POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

Derin öğrenme algoritmalarını kullanarak deepfake video tespiti

Deepfake video detection using deep learning algorithms

Yazar(lar) (Author(s)): *Şahin KORKMAZ[1], Mustafa ALKAN[2]*

ORCID[1]: 0000-0002-7197-4799

ORCID[2]: 0000-0002-9542-8039

To cite to this article: Korkmaz Ş., Alkan M., "Derin öğrenme algoritmalarını kullanarak deepfake video tespiti", ***Journal of Polytechnic***, 26(2): 855-862, (2023).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Korkmaz Ş., Alkan M., "Derin öğrenme algoritmalarını kullanarak deepfake video tespiti", ***Politeknik Dergisi***, 26(2): 855-862, (2023).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.1063104**

Derin Öğrenme Algoritmalarını Kullanarak Deepfake Video Tespiti

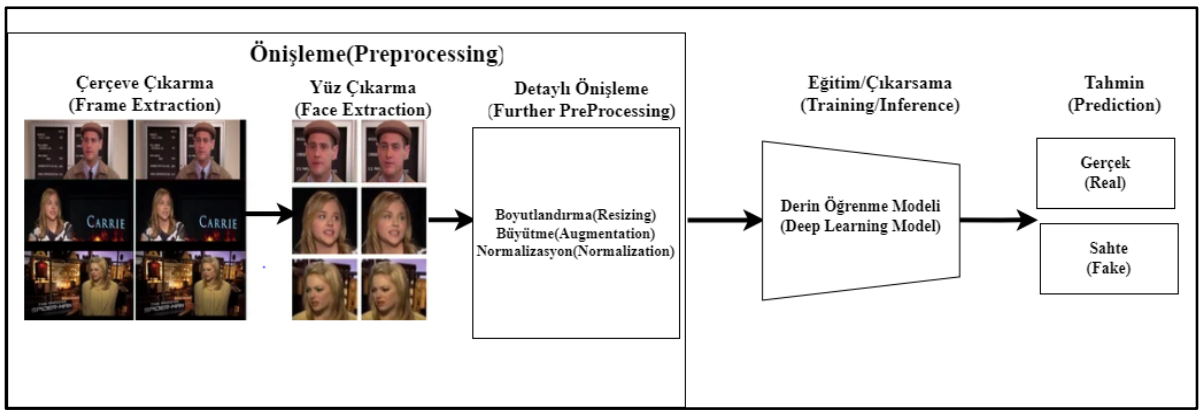
Deepfake Video Detection Using Deep Learning Algorithms

Önemli noktalar (Highlights)

- *DeepFake Kavramı (Deepfake Concept)*
- *Derin Öğrenme Algoritmaları (Deep Learning Algortihms)*
- *Evrişimsel Sinir Ağları (Convolutional Neural Networks)*
- *Çekişmeli Üretici Ağlar (Generative Adversarial Networks)*

Grafik Özet (Graphical Abstract)

Bu çalışmada, bir videonun sahte mi gerçek mi olduğunu tespit etmek için CNN ağlarından EfficientNet model ailesi kullanılarak Deepfake Video Tespit modeli geliştirilmiştir.



Şekil. Deepfake Video Tespit Mimarisi **/Figure.** DeepFake Video Detection Architecture

Amaç (Aim)

Sahte Video Tespitinin Yapılması./ Deepfake Video Detection.

Tasarım ve Yöntem (Design & Methodology)

Sahte Video Tespitinin Yapılması için Derin Öğrenme Algoritmalarından faydalanılarak yeni bir model geliştirilmiştir. / A new model has been developed by using Deep Learning Algorithms for Deepfake Video Detection.

Özgünlük (Originality)

Sahte videoların tespitinde daha önce kullanılmayan EfficientNet model ailesi eğitilmiş ve yeni bir model geliştirilmiştir. / The EfficientNet model family, which was not used before in detecting deepfake videos, was trained and a new model was developed.

Bulgular (Findings)

Sahte video tespitinde mevcut DFDC veri setinde yaklaşık %91'lik bir doğruluk elde edilmiştir. / An accuracy of approximately 91% was recorded using DFDC dataset in deepfake video detection.

Sonuç (Conclusion)

Geliştirilen model, literatür çalışmaları ile kıyaslandığında rekabetçi sonuçlar üretmiş, kullanılan veri setinin büyüklüğü göz önüne alındığında sahte video tespitinde litaratüre önemli bir metodoloji kazandırılmıştır. / The developed model produced competitive results when compared with related works, and an important methodology was brought to the literature in the detection of deepfake video, considering the size of the data set used.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Derin Öğrenme Algoritmalarını Kullanarak Deepfake Video Tespiti

Araştırma Makalesi / Research Article

Şahin KORKMAZ[1]*, Mustafa ALKAN[2]

[1]Bilişim Enstitüsü, Gazi Üniversitesi, Türkiye

[2]Teknoloji Fakültesi, Elektrik - Elektronik Mühendisliği Bölümü, Gazi Üniversitesi, Türkiye

(Geliş/Received : 30.01.2022 ; Kabul/Accepted : 07.03.2022 ; Erken Görünüm/Early View : 07.04.2022)

ÖZ

Deepfake videolar özellikle son yıllarda çok fazla dikkat çeken alanlardan birisidir. Sosyal ağların artan popülaritesi neticesinde mobil cihazların gelişmiş kameraları ile oluşturulan video ve görüntülerin düzenlenmesi ve paylaşılması geçmiş döneme göre daha erişilebilir bir düzeye ulaşmıştır. Deepfake tekniği ile oluşturulan ve sosyal ağlarda dağıtımı yapılan birçok sahte görüntü ve video sadece kişilerin özel hayatını değil, aynı zamanda toplum düzenini de tehdit etmektedir. İnsan yüzü, insanlar arası etkileşimde ve biyometrik tabanlı doğrulama sistemlerinde halihazırda önemli bir role sahiptir. Bu nedenle, yüz karelerinde küçük çaplı manipülasyonlar dahi, güvenlik uygulamalarına ve sayısal verilere olan güveni sarsabilecektir. Bu çalışmada, deepfake video tespiti modelinin oluşturulmasında bir sınıflandırma probleminin çözümü yaklaşımı benimsenmiştir. Öznitelik çıkarıcı olarak önceden eğitilmiş EfficientNet model ailesi kullanılmış ve tahminin çıktısını elde etmek için bunun üzerinde bir sınıflandırıcı eğitilmiştir. Modelin eğitilmesinde yine derin öğrenme tabanlı yöntemlerle üretilen ve en büyük deepfake veri setlerinden olan DFDC veri seti kullanılmıştır. Derin öğrenme algoritma ve kütüphanelerinden yararlanılmış ve belirlenen videonun gerçek mi yoksa sahte mi olduğuna karar veren yeni bir model ortaya koyulmuştur.

