**NASNetLarge CNN Modeli ile Derin Sahte Algılama**

**DeepFake Detection with NASNetLarge CNN Model**

**ÖZET**

Derin sahte görüntüler, son yıllarda bilgisayarlı görü ve derin öğrenme teknolojilerinin gelişimi ile ortaya çıkmış, yeni bir teknolojidir. Bu derin sahte teknolojilerinin gelişmesi pek çok soruna sebep olabileceğinden, derin sahtenin algılanması problemi ortaya çıkmış ve bu problemi çözmek için de birçok yöntem geliştirilmiştir. Bu çalışmada NASNetLarge CNN derin öğrenme modeli ile veri kümesi üzerinde özellik çıkarımı ve sınıflandırma yöntemi önerilmekte ve başarılı bir sonuç üretilmektedir. Proje aşamalarında ilk olarak Celeb-DFv2 veri kümesinde bulunan video karelerinden yüz bölgeleri çıkarılmış ve eğitime verilecek veri kümesi oluşturulmuştur. Ardından önerilen modelin eğitimi yapılarak modelin başarısı test edilmiştir. Çalışma sonuç olarak diğer yöntemlere göre daha küçük bir veri kümesi ile daha kolay bir model eğitimi sunmakta ve rekabetçi, başarılı bir sonuç üretmektedir.

**ABSTRACT**

The image of deep imagination is a new technology that has emerged with the state of the latest computerized computerized and deep learning technologies. Because of the simplicity of the methods of these deep forgery technologies to many problems, the problem of detection of deep forgery arises and can also be applied to many to solve this problem. In this study, to propose and successfully design the dataset with the NASNetLarge CNN deep learning model. Project first2 datasets extracted from video frames consisting of Celeb-DF clusters and datasets to be given to training. It has been tested on the appropriate model of this model. As a result, it will be easy and convenient to build a model with a smaller dataset than usefulness.

1. **GİRİŞ**

Günümüzde görüntü işleme ve derin öğrenme teknolojileri büyük gelişme göstermiştir. Bu iki teknolojinin beraber kullanımı ile bilgisayarlı görü ile derin öğrenme teknikleri kullanılarak birçok yeni gelişme sağlanmıştır. Bunların başında da sentetik görüntü üretimi gelmektedir[1]. Sentetik görüntüden kasıt derin sahte yani deepfake görüntülerdir. Derin sahte görüntüler de yine derin öğrenme yöntemleri ve özellikle GAN mimarisi kullanılarak üretilmektedir[2]. Üretilen bu görüntüler özellikle intikam pornoları ve siyasiler üzerinden yalan söylem oluşturmak üzere kullanılmakta, bu durum devletler, kurumlar ve bireyler için olumsuz sonuçlar doğurabilmektedir. Bunu engellemek için de yine derin öğrenme teknikleri kullanılarak derin sahte algılama algoritmaları gerçekleştirilmektedir. Her ne kadar derin sahte içerik üretme, derin sahte içeriği algılamadan daha önde ilerlesede bu konuda büyük başarılar elde edilmiştir.

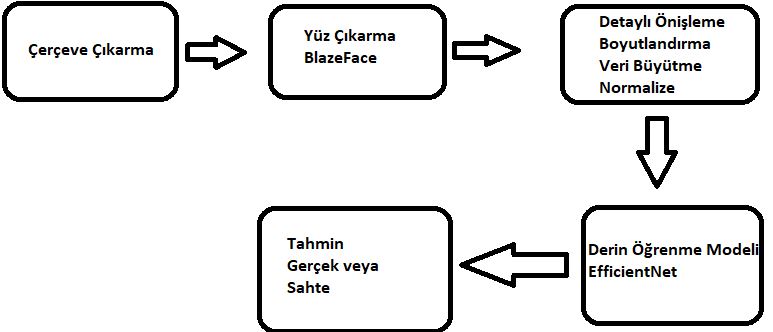
Derin sahte içerik algılama için birçok farklı yöntem kullanılabilmektedir. Klasik makine öğrenimi ve daha çok derin öğrenme yöntemleri kullanılmaktadır. Derin öğrenme yöntemlerinde ise genellikle Convolutional Neural Networks (CNN) yani Evrişimli Sinir Ağları mimarileri kullanılmaktadır. Bununla beraber bir makine öğrenmesi tekniği olan Support-Vector Machine (SVM) yani Destek Vektör Makinesi ile CNN veya CNN ile RNN veya CNN ile LSTM gibi farklı yöntemler de beraber kullanılabilmektedir. Derin sahte görüntülerin algılanmasında baş pozisyonundaki bozukluklar, dudak hareketleri, yüz bölgelerinden özellik çıkarımı – ki bu çalışmada bu yöntem kullanılmıştır – biyolojik sinyallerin tespiti, ses sinyallerinin incelenmesi gibi farklı yöntemler ile derin sahte algılama işlemlerinin başarıldığı görülmektedir[3][4].

