seti içindeki sınıf dengesizliği yeniden örnekleme metotları kullanılarak giderilmiş ve kurulan DSA modeline girdi olarak kullanılmıştır. Kurulan DSA modeli ile veriler eğitilmiş ve teste tabi tutulmuştur. Çalışmada kurulan tespit sistemi ile hem DSA modelinin performansı hem de yeniden örnekleme metotlarının performansı, model performans ölçütleri ile ortaya konmuştur ve Çizelge 2'de gösterilmiştir ve Çizelge 3' te ise önerilen çalışma ve aynı veri setini kullanan çalışmaların kullandığı modeller ve bu modellerden elde ettiği en yüksek oranlar gösterilmiştir.

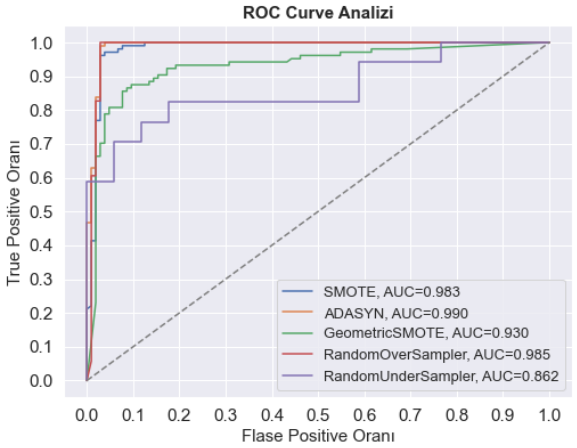
Modelin eğitilip test verisi ile test edilmesinden sonra elde edilen en yüksek doğruluk oranı %98,5 ile Random OverSampler algoritması ile elde edilmiştir. Yine modelin başarımını belirleyen kesinlik oranı %99, duyarlılık oranı %99, F1 skoru %99, ROC-AUC oranı %98,5 olarak elde edilmiştir. Doğruluk oranı en düşük olan algoritma ise; %82,3'lük doğruluk oranı elde edilen Random UnderSampler algoritması olmuştur. Bu algoritmanın diğer başarım kriterleri de çalışmada kullanılan diğer algoritmalara göre daha düşük oranda olup kesinlik oranı %83, duyarlılık oranı %82, F1 skoru %82, ROC-AUC oranı %86,2 olarak elde edilmiştir. Sentetik veri üretme algoritmaları SMOTE ve ADASYN algoritmaları ile %95 oranının üzerinde değerlere erişerek birbirine yakın sonuçlar alınmıştır. Ancak Geometric SMOTE algoritması %88'ler düzeyinde kalarak bu algoritmaların gerisinde kalmıştır. Kullanılan modelin yeniden örnekleme algoritmaları ile elde edilen doğruluk değerleri Şekil 6'da ve ROC-AUC eğrisi Şekil 7'de gösterilmiştir.

Çizelge 2. Model performans ölçütleri
Table 2. *Model performance metrics*

Algoritmalar	Kesinlik (Precision)	Duyarlılık (Recall)	F1-skoru (F1-score)	ROC AUC	Doğruluk (Accuracy)
SMOTE	0,96	0,96	0,96	0,983	0,961
Geometric SMOTE	0,88	0,88	0,88	0,930	0,884
ADASYN	0,98	0,98	0,98	0,990	0,976
Random UnderSampler	0,83	0,82	0,82	0,862	0,823
Random OverSampler	0,99	0,99	0,99	0,985	0,985