Anahtar Kelimeler: Sinir ağları, adli bilişim, derin öğrenme, sahte video.

Deepfake Video Detection Using Deep Learning Algorithms

ABSTRACT

Deepfake videos are one of the areas that have attracted a lot of attention, especially in recent years. As a result of the increasing popularity of social networks, editing and sharing of videos and images created with advanced cameras of mobile devices has reached a more accessible level than before. Many fake images and videos created with the deepfake techniques and distributed on social networks threaten not only the private life of individuals, but also the public order. The human face always has an important role in human interaction and biometric-based verification systems. Therefore, even minor manipulations of face frames can undermine trust in security applications and digital data In this study, a classification problem solution approach is adopted in the creation of the deepfake video detection model. Pre-trained EfficientNet model family is used as feature extractor and a classifier is trained on it to get the output of the prediction. The DFDC dataset, which is one of the largest deepfake datasets and produced by deep learning-based methods, was used to train the model. Deep learning algorithms and libraries have been used and a new model has been introduced that decides whether the determined video is real or fake.

Keywords: Neural networks, forensic science, deep learning, fake video.

1. GİRİŞ (INTRODUCTION)

Derin öğrenme, birkaç yıl öncesine kadar karmaşık veya imkânsız olarak kabul edilen bir dizi teknolojiyi insanoğlunun hizmetine sunmuştur. Derin öğrenme modellerinde ortaya çıkan ilerlemelerin neticesinde, yapay zekâ tabanlı uygulamalar çok çeşitli sorunların çözümünde etkin bir şekilde kullanılmaya başlanmıştır. Bu alanlardan bir tanesi de, yine yapay zekâ algoritmaları kullanılmak suretiyle üretilen ve literatürde "deepfake" olarak adlandırılan yüksek kaliteli sahte videoların tespit edilmesidir. DeepFake ifadesi "Deep" ve "Fake" kavramlarının birleşiminden oluşmaktadır. Buradaki "deep" kavramı "derin" anlamına karşılık gelmekte ve yapay zekâ algoritmalarından derin öğrenme mimarisine vurgu yapmaktadır. Diğer yandan "fake" ise "sahte" anlamını ifade etmektedir. Deepfake, derin öğrenme algoritmaları kullanılmak suretiyle sahte görsel ve işitsel içerik üretme tekniği olarak adlandırılmaktadır.[1] Fotoğraf düzenleme araçları ile daha önceleri de sahte görüntü ve video üretilmesi mümkün olsa da, gelinen noktada, bu sürece derin öğrenme tekniklerinin dâhil edilmesi işlemin zaman maliyetini düşürmüş ve gerçekçilik yüzdesini artırmıştır. Deepfake teknolojisinin her ne kadar yararlı kullanım alanları mevcut olsa da kötü amaçlı kullanımında kişilerin varlığı ve gerçekte mevcut olmayan faaliyetleri hakkında yanılsamalar yaratılabilmektedir. Sahte içeriklerin sosyal medya platformları vasıtasıyla yayılması ve siber zorbalık aracı haline gelmesi toplumda siyasi, sosyal ve mali

*Sorumlu Yazar (Corresponding Author)
e-posta : korkmaz2023@gmail.com

istikrarsızlıklara neden olmakla birlikte demokrasileri de tehdit eder hale gelebilmektedir.

Fotoğraf ve videolardaki yüzleri manipüle etmek her geçen gün daha kritik bir sorun olarak ortada durmaktadır. Bu kapsamda, intikam alınmak maksadıyla bir kişinin yüzünün sentezlenerek pornografik videolarda kullanılması, devlet yöneticilerine söylemedikleri sözlerden dolayı sorumluluk yüklenmesi, üst düzey şirket yöneticilerinin ses ve görüntülerinin manipüle edilerek finans piyasalarının sabote edilmesi gibi örnekler sorunu daha iyi tasvir etmek için verilebilir.[2]

Bu çalışmada, yüzlerin manipüle edilmesi suretiyle oluşturulan sahte videoların tespit edilmesi için derin öğrenme mimarilerinden yararlanılarak yeni bir model önerilmiştir. Modeli kurmadan önce problemin açık bir şekilde ifade edilmesi, istenilen çıktı, veri türü/boyutu, verinin özellik sayısı gibi faktörlerin değerlendirilmesi gerekmektedir. Bu kapsamda eğitim verilerinin boyutu, çıktının doğruluğu ve yorumlanabilirliği, algoritmaların hız ve ihtiyaç duyulan eğitim süreleri de göz önünde bulundurularak bir model tasarlanmıştır. Deepfake video tespiti, sınıflandırıcıların hem orijinal hem de değiştirilmiş videoları algıladığı bir sınıflandırma problemi olarak kabul edilmiştir. Video karelerindeki görsel tutarsızlıkları ve manipülasyonları tespit etmek suretiyle videonun gerçek mi yoksa sahte mi olduğuna karar vermek için EfficientNet ağ modeli kullanılmıştır. EfficientNet evrişimsel sinir ağlarını (CNN-Convolutional Neural Network) kullanan bir algoritmadır. EfficientNet, hem ImageNet hem de yaygın görüntü sınıflandırma aktarımı öğrenme görevlerinde en son teknoloji (State-of-Art-SOTA) doğruluğuna ulaşan verimli modeller arasındadır.

2. LİTERATÜR ARAŞTIRMASI (RELATED WORK)

Son yıllarda video karelerindeki görsel tutarsızlıklarını ve sahtelikleri tespit etmek için çeşitli yöntemler geliştirilmiştir. Deepfake videoları ve görüntüleri tespit etmek için literatürde CNN ve özellik tabanlı yöntemler kullanılmıştır.