Bu çalışmada derin sahte videoların tespit edilmesi için en başarılı veri kümelerinden olan Celeb-DFv2[14] veri kümesi kullanılmış, bu videolardan DLib kütüphanesi ile yüz bölgeleri çıkarılmıştır. Ardından elde edilen veri kümesi ile NASNetLarge CNN modeli tekrar eğitilerek derin sahte algılama modeli tamamlanmıştır. NASNetLarge modeli bir AutoML (otomatik makine öğrenmesi) tekniği olan NAS (Neural Architecture Search) yöntemi ile CIFAR-10 veri kümesini sınıflandırmak için otomatik yapak zeka ile oluşturulmuştur[5]. Ardından ImageNet veri kümesi ile modelin başarısı kanıtlanmıştır. Bu çalışmanın başlıca katkıları ise şu şekildedir:

1. Derin öğrenme yöntemleri arasında görüntü üzerinden özellik çıkarımı ve sınıflandırma yapma konusunda en başarılı yöntem olan CNN mimarilerinden NASNetLarge modelinin derin sahte algılamada ki başarısının kanıtlanması.
2. Derin sahte algılamada sadece yüz bölgesinden özellik çıkarımı ve sınıflandırma yöntemini kullanan çalışmalar arasında en zorlu veri kümelerinden olan Celeb-DFv2 veri kümesi ile rekabetçi sonuçlar elde edildiğini göstermek ve yüksek başarı oranı sağlamak.
3. Bu problemin daha küçük veri kümesi ile de başarılı bir şekilde çözülebildiğini göstermek.
4. **LİTERATÜR ARAŞTIRMASI**

Son yıllarda bilgisayarlı görü ile derin öğrenme teknolojilerinin büyük gelişim göstermesi ile derin sahte algılama konusunda da birçok yöntem geliştirilmiştir. Derin sahte video veya görüntüleri tespit etmek için çoğunlukla CNN tabanlı yöntemler kullanılmıştır.

Ş. Korkmaz ve M. Alkan tarafından yapılan çalışmada CNN mimarisi kullanan bir model olan EfficientNet ailesi ile video karelerinden özellik çıkarımı yapılmış ve bir sınıflandırıcı ağa beslenmiştir. Video karelerinden elde edilen görüntülerden yüz bölgeleri çıkarılarak elde edilen veri kümesi üzerinde belirli veri büyütme teknikleri uygulanmış, ardından model parametrelerinin eğitimi devre dışı bırakılarak EfficientNet ile eğitim yapılmıştır. DFDC veri kümesi ile yapılan deneyler sonucunda en yüksek ‘accuracy’ değeri EfficientNetB4 ile 0.918 olarak elde edilmiştir. Sonuç olarak bu yöntem EfficientNet modelinin derin sahte algılama için pozitif sonuç üreten bir model olduğunu göstermektedir. Şekil 2.1’de modelin blok şeması verilmiştir[6].