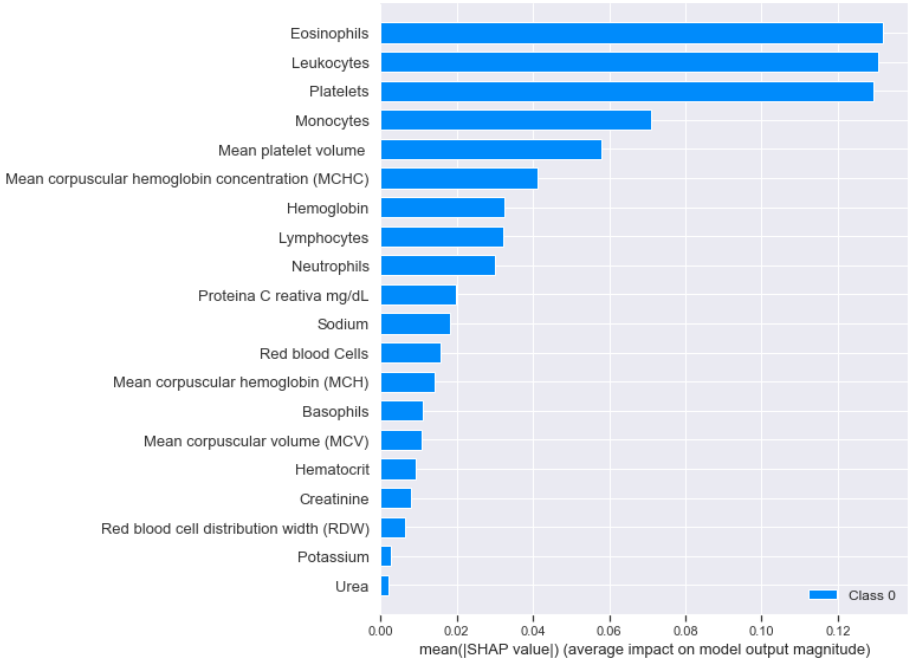


Şekil 6. Doğruluk değerleri
Figure 6. *Accuracy values*



Şekil 7. ROC-AUC eğrisi
Figure 7. *ROC-AUC Curve*

Şekil 8, COVID-19'un pozitif durumları ile ilgili olarak özellik önemini, her bir özelliğin modelin doğruluğuna katkısını göstermektedir. Bu grafikte sadece doğruluğu en yüksek olan Random OverSampler algoritmasına ait modelde elde edilen SHAP değerleri gösterilmiştir. Buna göre "Eosinophils", "Leukocytes", "Plateles" niteliklerinin COVID-19 pozitif vakaların belirlenmesinde tahmin modeline en fazla katkıda bulunduğu görülmektedir. En az katkısı olan niteliklerin ise "Potassium" ve "Urea" olduğu görülmektedir.



Şekil 8. SHAP değerleri
Figure 8. *SHAP values*

Çizelge 3. İlgili çalışmaların yöntemleri ve model performans ölçütleri
Table 3. *Methods of related works and performance metrics*

Çalışmalar	Model	Doğruluk (Accuracy)	Duyarlılık (Recall)	Özgüllük (Specificity)	ROC AUC	F1-Skoru (F1-Score)
Soares, 2020	SVM		70,25%	85,98	86,78%	-
Banerjee vd., 2020	ANN, LR, GLMNET, RF	61-90	43-85	81-91	65-95	-
de Moraes Batista vd., 2020	NN, RF, GBT, LR, SVM		68	85	85	78
de Freitas Barbosa vd., 2021	MLP, SVM, RT, RF, BN, NB	95,159	96,8	93,6		
Avila vd., 2020	NB	76,6				
Yavaş vd., 2020	YSA	90	90			91
Schwab vd., 2020	LR, NN, SVM, RF, GB		75	59	66	
Czako vd., 2020	RF, ET	98.7			98,3	98,2
AlJame vd., 2020	ET, RF, LR, XGBoostT	99,94	99,38	99,99	99,94	
Önerilen Çalışma	**DSA**	**98,5**	**99**		**98,5**	**99**

SONUÇLAR (CONCLUSIONS)

COVID-19 hastalarının erken dönemde tespit edilmesi ve hastaya zamanında müdahale edilmesi salgının yayılmasının önlenmesi için kritik bir öneme sahiptir. Son yapılan çalışmalarda, COVID

hastalarının ilk taraması için rutin kan testlerinin kullanıldığı ortaya konmuş ve kan testlerinin nispeten hızlı, daha ucuz olması ve birçok sağlık kuruluşlarında kolayca yapılabilir olması gerçeğiyle desteklenmiştir. Bu araştırmalarda Navie Bayes (NB), Bayes Ağları (Bayesian Networks: BN), Yapay Sinir Ağları (ANN:Artificial Neural Network), Rastgele Orman (RF: Random Forest), Ekstra Ağaçlar (Extra Trees: ET), Lojistik Regresyon (Logistic Regression: LR), Destek Vektör Makineleri (Support Vector Machines: SVM), Gradient Boosted Ağaçları (Gradient Boosted Trees: GBT), XGBoost, Gradient Boosting (GB), Kollektif Öğrenme (Ensemble Learning: EL), GLMNET (Lassoelastic-Net Regularized Generalized Linear), Çok Katmanlı Algılayıcı (Multi Layer Perceptron: MLP) gibi farklı ML algoritmaları kullanılmıştır.