Bayar ve Stamm tarafından derin öğrenme tekniklerini kullanarak görüntüde gerçekleştirilen sahtelikleri tespit etmek için CNN mimarisi önerilmiştir. CNN mimarisi, manipüle edilmiş nitelikleri çıkarmak için tasarlanmıştır. Önerilen model, gri tonlamalı görüntü oluşturarak toplanan bir veri setinde (12 farklı modelden) test edilmiştir. Model, çoklu sınıflandırma için % 99,10 doğruluk (dört farklı tür sahtecilik tespit etmiştir) ve ikili sınıflandırma için ise % 99,31 doğruluk ile sonuç vermiştir.[3]

Afchar ve arkadaşları, görüntülerin mezoskopik özelliklerine odaklanmak için daha az sayıda katmanla tasarlanmış MesoNet adında başka bir CNN tabanlı model önermiştir. Videoların sıkıştırılması nedeniyle ve makroskopik düzeyde görüntü gürültüsüne dayalı mikroskobik analiz mümkün olmadığından, mezoskopik unsurlara odaklanmak için bir ara yöntem benimsemiştir. Evrişimsel sinir ağ katmanları kullanan Meso-4 ve Inception modüllerine dayanan MesoInception-4 adlı iki farklı model ortaya koyulmuştur. Bu modellerin değerlendirilmesi internetten toplanan deepfake videolar üzerinde yapılmakta ve Deepfake videoları için ortalama %98 doğrulukla tespit oranı elde edilmektedir. [4]

Zhou ve arkadaşları tarafından sahte yüz tespiti için iki akışlı (yüz sınıflandırması ve yama üçlüsü) ağ önerilmiştir. Yüz görüntülerini gerçek veya sahte olarak sınıflandırmak için bir CNN modeli eğitilmiştir. Önerilen model, FaceSwap ve SwapMe araçları kullanılarak ilk defa toplanan bir veri kümesi üzerinde değerlendirilmiştir. Oluşturmuş oldukları veri setlerinde %92.7'lik bir doğruluk elde edilmiştir.[4]

Li ve arkadaşları, sahte yüz videolarını göz kırpma özelliği temelinde ortaya çıkarmak için sinir ağlarını kullanan bir yöntem önermiştir. İlk adım, videonun her karesindeki yüzlerin algılanmasını içermektedir. Bir sonraki adımda, yüzdeki odak noktalarının yönelimindeki değişiklikler ve kafa hareketlerini azaltmak için tespit edilen yüzler aynı koordinat sistemine hizalanmıştır. [6] Daha sonra, uzun süreli tekrarlayan evrişimsel sinir ağı (Long-term Recurrent Convolutional Networks-LRCN) kullanılarak videonun her karesinde göz kırpma tespiti yapılmıştır.

Guera ve Delp, deepfake videoları otomatik olarak tespit etmek için zamansal farkındalığa sahip bir sistem önermiştir. Video karelerindeki nitelikler önerilen sistem tarafından CNN kullanılarak çıkarılmıştır. Zamansal dizi analizi için ise bir LSTM kullanılmıştır. Buradaki husus, manipülasyonlar kare kare oluşturulduğunda, her karenin diğer karelerle karşılaştırılması neticesinde kareler arasında tutarsızlıklar bulunmakta olup zamansal farkındalık bulunmadığının gösterilmesidir. Tespit edilen tutarsızlıklar, deepfake tespitinde kullanılmış ve veri setlerinde %97'lik bir doğruluk elde edilmiştir.[7]

Yang ve arkadaşları, fotoğraflarda ve videolarda sadece merkezi bölgeleri ve yüzün tüm niteliklerini kullanarak başın duruşunu karşılaştırmıştır. Baş duruşlarının farklılıklarını bir nitelik vektörü olarak sahte ve orijinal görüntüleri veya videoları tespit etmek için ikili sınıflandırıcı destek vektör makinesini (Support Vector Machine-SVM) eğitmek için kullanmışlardır. SVM modelinin, DARPA(Defense Advanced Research Projects Agency) ve UADFV veri kümeleri için sırasıyla %84.3 ve %89'luk bir sonuç elde ettiği ortaya koyulmuştur.[8]

Deepfake video veya görüntü tespiti alanındaki araştırmalarında, genel olarak CNN ve SVM sınıflandırıcıları kullanılmıştır.

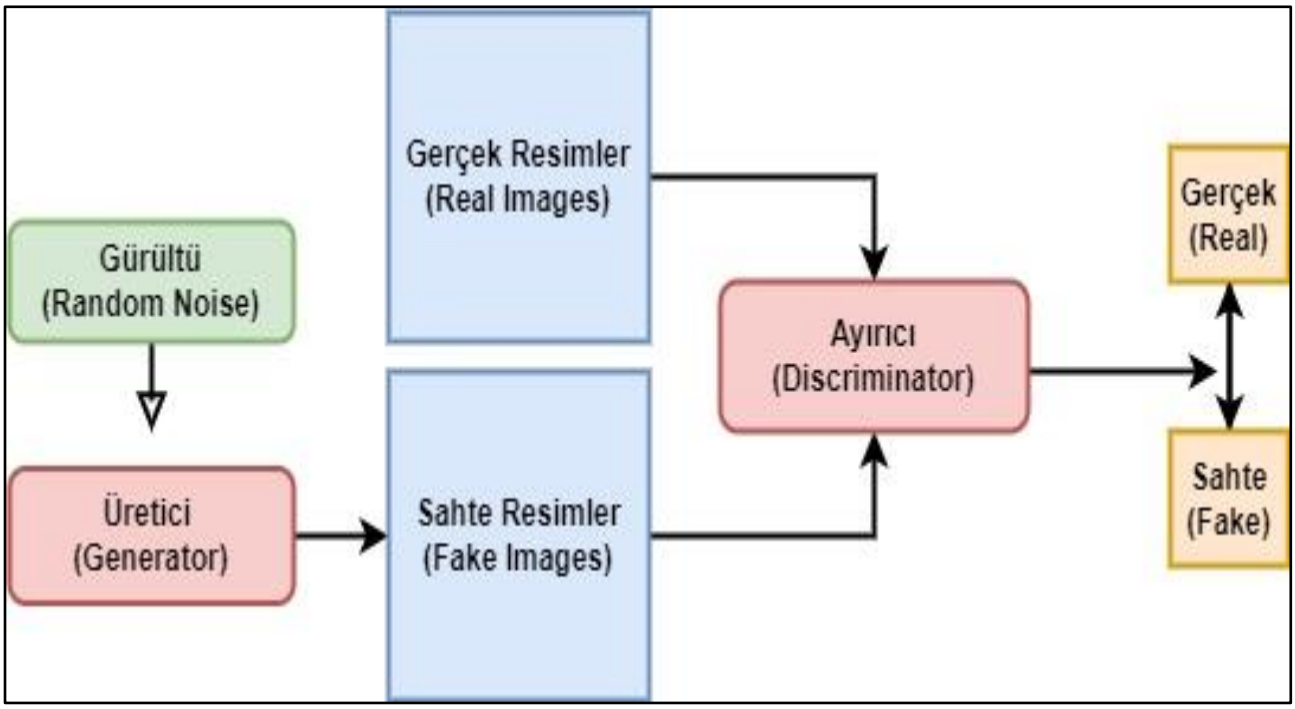
3. DEEPFAKE VİDEO TESPİTİ (DEEPFAKE VIDEO DETECTION)