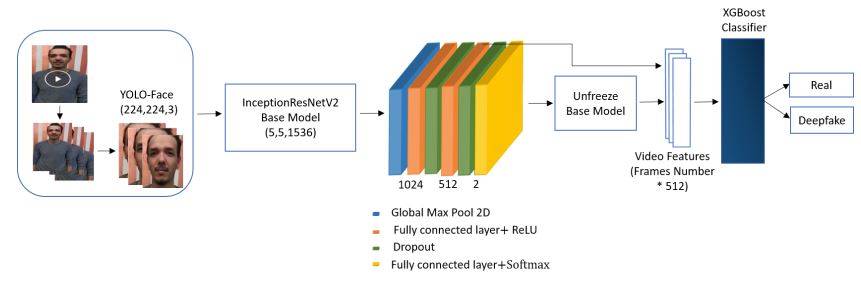
**Şekil 2.1.** EfficientNet[6] Modelin Blok Şeması

O. Giudice ve arkadaşları tarafından yapılan çalışmada, farklı çalışmalardan edinilen bilgilere göre insan yüzlerinin derin sahte görüntülerinde yapılan deneylere göre verilen uzamsal frekanslarda üretici sürecin uygun bir imzasının gömülü olduğu kanıtlanmıştır. Bu çalışmada kullanılan asıl yöntem derin sahtenin oluşturulması sırasında Generative Adversarial Network (GAN) motorlarının bıraktığı izlerin, geçici frekanslar analiz edilerek tespit edilmesidir. Bu benzersiz parmak izleri GAN Spesifik Frekansları (GSF) olarak adlandırılır. Bu izlerin yakalanabilmesi için ayrık kosinüs dönüşümü aracılığı ile AC katsayıları hesaplanmaktadır. Bu AC katsayılarının analizi ile etkileyici genelleme ve sınıflandırma sonuçları elde edilmektedir. Bu çalışmada CelebA ve FFHQ veri kümelerindeki gerçek görüntüler kullanılarak farklı yöntemler ile derin sahte görüntüler oluşturulmuş ve bu verilerden veri kümesi oluşturulmuştur. Bu görüntüler yansıtma, ölçekleme, döndürme, rastgele renk, konum ve boyutlu dikdörtgen ekleme gibi farklı saldırı türleri ile değiştirilmiştir. Bu yöntem ile yüksek başarı sağlanmıştır[8].

T. Zhao ve arkadaşlarının çalışmasındaki derin sahte algılama yöntemi görüntülerin derin sahte oluşturma sürecinden geçtikten sonra korunabileceği ve çıkarılabileceği hipotezine dayanmaktadır. CNN’i derin sahte algılamak üzere eğitmek için çift yönlü kendi kendine tutarlı öğrenme (Pair-Wise Self-Consistency Learning - PCL) adı verilen temsili öğrenme yaklaşımı sunulmaktadır. PCL için gerekli olan zengin açıklamalı veri setini oluşturmak için tutarsız görüntü oluşturucu (I2G) kullanılmaktadır. Bu yöntemde genellikle artefaktları kullanan diğer yöntemlerin aksine, sahte görüntülerdeki kaynak özelliklerin tutarsızlıkları kullanılır. Kavramsal olarak görüntüler kaynaklarını benzersiz olarak tanımlayabilen içerikten bağımsız, uzamsal olarak yerel bilgiler taşımaktadır. Bu nedenle sahte bir görüntü farklı konumlarda farklı kaynak görüntüler içerirken, bozulmamış görüntünün tüm konumlarda tutarlı olması beklenir. Yerel kaynak özellikleri çıkarılarak ve görüntünün tutarlılığı ölçülerek derin sahte tespit edilebilmektedir[9].