Bu çalışma ile de DSA modeli kullanılmış, dengesiz sınıf dağılımının giderilmesi noktasında yeniden örnekleme algoritmalarının performansları da değerlendirilmiştir. %82,3 ile %98,5 oranında doğruluk değerlerine erişilmiştir. %98,5 ile Random OverSampler algoritması ile en yüksek doğruluk oranı elde edilmiştir. Modelin performans ölçütlerinden kesinlik oranı %99, duyarlılık oranı %99, F1 skoru %99, ROC-AUC oranı %98,5 olarak elde edilmiştir. Önerilen çalışmada elde edilen ölçütlerin oldukça başarılı olduğu görülmektedir. Her ne kadar AlJame vd., 2020 çalışmasının sonucunda paylaştıkları ölçütlerin %99,9 seviyelerinde olsa da; kullandıkları 19 nitelikten sadece 2 tanesi tam dolu, diğerleri ise %89,32 ile %99,77 oranında eksik veri içermektedir. Örneğin, "Albumin" niteliği için 5644 satırdan sadece 13 tanesi veri içermektedir. Dolayısıyla çalışma evrenini tam olarak temsil edip etmediği tartışılmalıdır.

Kurulan modelin bir sağlıkçı gözüyle yorumlanabilmesi noktasında modelin başarısında etkisi olan niteliklerin önem sıralaması ortaya konmuştur. Ayrıca kurulan modeli kullanarak yeni girilecek veriler için tahmin yapabilmek amacıyla bir uygulama geliştirilmiştir.

Bununla birlikte, COVID-19 teşhisinde ML modellerinin otomatik ve doğru olarak ilerletilebilmesi ve geliştirilmesi için, sağlık uzmanlarının rehberliğinde kan değerlerinin yanı sıra farklı semptomları ve/veya tıbbi görüntüleme cihazlarından alınan veriler ile birleştirildiği yüksek kaliteli veri kümeleri ile çalışmaların yapılması faydalı olacaktır.

KAYNAKLAR (REFERENCES)

Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M., Hossain, M. S., 2020, "Study of different deep learning approach with explainable ai for screening patients with COVID-19 symptoms: Using ct scan and chest x-ray image dataset", *arXiv preprint arXiv:2007.12525*.

AlJame, M., Ahmad, I., Imtiaz, A., Mohammed, A., 2020, "Ensemble learning model for diagnosing COVID-19 from routine blood tests", *Informatics in Medicine Unlocked*, Vol. *21*, pp 100449.

Ankara, N., Sahïnturk, H., 2019, "Dengesiz Kredi Skorlama Veri Setlerinde Kolektif Öğrenme Algoritmalarının Performans Değerlendirmesi", *PressAcademia Procedia*, Vol. *9*, No. 1, pp 180-185.

Avila, E., Dorn, M., Alho, C. S., Kahmann, A., 2020, "Hemogram Data as a Tool for Decision-making in COVID-19 Management: Applications to Resource Scarcity Scenarios", *ArXiv:2005.10227*.

Banerjee, A., Ray, S., Vorselaars, B., Kitson, J., Mamalakis, M., Weeks, S., Baker, M., Mackenzie, L. S., 2020, "Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population", *International Immunopharmacology*, Vol. *86*, pp 106705.

Barros, P., Parisi, G. I., Weber, C., Wermter, S., 2017, "Emotion-modulated attention improves expression recognition: A deep learning model", *Neurocomputing*, Vol. *253*, pp 104-114.

Bogu, G. K., Snyder, M. P., 2021, "Deep learning-based detection of COVID-19 using wearables data", *MedRxiv*, pp 2021.01.08.21249474.

Cabitza, F., Campagner, A., Ferrari, D., Di Resta, C., Ceriotti, D., Sabetta, E., Colombini, A., De Vecchi, E., Banfi, G., Locatelli, M., Carobene, A., 2021, "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests", *Clinical Chemistry and Laboratory Medicine (CCLM)*, Vol. *59*, No. 2, pp 421-431.

Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., Di Napoli, R., 2020, "Features, Evaluation, and Treatment of Coronavirus", *StatPearls*, Treasure Island (FL): StatPearls Publishing.

Chassagnon, G., Vakalopoulou, M., Paragios, N., Revel, M.-P., 2020, "Artificial intelligence applications for thoracic imaging", *European journal of radiology*, Vol. *123*, pp 108774.

Civit-Masot, J., Luna-Perejón, F., Domínguez Morales, M., Civit, A., 2020, "Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images", *Applied Sciences*, Vol. *10*, No. 13, pp 4640.

Cui, Z., Xue, F., Cai, X., Cao, Y., Wang, G., Chen, J., 2018, "Detection of malicious code variants based on deep learning", *IEEE Transactions on Industrial Informatics*, Vol. *14*, No. 7, pp 3187-3196.

Czako Z., Sebestyen G., Hangan A., 2020, "Potenciális COVID-19 fertőzés automatikus felismerésé hagyományos véranalízis alapján", *XXI. Energetika-Elektrotechnika – ENELKO és XXX. Számítástechnika és Oktatás – SzámOkt Multi-konferencia,* pp 57–62.

de Freitas Barbosa, V. A., Gomes, J. C., de Santana, M. A., Albuquerque, J. E. de A., de Souza, R. G., de Souza, R. E., dos Santos, W. P., 2021, "Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests", *Research on Biomedical Engineering*.

de Moraes Batista, A. F., Miraglia, J. L., Rizzi Donato, T. H., Porto Chiavegatto Filho, A. D., 2020, "*COVID-19 diagnosis prediction in emergency care patients: a machine learning approach*" (preprint),Epidemiology. https://doi.org/10.1101/2020.04.04.20052092

Dlotko, P., Rudkin, S., 2020, "*Covid-19 clinical data analysis using Ball Mapper*" (preprint),Intensive Care and Critical Care Medicine. https://doi.org/10.1101/2020.04.10.20061374

Fan, B. E., Chong, V. C. L., Chan, S. S. W., Lim, G. H., Lim, K. G. E., Tan, G. B., Mucheli, S. S., Kuperan, P., Ong, K. H., 2020, "Hematologic parameters in patients with COVID-19 infection", *American journal of hematology*, Vol. *95*, No. 6, pp E131-E134.

Formica, V., Minieri, M., Bernardini, S., Ciotti, M., D'Agostini, C., Roselli, M., Andreoni, M., Morelli, C., Parisi, G., Federici, M., Paganelli, C., Legramante, J. M., 2020, "Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2", *Clinical Medicine*, Vol. *20*, No. 4, pp e114-e119.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches", *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. *42*, No. 4, pp 463-484.

Gao, Y., Li, T., Han, M., Li, X., Wu, D., Xu, Y., Zhu, Y., Liu, Y., Wang, X., Wang, L., 2020, "Diagnostic utility of clinical laboratory data determinations for patients with the severe COVID-19", *Journal of medical virology*, Vol. *92*, No. 7, pp 791-796.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017, "Learning from class-imbalanced data: Review of methods and applications", *Expert Systems with Applications*, Vol. *73*, pp 220-239.

Hammoudi, K., Benhabiles, H., Melkemi, M., Dornaika, F., Arganda-Carreras, I., Collard, D., Scherpereel, A., 2020, "Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19", *arXiv preprint arXiv:2004.03399*.

Jacobi, A., Chung, M., Bernheim, A., Eber, C., 2020, "Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review", *Clinical Imaging*, Vol. *64*, pp 35-42.

Kaggle, Einstein Data4u, 2020. https://www.kaggle.com/einsteindata4u/covid19, Ziyaret Tarihi: 20 Aralık 2020.

Kamal, K. C., Yin, Z., Wu, M., Wu, Z., 2021, "Evaluation of deep learning-based approaches for COVID-19 classification based on chest X-ray images", *Signal, Image and Video Processing*, pp 1-8.

Kolosnjaji, B., Zarras, A., Webster, G., Eckert, C., 2016, "Deep learning for classification of malware system call sequences", *Australasian Joint Conference on Artificial Intelligence*, Cham, ss: 137–149, 2016.

Loey, M., Smarandache, F., Khalifa, N. E. M., 2020, "Within the Lack of Chest COVID-19 X-ray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning", *Symmetry*, Vol. *12*, No. 4, pp 651.