Deepfake içerik üretmek için birçok yöntem vardır. İnternet ortamı, amatör seviyeden profesyonel seviyeye kadar çeşitli teknik becerilere sahip kullanıcılar tarafından sahte video veya görüntü oluşturmak için birçok

uygulamaya sahiptir. Bu sahte içerikler yine Otomatik Kodlayıcı (Autoencoders) adı verilen bir tür CNN mimarisi tarafından üretilmektedir. Bu yöntemin daha da gelişmişi ise, Şekil-1'de mimarisi gösterilen, sahte görüntü ve videoların kalitesini daha da artıran, denetimsiz bir derin öğrenme algoritması olan Çekişmeli Üretici Ağlar (GAN-Generative Adversarial Network)'dır. Daha önceleri Photoshop gibi yöntemler kullanılarak oluşturulan sahte medyalar, GAN ile sonuçların oldukça gerçekçi olduğu otomatik bir sürece dönüşmüştür.

katmanlardan yararlanır. Özetle, üreticinin amacı, mümkün olduğunca gerçekçi görüntüler üretmek ve ayırıcıyı, oluşturulan görüntülerin gerçek olduğunu düşünmesi için başarılı bir şekilde kandırmaktır.

DeepFake tespiti için önerilen yöntemler arasında Evrişimsel Sinir Ağları en başarılı yöntemler arasında kabul edilmektedir. CNN, bir görüntüden nitelikleri çıkarmada ve bu nitelikleri daha sonra çeşitli uygulamalara sunmakta özel yeteneği olan bir ağ yapısına sahiptir.



Şekil 1. Geleneksel GAN Mimarisi (Architecture of GAN)

Bu içerikler üretilirken, genellikle üst üste bindirme için temel olarak kullanılan yüzdeki sabit alanları değiştiren özel bir teknik kullanılmaktadır. Farklı deepfake videolar oluşturmak için derin öğrenme algoritmaları kullanmış olsalar dahi düzenleme işlemi sırasında geride bazı tutarsızlıklar bırakmaktadır. Sıkıştırma, ışık farklılıkları, dudak ve göz hareketleri ve zamansal farklılıklar gibi faktörler bu tutarsızlıklara örnek olarak verilebilir. [9]

Şekil-1'deki GAN mimarisi incelendiğinde, GAN'da yer alan temel iki bileşen vardır. Bunlar, görüntüleri oluşturan üretici ve oluşturulan görüntünün sahte mi yoksa gerçek mi olduğunu sınıflandıran ayırıcıdır. Üretici, girdi olarak rastgele bir gürültü vektörü olan gizli bir örnek alır. Esasen evrişimli katmanların tersi olan evrişimsiz katmanlardan yararlanarak bir görüntü üretilir. Evrişimli katmanlar, bir girdiden özellikleri çıkarmaktan sorumludur, evrişimsiz katmanlar, özellikleri girdi olarak alırken tersini gerçekleştirir ve çıktı olarak bir görüntü üretir.

Ayırıcının temel görevi, üretilen görüntünün gerçek (1) veya sahte (0) olup olmadığını tahmin etmek olduğu için, görüntü sınıflandırması işlemlerinde evrişimli Ayırt edici uzamsal ve görsel nitelikler, EfficientNet CNN modeli kullanılarak çıkarılmaktadır. Çıkarılan bu nitelikler video karelerindeki görsel kusurları tespit etmeye yardımcı olmakta ve ardından gerçek ve sahte videolar arasında ayrım yapmak üzere sınıflandırıcı algoritma kullanılmaktadır.

4. DENEYSEL ÇALIŞMALAR (EXPERIMENTS)

Bu bölümde yapılan çalışma ile ilgili detaylar paylaşılacaktır. Öncelikle veri seti tanımlanmakta ve ilerleyen süreçte yapılacak araştırmalarda kullanılabilmesi için çalışma kapsamında kullanılan parametrelerin detayları verilecek ve elde edilen sonuçlar analiz edilecektir.

4.1. Veri Seti (Dataset)

Deepfake tespit yöntemlerinin değerlendirilmesi için çok az sayıda veri kümesi mevcuttur. Bu çalışmada, DFDC (Deepfake Detection Challenge) veri seti kullanılacaktır.

DFDC (Deepfake Detection Challenge)

DFDC , manipüle edilmiş medyayı tespit etmek, yeni ve daha iyi modeller oluşturmak maksadıyla bir Kaggle yarışması için yayınlanan eğitim veri setidir. 100.000'den fazla videodan oluşan veri seti Facebook tarafından oluşturulmuştur.[10]

Gerçek videolar, görsel değişkenlik getirmek için rastgele arka planlarla kaydedilen çeşitli eksenlerdeki (cinsiyet, cilt tonu, yaş vb.) çeşitliliği dikkate alarak oluşturulmuştur. Sahte videolar, gerçek olanlardan başlayarak ve farklı deepfake teknikleri, örneğin farklı yüz değiştirme algoritmaları uygulanarak oluşturulur. Video dizi uzunluğu yaklaşık 300 karedir ve sınıflar yaklaşık 100.000 sahte ve 19.000 gerçek olacak şekilde veri seti sahte olana karşı oldukça dengesizdir. Veri setinin boyutu yaklaşık olarak 470 GB, olup her klasör 10GB boyutunda 50 adet mp4 formatında dosyadan oluşmaktadır.[10]

DFDC gibi büyük bir veri setinin seçilmesinin nedeni, algoritmaların yeterince sağlam ve istikrarlı çıktılar üretmesinin amaçlanmasıdır.

Veri setinin dezavantajları değerlendirildiğinde, özellikle sahte videoların oluşturulma sürecinde hangi algoritmaların kullanıldığı ve ses-görüntü ikilisinden hangisinin değiştirildiğinin bilinmemesi sayılabilir.