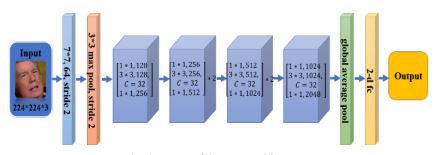
S. Fung ve arkadaşları tarafından yapılan çalışmada denetimsiz karşılaştırmalı öğrenme yoluyla yeni bir derin sahte algılama yöntemi tasarlanmaktadır. Önce bir görüntünün iki farklı dönüştürülmüş versiyonu oluşturulmakta ve bu görüntüler iki ardışık alt ağa, yani bir kodlayıcı ve bir projeksiyon başına beslenmektedir. Denetimsiz eğitim, projeksiyon başı çıktılarının, uygunluk derecesinin maksimize edilmesiyle elde edilmektedir. Derin sahte tespiti için denetimsiz öğrenme daha nadir incelenen bir konudur, bu çalışmada bu yöntem ile rekabetçi sonuçlar elde edilmiştir. Eğitim sırasında etiketler bilinmediğinden denetimli öğrenmeye göre daha zordur. Çalışmada farklı veri kümeleri üzerinde deneyler yapılmış ve Celeb-DF veri kümesi ile 90.5 ‘AUC’ sonucu üretilmiştir[10].

A. Ismail ve arkadaşları tarafından yapılan çalışmada ise State-of-the-Art denilen başarısı kanıtlamış CNN modelleri üzerine XGBoost sınıflandırıcı olarak eklenerek deneyler yapılmıştır. Asıl önerilen yöntem YOLO-InceptionResNetv2-XGBoost (YIX) olarak gösterilmektedir. YOLO yüz detektörü ile video karelerinden yüz bölgesi çıkarılmakta, InceptionResNetv2 ile bu yüz görüntülerinden özellik çıkarımı yapılmakta ve son olarak XGBoost ile sınıflandırma işlemi yapılmaktadır. Bu çalışmada özellik çıkarmak için pek çok farklı CNN modeli denenmiş ve en iyi sonuç InceptionResNetv2 ile elde edilmiştir. Bu çalışmada Celeb-DF ve FF++ veri kümelerinden belirli sayıda veri alınmış ve ortak bir veri kümesi oluşturulmuştur. Sonuç olarak ise ‘accuracy’ değeri 0.90 olarak elde edilmiştir. Şekil 2.2’de modelin blok şeması verilmiştir[7].



**Şekil 2.2.** YIX[7] Modelin Blok Şeması

D. Gong ve arkadaşları tarafından yapılan çalışmada ise 20 ağ katmanından oluşan DeepfakeNet adlı bir derin sahte algılama modeli önerilmektedir. Ağ blok yapısını, yani ResNeXt'in blok yapısını tasarlamak için ResNet'in yığınlama fikrini ve Inception'ın bölme-dönüştürme-birleştirme fikirlerini benimsemektedir. Veri ön işleme modülünde önce video veri seti işlenir. Ardından yüz görüntülerinin özellikleri CNN tarafından çıkarılır. Yeterli eğitim ve doğrulamadan sonra DeepfakeNet modeli daha iyi sonuçlar elde etmek için sürekli olarak geliştirilir. Daha doğru sonuçlar elde edebilmek için, veri kümesini geliştirmenin yaygın yöntemlerinden olan germe, döndürme, çevirme ve parlaklık değiştirme yöntemleri ile veriler çoğaltılır. Çalışma, FaceForensics++, Kaggle ve TIMIT veri kümelerinin bazı verilerini kullanmaktadır ve 96,69 ‘accuracy’ elde edilmiştir[11].



**Şekil 2.3.** DeepfakeNet[11] Model Mimarisi

X. Jin ve arkadaşları ise derin sahte algılama için biyoloiik sinyallerin incelenmesini önermiştir. Bu çalışmada, gerçek ve sahte videolar arasındaki biyolojik sinyal farkını üç boyutlu olarak çıkaran, çok boyutlu biyolojik sinyallerle yeni bir tespit yöntemi önerilmektedir. Diğer teknolojilerle karşılaştırıldığında, önerilen yöntem yalnızca sahte video bilgilerini çıkarır ve belirli bir üretim yöntemiyle sınırlı değildir, bu nedenle sentetik yöntemlerden etkilenmez ve iyi uyarlanabilirliğe sahiptir. Bu makaleye göre son çalışmalar, kalp atış hızı sinyallerinin gerçek ve sahte videoları etkili bir şekilde ayırt etmek için kullanılabileceğini göstermektedir. Bu nedenle videolardaki biyolojik sinyalleri ayıklayıp analiz ederek gerçek ve sahte videoları sınıflandırılmaktadır. Yöntemin eğitimi için üç genel veri kümesi kullanılmıştır, FF++, DFD, UADFV[12].