Lundberg, S., Lee, S.-I., 2017, "A unified approach to interpreting model predictions", *arXiv preprint arXiv:1705.07874*.

Maghdid, H. S., Asaad, A. T., Ghafoor, K. Z., Sadiq, A. S., Khan, M. K., 2020, "Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms", *arXiv preprint arXiv:2004.00038*.

Mezgec, S., Eftimov, T., Bucher, T., Seljak, B. K., 2019, "Mixed deep learning and natural language processing method for fake-food image recognition and standardization to help automated dietary assessment", *Public health nutrition*, Vol. *22*, No. 7, pp 1193-1202.

Mohammad, Tayarani, 2020, "Applications of Artificial Intelligence in Battling Against Covid-19: A Literature Review", *Chaos, Solitons & Fractals*, Vol. *142*, pp 110338.

Rajaraman, S., Siegelman, J., Alderson, P. O., Folio, L. S., Folio, L. R., Antani, S. K., 2020, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-rays", *arXiv preprint arXiv:2004.08379*.

Rodríguez-Pérez, R., Bajorath, J., 2020, "Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions", *Journal of computer-aided molecular design*, Vol. *34*, No. 10, pp 1013-1026.

Schwab, P., DuMont Schütte, A., Dietz, B., Bauer, S., 2020, "Clinical Predictive Models for COVID-19: Systematic Study", *Journal of Medical Internet Research*, Vol. *22*, No. 10, pp e21439.

Shilbayeh, S. A., Abonamah, A., Masri, A. A., 2020, "Partially versus Purely Data-Driven Approaches in SARS-CoV-2 Prediction", *Applied Sciences*, Vol. *10*, No. 16, pp 5696.

Shoeibi, A., Khodatars, M., Alizadehsani, R., Ghassemi, N., Jafari, M., Moridian, P., Khadem, A., Sadeghi, D., Hussain, S., Zare, A., Sani, Z. A., Bazeli, J., Khozeimeh, F., Khosravi, A., Nahavandi, S., Acharya, U. R., Shi, P., 2020, "Automated Detection and Forecasting of COVID-19 using Deep Learning Techniques: A Review", *ArXiv:2007.10785*.

Shorten, C., Khoshgoftaar, T. M., Furht, B., 2021, "Deep Learning applications for COVID-19", *Journal of Big Data*, Vol. *8*, No. 1, pp 1-54.

Singh, D., Kumar, V., Yadav, V., Kaur, M., 2020, "Deep Neural Network-Based Screening Model for COVID-19-Infected Patients Using Chest X-Ray Images", *International Journal of Pattern Recognition and Artificial Intelligence*, pp 2151004.

Soares, F., 2020, "A novel specific artificial intelligence-based method to identify COVID-19 cases using simple blood exams", *MedRxiv*.

Syeda, H. B., Syed, M., Sexton, K. W., Syed, S., Begum, S., Syed, F., Prior, F., Yu Jr, F., 2021, "Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review", *JMIR medical informatics*, Vol. *9*, No. 1, pp e23811.

T.C. Sağlık Bakanlığı, 2020. https://covid19.saglik.gov.tr/TR-66300/covid-19-nedir-.html, Ziyaret Tarihi: 31 Aralık 2020.

Tokmak, M., Küçüksille, E. U., 2019, "Kötü Amaçlı Windows Çalıştırılabilir Dosyalarının Derin Öğrenme İle Tespiti", *Bilge International Journal of Science and Technology Research*, Vol. *3*, No. 1, pp 67-76.

Vogels, C. B., Brito, A. F., Wyllie, A. L., Fauver, J. R., Ott, I. M., Kalinich, C. C., Petrone, M. E., Casanovas-Massana, A., Muenker, M. C., Moore, A. J., 2020, "Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT–qPCR primer–probe sets", *Nature microbiology*, Vol. *5*, No. 10, pp 1299-1305.

Yavaş, M., Güran, A., Uysal, M., 2020, "Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması", *European Journal of Science and Technology*, No. Özel Sayı, pp 258-264.

Zeyer, A., Doetsch, P., Voigtlaender, P., Schlüter, R., Ney, H., 2017, "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* ss. 2462-2466,IEEE.

Zheng, C., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., Wang, X., 2020, "Deep learning-based detection for COVID-19 from chest CT using weak label", *medRxiv*.