Eğitim için ilk 35 klasör, doğrulama için 36'dan 40'a kadar olan klasörler ve test için son 10 klasör kullanılarak veri seti klasör yapısına göre ayrılmıştır.

4.2. Ağlar (Networks)

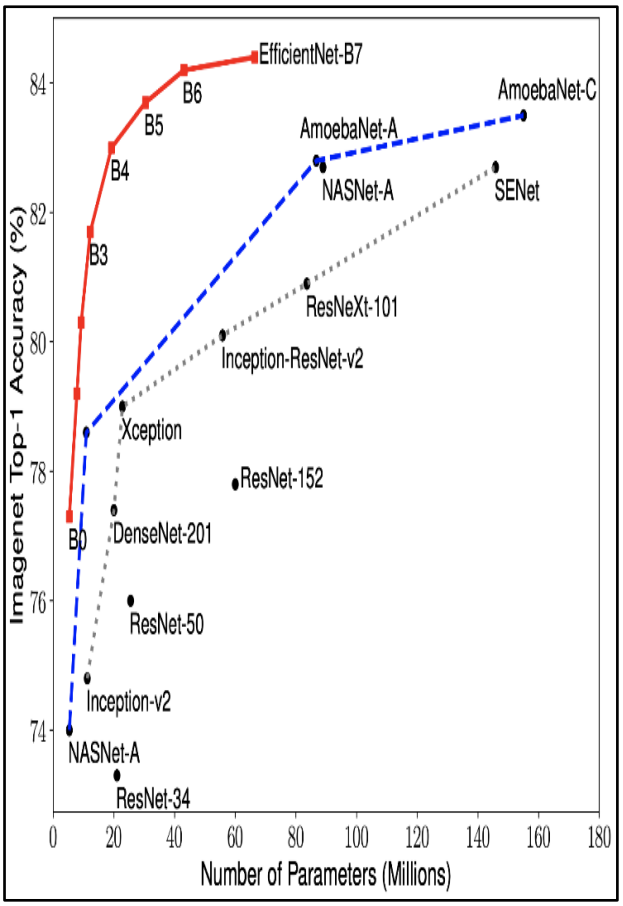
EfficientNet, modeli ölçeklendirmek için buluşsal bir yol sunarak, çeşitli ölçeklerde verimlilik ve doğruluğun iyi bir kombinasyonunu temsil eden bir model ailesi (B0-B7) sağlamaktadır. Temelde, EfficientNet daha az parametreyle daha verimli sonuçlar elde edebilen bir modeldir. Bu tür bir ölçekleme, geniş kapsamlı hiperparametre aramasından kaçınırken verimlilik odaklı temel modelin (B0) her ölçekte modelleri aşmasına olanak tanımaktadır. Ayrıca model B0'dan B7'ye doğru gittikçe başarım oranı ve performansı artmaktadır.[11]

Deneyde EfficentNet model ailesi (B0 – B5) ile çalışılmıştır. Her model, veri seti üzerinde ayrı ayrı eğitilmiş ve test edilmiştir.

EfficientNet model ailesinin ImageNet'teki diğer mevcut CNN'lerle karşılaştırılması Şekil-2'de gösterilmiştir.Genel olarak, ağlar incelendiğinde, EfficientNet modelleri, parametre boyutunu ve FLOPS'u (Floating Point Operations Per Second) bir büyüklük sırasına göre azaltarak, mevcut CNN'lere göre hem daha yüksek doğruluk hem de daha iyi verimlilik elde etmektedir. Yani, EfficientNets, ResNet, Densenet, Inception-v4 ve NASNet dâhil olmak üzere mevcut modellerden çok daha az sayıda parametre ile tutarlı bir şekilde daha iyi doğruluk elde etmektedir.

4.3. Görüntü Sınıflandırma Kurulumu (Image Classification Setup)

Bu çalışmada önerilen deepfake video tespit sistem mimarisi Şekil-3'te ayrıntılı olarak gösterilmiştir. Önerilen modelde video karelerindeki yüzlerin tespiti BlazerFace yüz detektörü ile yapılmaktadır.



Şekil 2. EfficientNet Ağının Diğer Ağlarla Karşılaştırılması (Comparison of the EfficientNet network with other networks)

Çerçeve Çıkarma (Frame Extraction)**:**

Deneyde, her video için yalnızca sınırlı sayıda kare dikkate alınmıştır. Donanım ve zaman kısıtlamaları hesaba katıldığında, hem eğitim hem de test aşamaları için her bir diziden analiz edilen çerçeve sayısı 32 ile sınırlandırılmıştır. Bu sınırlandırma uygulandığında bile, veri setinden 3,8 milyon kare elde edilmiştir. Tüm sahne bilgilerinin deepfake algılama işlemi için yararlı olmadığı bilindiği için analize esas olarak öznenin yüzünün bulunduğu bölgeye odaklanılmıştır.

Yüz Çıkarma (Face Detector)**:**

Sonuç olarak, bir ön işleme adımı olarak deneylerde, MTCNN dedektöründen daha hızlı olduğu kanıtlanan BlazeFace çıkarıcıyı kullanarak sahnede bulunan deneklerinin yüzleri her kareden çıkarılmıştır. Birden fazla yüz tespit edilmesi durumunda en iyi güven puanına sahip yüz çıkartılmıştır.

Detaylı Önişlem (Further Preprocessing)**:**

Bu işlemler yüz dedektörü resimde bir yüz bulamadığı durumlarda, eğitim ve doğrulama sırasında modelleri genellemek ve sağlamlığını artırmak için kullanılmaktadır. Resimde yüz bulunma olasılığı bazı

durumlarda %50'nin altına düşmektedir. Bu sorunu çözmek için, resmin yüzün bulunma olasılığının en düşük olduğu tarafları kırpma ve ardından resmi büyütme işlemi yapılmıştır. Ayrıca girdi olarak alınan yüzlerde yakınlaştırma, döndürme, yatay çevirme, parlaklık ve renk tonu değişiklikleri, gürültü ekleme ve son olarak JPEG sıkıştırması dahil olmak üzere birkaç küçük rasgele dönüşüme tabi tutulmuştur. Özellikle, tüm önişlem adımlarında Pytorch platformu ile birlikte Albumentation Kütüphanesi kullanılmıştır.[12]

Ağlar için elde edilen girdi, 224 × 224 piksel boyutundaki kare renkli görüntü olacak şekilde ayarlanmıştır. Çünkü veri setindeki farklı boyutlardaki görüntülerin hepsinin model için tutarlı bir boyutta olması gerekmektedir. Ayrıca görüntüler Pytorch tarafından işlenecek tensörlere dönüştürülmüş ve tüm görüntüler normalize edilmiştir. Normalizasyon Fonksiyonu vasıtasıyla 0.5'lik bir ortalama ve standart sapma ile tüm görüntüler normalize edilmiştir.