B. Xu ve arkadaşları tarafından yapılan çalışmada ise video karesinden elde edilen görüntünün doku özellikleri incelenerek derin sahte tespiti yapılmaktadır. Bazı derin sahte videolarının yüz dokusu detaylarından yoksun olduğu gözlemlerine dayalı olarak, derin sahte videolarını tespit etmek için geleneksel makine öğrenme teknolojilerini kullanan yeni bir yöntem önerilmiştir. İlk olarak, doku özellikleri, yüz dokusu ayrıntılarını temsil edebilen her karenin yüz bölgesinden görüntü gradyanı, standart sapma, gri seviyeli ortak oluşum matrisi ve dalgacık dönüşümü kullanılarak çıkarılmaktadır. İkinci olarak, doku özelliklerine dayalı olarak, derin sahte videoların tespitini gerçekleştirmek için SVM sınıflandırıcı kullanılmaktadır. Bu yöntem farklı veri kümelerinde yüksek başarılar elde etmiş fakat Celeb-DF veri kümesinde 75.7 ‘accuracy’ sonucu elde etmiştir[13].

Literatürdeki araştırmalarında görüldüğü üzere derin sahte algılamada genellikle derin öğrenme tekniklerinden CNN mimarileri tercih edilmekte, farklı yöntemlerde SVM[17][19], RNN[20] ve LSTM[18][21][22] gibi mimarilerde kullanılmaktadır.

Ayrıca derin sahte algılama için en fazla kullanılan veri kümeleri, çerçeve sayıları ve oluşturulma yöntemleri tablo 2.1’de görülmektedir[23]. Derin sahte algılama teknolojilerinin büyümesi ve iyileştirilmesinde büyük katkıları olan ilk veri kümeleri UADFV[24] ve DF-TIMIT[25]’tir. FaceForensics++[26] ise Youtube’dan indirilen gerçek görüntüler ve bu görüntülerin dört farklı yöntem kullanılarak sahteleştirilmesi ile oluşturulmuştur. Google tarafından Jigsaw ile işbirliği içinde yayınlanan DeepFakeDetection[26] veri seti (DFD), 16 farklı sahnede 28 farklı oyuncudan 363'ün üzerinde orijinal videonun yanı sıra derin sahte kullanan 3000'den fazla manipüle edilmiş video içermektedir. DFDC[27] veri kümesi yaklaşık 128 bin videodan oluşmakta ve 10 milyonlarca çerçeve barındırmaktadır. Tabloda DFDC’nin örnek veri kümesi verilmiştir. Celeb-DF[14] veri seti, farklı yaş, etnik grup ve cinsiyete sahip genelde ünlülerin bulunduğu 590 gerçek video ve bunların sentezlenmesiyle oluşturulan 5639 derin sahte video sağlar. Bu çalışmada kullanılan Celeb-DFv2 veri kümesinde ise Youtube’dan alınmış ekstra 300 gerçek video bulunmaktadır.

**Tablo 2.1.** Veri Kümeleri Oluşturulma Yöntemleri ve Video-Çerçeve Sayıları

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Veri Kümesi | G. Video | G. Çerçeve | S. Video | S. Çerçeve | Yöntem |
| UADFV | 49 | 17.3B | 49 | 17.3B | FakeAPP |
| DF-TIMIT | 320 | 34B | 320 | 34B | Faceswap-GAN |
| FF++ | 1000 | 509.9B | 1000 | 509.9B | Deepfakes, Face2Face  FaceSwap, NeuralTextures |
| DFD | 363 | 315.4B | 3068 | 2.2M | FF++’a Benzer |
| DFDC-Preview | 1131 | 488.4B | 4113 | 1.7M | Deepfake, GAN-tabanlı ve  Doğrusal Olayan Yöntemler |
| Celeb-DF | 590 | 225.4B | 5639 | 2.1M | Geliştirilmiş Derin Sahte Sentez Algoritması |

Önerilen model, CNN tabanlı özellik çıkarma ve sınıflandırma yöntemleri arasında Celeb-DFv2 veri kümesi üzerinde, daha az veri kullanarak yüksek başarı elde etmektedir. Veri kümesini azaltarak eğitim sürecinin hızlandırılması ve Celeb-DFv2 veri kümesi üzerinde yüksek başarılı sonuç geliştirmesi, önerilen modelin sunduğu katkılardır.