Deepfake Sınıflandırıcı (Deepfake Classifier)**:**

Veri seti üzerinde yapılan tüm adımlardan sonra transfer öğrenme metodu kullanılmak suretiyle EfficientNet modeli kurulmuştur. Transfer öğrenimini gerçekleştirmenin birkaç yolu mevcuttur. Bunlar, önceden eğitilmiş bir model kullanmak veya yeni bir model geliştirmektir.

Önceden eğitilmiş bir model ise iki şekilde kullanılabilmektedir. İlk olarak, oluşturulan model için başlangıç parametreleri olarak önceden eğitilmiş ağırlıklar (weights) ve sapmalar(biases) kullanılabilir ve ardından bu ağırlıklarla modeli eğitebilirsiniz. Diğer yol ise, önceden eğitilmiş modeli kullanarak özellik çıkarımı yapmaktır. Giriş görüntünüzden özellikleri çıkarmak için önceden eğitilmiş modelin parametreleri kullanılır ve bunun üzerine bir sınıflandırıcı eğitilebilir.

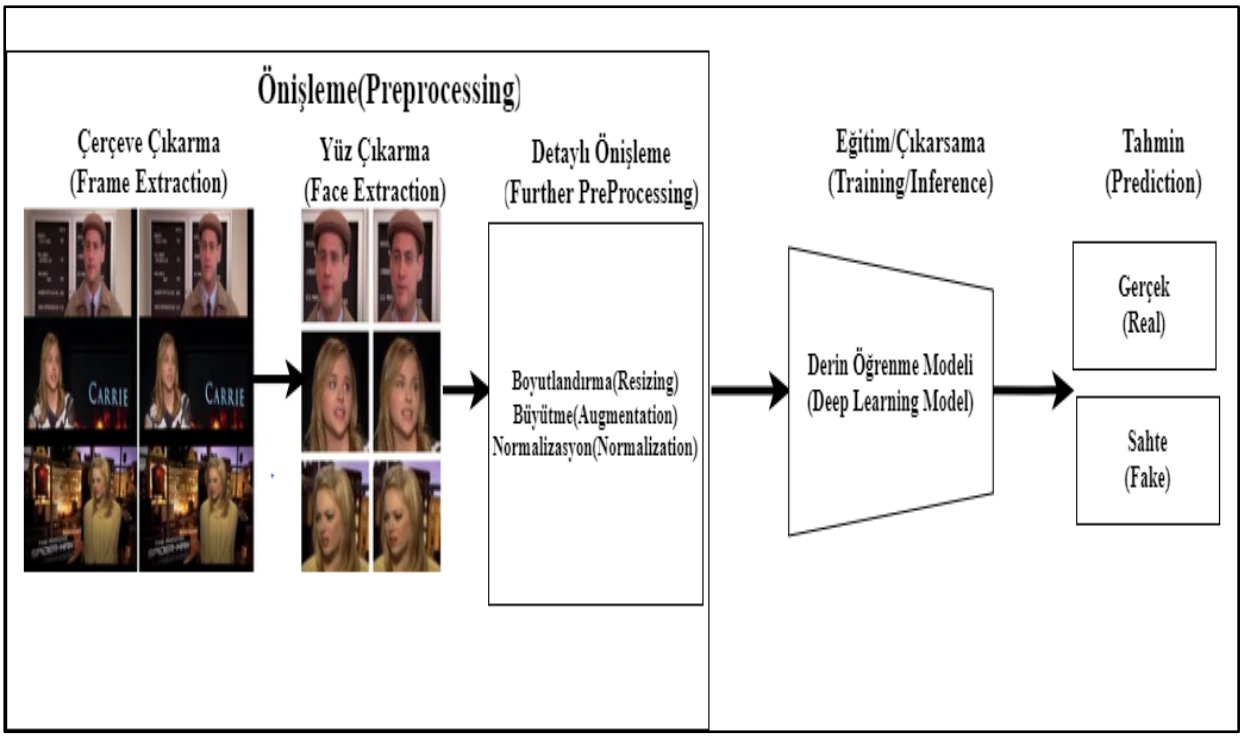
Diğer bir seçenek de, elde mevcut veri miktarı az ise, büyük miktarda veri içeren benzer bir sorun için geliştirilen modelin ağırlıklarının kullanılması suretiyle yeni modelin eğitilmesidir.

Bu çalışmada, öznitelik çıkarıcı olarak önceden eğitilmiş EfficientNet model ailesi kullanılmış ve tahminin çıktısını elde etmek için bunun üzerine bir sınıflandırıcı eğitilmiştir.

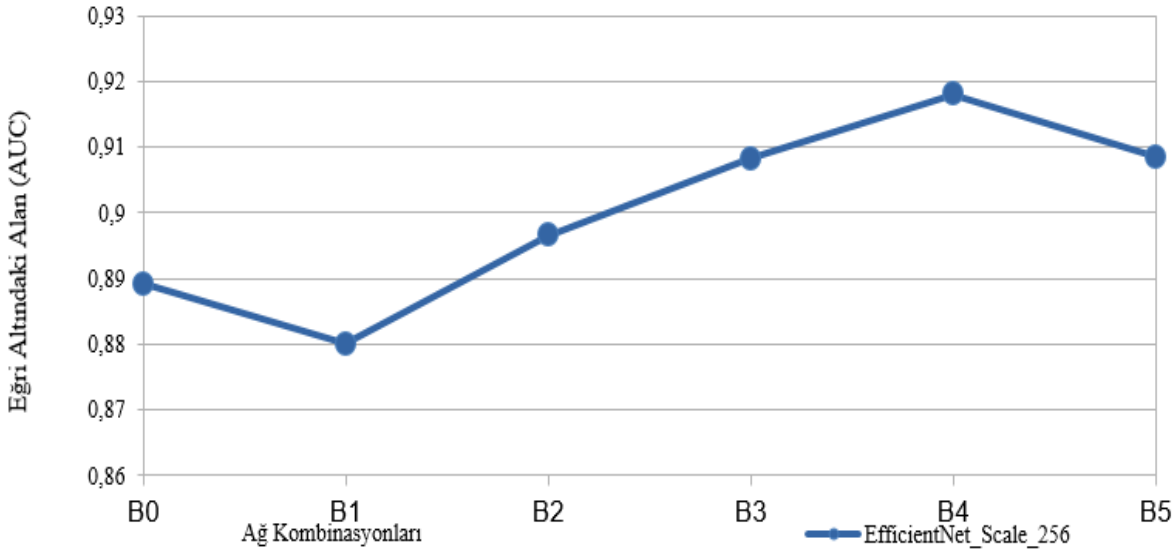
EfficientNet ağ ailesinden B0 ağ modeline göre çalışmayı açıklayabiliriz. Öncelikle EfficientNet-B0 modelini önceden eğitilmiş ağırlıkları ile içe aktarılmıştır. Daha sonra, verilerimizden öznitelikleri çıkarmak için önceden eğitilmiş parametreleri kullanılacağı için modelin parametrelerinin eğitimi devre dışı bırakılmıştır.

EfficientNet mimarisi katmanlarında Batchnorm katmanı, tüm veri grubunu argüman olarak verilen nöron sayısına normalleştirir ve modelin karmaşıklığını azaltır. Dropout katmanı ise modeldeki bazı nöronları argüman olarak verilen değerin olasılığıyla sıfırlar. Doğrusal (Lineer) katman, tamamen bağlı basit bir sinir ağı katmanıdır. Her bloktaki son katmanda lineer aktivasyon fonksiyonu kullanılarak bilgi kaybının önlenmesi sağlanmıştır.

Eğitim aşaması için kullanılan diğer fonksiyonlar Kayıp (Loss Function) ve Optimizasyon (Optimizer) fonksiyonlarıdır. Optimizasyon fonksiyonu için öğrenme



Şekil 3. Deepfake Video Tespit Mimarisi (Arthitecture of Deepfake Video Detection)



Şekil 4. Ağ Kombinasyonlarıyla Elde Edilen Eğri Altındaki Alan(Area Under the Curve for the Different Combinations of Network)

oranının değeri de tanımlanır. Farklı öğrenme oranlarının modeli farklı şekillerde nasıl etkilediğini görmek için bu değer değiştirilebilir. Modeller varsayılan parametrelerle ($\beta1 = 0.9$ ve $\beta2 = 0.999$), Adam Optimizer kullanılarak 256×256×3 boyutunda 32 yüzden oluşan ardışık yığınlarla eğitilmiştir. $1*10^{-3}$'lük ilk öğrenme oranı, 1 x 10^{-10}'a kadar her 250 yinelemede 10'a bölünmüştür. Toplam 30.000 yineleme ile 120 epoch (Eğitim Tur Sayısı) ile eğitim gerçekleşmiştir. Eğitim süresi yaklaşık olarak 50 saat sürmüştür. Her ağ PyTorch 1.10.0 modülü kullanılarak Python 3.7 ile gerçeklenmiştir. PyTorch, derin öğrenme modelleri oluşturmamıza yardımcı olan Python destekli bir kütüphanedir.

Deneyler, Intel Core i9-10900X ve NVIDIA Titan RTX ile donatılmış bir makinede gerçekleştirilmiştir

4.4. Görüntü Sınıflandırma Sonuçları (Image Classification Results) :

DFDC veri setinde eğitilmiş her bir ağın sınıflandırma puanları Çizelge-1'de gösterilmektedir.

Model	**Loss**	**AUC**
EfficientNetB0	0.451217	0.889168
EfficientNetB1	0.447142	0.880076
EfficientNetB2	0.418509	0.896611
EfficientNetB3	0.453147	0.908240
EfficientNetB4	0.451067	**0.918077**
EfficientNetB5	**0.400353**	0.908452

Çizelge 1. Eğitilen Ağın Sınıflandırma Puanları(Classification Scores of the Trained Network)

Her ağ ile her bir çerçeveyi bağımsız olarak ele alarak Şekil-4'te görüldüğü gibi yaklaşık %90 civarında oldukça benzer bir puana ulaşılmıştır. Veri seti çok düşük çözünürlükte çıkarılan bazı yüz görüntülerini içerdiğinden daha yüksek bir puanın mevcut veri setine göre elde edilmesi beklenmemiştir. Mevcut sonuçlar göz ününe alındığında, EfficientNetB5 modeli Loss değeri düşük olsa da, hem parametre sayısı hem de AUC sonucu dikkate alındığında EfficientNetB4 modeli hem AUC değeri olarak hem de parametre sayısı olarak uygun bir seçim olacaktır.

Önerilen modelden elde edilen sonuçlar ile özellikle son yıllarda yapılmış literatürdeki çalışmaların sonuçlarının karşılaştırma çizelgesi, Çizelge-2 ile verilmiştir. Bu karşılaştırmaya kullanılan yöntem, sınıflandırıcı, en iyi performans ve kullanılan veri setleri eklenmiştir. Şunu belirtmekte fayda vardır ki, bahse konu karşılaştırmada AUC değerlendirme ölçütü esas alınmıştır, ancak her modelin farklı veri setleri ile çalışması karşılaştırmayı zorlaştırmaktadır

Ayrıca bahse konu çalışmalara ilişkin detaylar literatür araştırması bölümünde sunulmuştur. Literatürdeki bu yöntemlerin sonuçları incelendiğinde derin öğrenme algoritmalarının kullanılması suretiyle niteliklerin çıkarılması ve bu kapsamda CNN mimarilerinin sınıflandırıcı olarak kullanılmasına yönelik yaklaşımların daha iyi ve istikrarlı sonuçlar ortaya koyduğu aşikârdır.

Bu değerlendirmeler çerçevesinde, deepfake tespit modeli olarak CNN mimarisinin seçilmesinin ne kadar isabetli bir yaklaşım olduğu görülmektedir. Çizelge-1 ve Şekil-4'teki test sonuçları mevcut literatür araştırmalarını hem teyit eder hem de destekler nitelikte bir sonuç ortaya koymuştur. Ancak çalışmalarda kullanılan farklı veri setlerinin bu sonuçların kıyaslanması noktasında tahdit oluşturduğunu da ifade etmek gereklidir.