1. **ÖNERİLEN YÖNTEM VE DENEY SONUÇLARI**
   1. **Önerilen Yöntem**

Literatürdeki araştırmalardan yola çıkarak, derin öğrenme teknikleri ile derin sahte algılamanın yüksek başarı ile sağlandığı anlaşılmaktadır. Bu araştırmalar sonucunda önerilen yöntem de CNN ile özellik çıkarımı ve sınıflandırma olarak belirlenmiş ve henüz derin sahte algılama için kullanılmamış, görüntü sınıflandırma işleminde kendini kanıtlamış bir yöntem olan NASNetLarge modelinin kullanılmasına karar verilmiştir. Önerilen yöntemde sırası ile öncelikle Celeb-DFv2 veri kümesindeki video karelerinden yüz bölgeleri çıkarılacak ve oluşturulan bu veri kümesi veri arttırma teknikleri ile büyütülecektir, ardından NASNetLarge ile veri kümesi üzerinden özellik çıkarımı ve sınıflandırma yapılacaktır. Şekil 3.1’de gerçekleştirilen çalışmanın aşamaları gözükmektedir. Bu çalışmanın sunduğu asıl katkı dördüncü aşamada görülmekte olan NASNetLarge modelinin derin sahte algılamak için eğitilmiş olmasıdır.



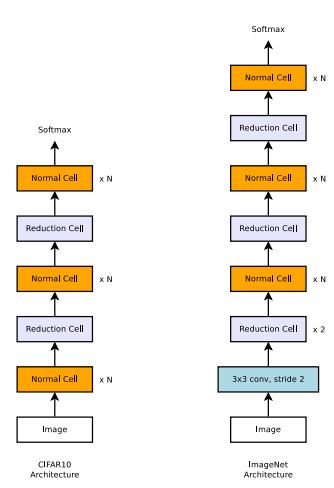
**Şekil 3.1.** Modelin Gerçekleştirme Aşamaları

CNN mimarileri görüntüyü çeşitli katmanlarda işlemektedir. Görüntü üzerinden özellikleri saptamak için konvülisyon katmanı bulunmaktadır. Bu katmanda görüntü üzerinde filtre ile maskeleme işlemi uygulanır ve uygulanan maskeye göre görüntünün özellikleri çıkarılır. Örneğin maskeleme ile bir görüntünün kenar bölgeleri çıkarılabilir. Bu katmandan sonra CNN mimarilerinde genelde pooling katmanı gelmektedir. Bu katmanda da bir filtre boyutu seçilir ve bu filtre görüntü üzerinde gezerek, görüntünün daha önemli bölgelerini seçerek küçültülmesini sağlar. Bu iki işlem defalarca gerçekleşir ve görüntüyü sınıflandırmaya yarayacak önemli ve önemsiz özelliklerin çıkarımı yapılır. Önerilen yöntem de bu şekilde işleyen bir CNN modeli olan NASNetLarge modelini kullanmaktadır.