Bu çalışma, sınıflandırma problemi çerçevesinde çözüm arayışında olması nedeniyle, sınıflandırma işlemi, veri kümesinin küçük ve dengesiz olması durumunda zorlaşmakta ve sınıflama performansı direkt etkilenmektedir. Veri setinin küçük olması, sınıflar arasında dengesizlik olması gibi dezavantajlar nedeniyle kıyaslama aslında bize genel bir bakış açısı sağlamaktadır. Öyle ki, sınıflama algoritmaları, veri setlerinin yeterli büyüklüğe sahip, dengeli olduğu varsayımı üzerine geliştirilmiştir.[14] Önerilen model, diğer literatür çalışmalarına göre daha büyük veri seti üzerinden eğitilmiş olup, sonuçlar daha gerçekçi ve tutarlıdır.

5. SONUÇ (CONCLUSION)

Çalışmada, sahte medyayı tespit etmek için performansı yüksek yeni bir metodoloji tanıtılmıştır. Önerilen yöntem, litetatürde belirtilen çalışmalar ile kıyaslandığında rekabetçi sonuçlar ortaya koymuştur.

Tüm derin öğrenme modellerinin en büyük sorunlarından biri, eğitim verilerine çok fazla uyum sağlamasıdır. Dolayısıyla bahse konu modellerde bilinen örnekler üzerinden bilinmeyen örnekleri değerlendirmek için genelleme yapma eğilimi her zaman mevcuttur. Yani deepfake tespit modellerinin kontrollü senaryolar altında elde ettiği başarım, yüz manipülasyon tekniklerinde meydana gelen gelişmeler ile değerlendirildiğinde, farklı manipülasyon tekniklerle üretilen sahte içeriklere karşı farklı iki veya daha fazla modelden gelen tahminleri birleştiren tahmine dayalı modellerin kullanılması (Ensemble Yöntem) büyük önem arz etmektedir.

Bu kapsamda, deepfake oluşturma sürecindeki gelişmelere ayak uydurmak için başarımı yüksek, performansı etkin derin öğrenme tabanlı ensemble metodu kullanarak deepfake tespit modelleri oluşturmaya devam edilmesi hedeflenmektedir.

Çalışma	Metot	Sınıflandırıcı	AUC	Veri Seti
Yang vd.	Baş Pozisyonu Öznitelikleri (Head Pose Features)	SVM	84.0%	UADF ve DARPA
Zhou vd.	Steganaliz ve Derin Öğrenme Öznitelikleri (Steganalysis and Deep Learning Features	CNN+SVM	92.7	DFDC Celeb-DF
Li vd.	Göz Kırpma Hareketi (Eye Blinking)	LRCN	75.5	DFDC Preview
Guera ve Delp	Görüntü ve Zamansal Öznitelikler (Image and Temporal Features)	CNN+RNN	97.1	Araştırmacılar Tarafından Oluşturulan Veri Seti Kullanılmıştır.
Afchar vd.	Mezoskopik Öznitelikleri (Mesoscopic Features)	CNN	90	FF++
Nguyen vd.	Derin Öğrenme Öznitelikleri (Deep Learning Features)	Capsule Networks	96.6	FF++
Matern vd.	Görsel Öznitelikleri (Visual Features)	Logistic Regression MLP	77.0	DeepFakeTIMIT
Önerilen Model	Derin Öğrenme Öznitelikleri (Deep Learning Features)	CNN	92	DFDC

Çizelge 2.Önerilen Modelin Diğer Modellerle Kıyaslanması(Comparison Between Proposed Model and Other Models)[13]

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Şahin KORKMAZ: Deneyleri yapmış ve sonuçlarını analiz etmiştir. Makalenin yazım işlemini gerçekleştirmiştir

Mustafa ALKAN: Deney sonuçlarını analiz etmiştir. Makalenin yazım sürecine katkı sağlamıştır.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST**)**

Bu çalışmada herhangi bir çıkar çatışması yoktur.

KAYNAKLAR (REFERENCES)

[1] Berk, M ., "Dijital Çağın Yeni Tehlikesi Deepfake" . ***OPUS Uluslararası Toplum Araştırmaları Dergisi*** , 16 (28): 1508-1523, (2020).

[2] Westerlund, M., "The Emergence of Deepfake Technology: A Review." ***Technology Innovation Management Review.*** , 39-52. 10.22215/timreview/1282, (2019).

[3] Bayar, B., Stamm M. C., "A deep learning approach to universal image manipulation detection using a new convolutional layer", ***4th ACM Workshop Inf. Hiding and Multimedia Secure***, 5–10, (2016).

[4] Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., "MesoNet: a Compact Facial Video Forgery Detection Network", ***IEEE International Workshop on Information Forensics and Security (WIFS)***, (2018).

[5] Zhou, P., Han, X., Morariu, V. I., Davis, L. S., "Two-Stream Neural Networks for Tampered Face Detection," [Online]. ***İnternet:http://arxiv.org/abs/1803.11276***, (2018).

[6] Li, Y., Chang, M., Lyu, S., "Exposing AI created fake videos by detecting eye blinking", ***IEEE Int. Workshop on Information Forensics and Security (WIFS)***, 1 –7, (2018).

[7] Güera, D., Delp, E. J., "Deepfake video detection using recurrent neural networks", ***15th IEEE Int. Conf. Adv. Video and Signal Based Surveill. (AVSS)***, 1 –6 Google Scholar, (2018).

[8] Li, Y., Yang, X., Lyu, S., "Exposing deep fakes using inconsistent head poses", ***IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)***, 8261 –8265, (2019).

[9] Li, Y., Chang, M., Lyu, S., "Exposing AI Generated Fake Face Videos by Detecting Eye Blinking", ***arXiv:1806.02877v2 [cs.CV]***, (2018).

[10] "Deepfake Detection Challenge/ Kaggle."**Internet:https://www.kaggle.com/c/deepfake-detection-challenge** (Erişim Tarihi: 20 Ekim 2021).

[11] Tan, M., Le, Q.V., "EfficientNet: Rethinking model scaling for convolutional neural networks",***36th Int. Conf. Mach. Learn. ICML 2019-June***, pp. 10691–10700,(2019).

[12] "Albumentation Library1[Online]. İnternet: ***https://albumentations.ai/docs/examples/pytorch_classification/*** (Erişim Tarihi:17 Ağustos 2021).

[13] Ruben, T., Rodriguez, V., Fierrez, R., Morales, J., Garcia, A. O.,"DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection", ***Information Fusion***, (2020).

[14] Par, Ö., Akçapınar, S., Sever, H., "Sınıflandırmada Küçük ve Dengesiz Veri Kümesi Problemi/Small and Unbalanced Data Set Problem in Classification", ***IEEE 27th Signal Processing and Communications Applications Conference (SIU)***, (2019).