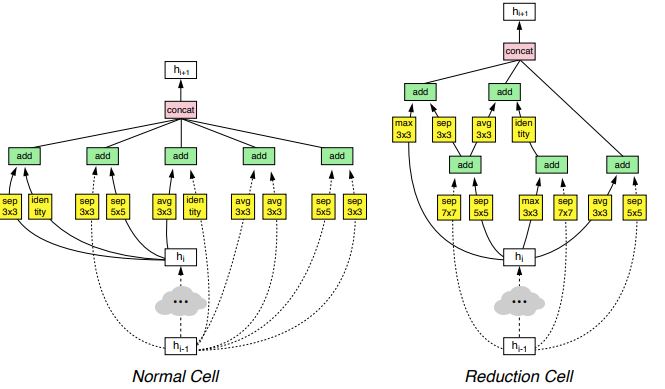
* 1. **NASNetLarge Mimarisi**

Sinir ağı görüntü sınıflandırma modelleri geliştirmek genellikle önemli mimari mühendislik gerektirir. NASNetLarge oluşturulurken, model mimarilerini doğrudan ilgilenilen veri kümesi üzerinde öğrenmek için bir yöntem üzerinde çalışılmıştır. Bu yaklaşım, veri kümesi büyük olduğunda pahalı olduğu için, küçük bir veri kümesinde bir mimari yapı bloğu aramayı ve ardından bloğu daha büyük bir veri kümesine aktarmayı kullanmış, küçük veri kümesi olarak CIFAR-10, büyük veri kümesi olarak ImageNet kullanmıştır. Deneylerde, CIFAR-10 veri kümesinde en iyi evrişim katmanı (veya "hücresi") aranmış ve daha sonra bu hücrenin daha fazla kopyası toplanarak ImageNet veri kümesine uygulanmıştır. Yaklaşım, mimari konfigürasyonları optimize etmek için pekiştirmeli öğrenme arama yöntemini kullanan, yakın zamanda önerilen Nöral Mimari Arama (NAS) çerçevesinden esinlenmiştir. Arama uzayındaki tüm kıvrımlı ağlar, aynı yapıya ancak farklı ağırlıklara sahip kıvrımlı katmanlardan (veya “hücreler”) oluşur. Bu nedenle, en iyi evrişimsel mimarileri aramak, en iyi hücre yapısını aramaya indirgenmiştir.

Yaklaşımda, konvolüsyonel ağların genel mimarileri manuel olarak önceden belirlenir. Her bir evrişim hücresinin aynı mimariye sahip olduğu, ancak farklı ağırlıklara sahip olduğu, birçok kez tekrarlanan evrişim hücrelerinden oluşurlar. Herhangi bir boyuttaki görüntüler için kolayca ölçeklenebilir mimariler oluşturmak üzere, bir özellik haritasını girdi olarak alırken iki ana işleve hizmet edecek iki tür evrişim hücresi kullanılır: aynı boyutta bir özellik haritası döndüren evrişim hücreleri ve özellik haritası yüksekliği ve genişliğinin iki kat azaltıldığı bir özellik haritası döndüren evrişim hücreleri. Evrişimli hücrelerin birinci tipi ve ikinci tipi sırasıyla Normal Hücre ve İndirgeme Hücresi olarak adlandırılmaktadır. Bu hücre yapıları NAS modeli ile oluşturulmuş ve elde edilen NASNetLarge modeli ile %82,7 top-1 ve %96,2 top-5'lik son teknoloji doğruluk elde edilmiştir. Bu sonuç, 9 milyar daha az FLOPS'a sahipken, insan tarafından icat edilen en iyi mimarilere göre ilk 1 doğrulukta %1.2'lik bir iyileşme anlamına gelir. CIFAR-10'un kendisinde NASNet, aynı zamanda son teknoloji olan %2,4 hata oranına ulaşır. Şekil 3.2’de modelin hücre mimarisi ve 3.3’de hücrelerin yapısı görülmektedir[5].



**Şekil 3.2.** NASNetLarge Modelinin Hücre Mimarisi[5]

****

**Şekil 3.3.** Hücrelerin Yapısı[5]

* 1. **Önerilen Yöntemin Katman Yapısı**

Bu modelin ImageNet veri kümesindeki bu başarılı sonuçları göz önüne alındığında derin sahte algılama için etkili bir CNN modeli olacağı anlaşılmış ve önerilen yöntemde özellik çıkarımı için kullanılmıştır. NASNetLarge parametreleri ‘ImageNet’ ağırlıkları kullanılarak, oluşturulan veri kümesi ile tamamen baştan eğitilmiş, model sonuna öncelikle bu modelin çıktılarını sınıflandırıcı katmana girdi olarak verebilmek için bir düzleştirici(flatten) katman eklenmiş, ardından aşırı öğrenme (overfitting) olayını önlemek amacı ile bir bırakma (dropout) katmanı eklenmiş, son olarak bir tane yapay nörona sahip ve aktivasyon fonksiyonu ‘sigmoid’ olan bir fully connected layer yani tam bağlı katman eklenerek sınıflandırma işlemi yapılmıştır. Ayrıca model derlenirken kayıp(loss) fonksiyonu olarak ikili sınıflandırma problemlerinde kullanılan ‘binary\_crossentropy’, optimize edici (optimizer) olarak ‘Adam’ kullanılmış ve öğrenme oranı(learning rate), beta1 ve beta2 parametreleri sırası ile ‘1e-5’, 0.9, 0.999 olarak seçilerek optimize bir sonuç elde edilmiştir. Tablo 3.1’de model katmanları ve parametreleri görülmektedir.

**Tablo 3.1.** Önerilen Modelin Katman Detayları

|  |  |  |
| --- | --- | --- |
| Katman (Tip) | Çıktı | Parametre Sayısı |
| NASNet (Functional) | (None, 11, 11, 4032) | 84916818 |
| Flatten\_1 (Flatten) | (None, 487872) | 0 |
| Dropout\_1 (Dropout) | (None, 487872) | 0 |
| Dense\_1 (Dense) | (None, 1) | 487873 |
| Toplam Param : 85,404,691 | Eğitilebilir Param: 85,208,023 | Eğitilemez Param: 196,668 |

* 1. **Deney Sonuçları**

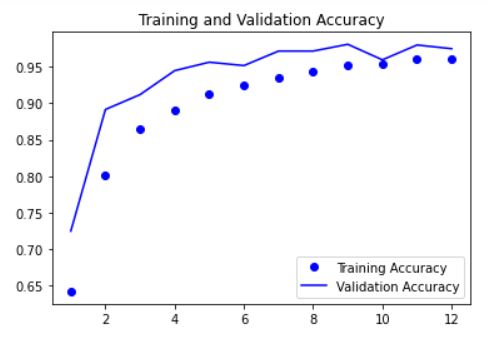
**Veri kümesinin seçilmesi.** Öncelikle veri kümesi araştırması yapılmış ve veri kümesi belirlenmiştir. Daha önce incelenen çalışmalara da bakarak en fazla kullanılan veri kümeleri incelenmiş ve insan gözü ile ayırt etmekte en çok zorlanılan, daha gerçekçi sahte videolara sahip Celeb-DFv2 veri kümesi tercih edilmiştir. Ayrıca yine incelenen çalışmalar göz önüne alındığında, CNN ile derin sahte algılama yöntemleri açısından en düşük başarı sağlanan veri kümesi olduğu anlaşılmıştır[4][10]. Bu sebeplerden dolayı Celeb-DFv2 veri kümesi seçilmiş ve yüksek bir başarı elde etmek amaçlanmıştır. Bu veri kümesinde 5639 sahte ve 890 gerçek video bulunmaktadır.

**Veri kümesinin ayarlanması.** Bu aşamada öncelikle her video ayrı ayrı açılmış ve DLib yüz detektörü kullanılarak yüz bölgeleri çıkarılmıştır. Bu çalışmada küçük bir veri kümesi ile başarı elde edilmek istendiğinden, videolardaki her kare işlenmemiştir. Yaklaşık olarak videoların her saniyesinden bir yüz görüntüsü elde edilmiş, ardından elde edilen görsellerden el ile seçim yapılarak veri kümesi biraz daha küçültülmüştür. Normal şartlarda sahte video çok olduğundan sahte görüntü sayısı çok daha fazla olmaktadır. El ile seçim kısmında bu durum düzeltilmiş, birbirine yakın sayıda sahte ve gerçek görüntü alınmıştır. Tablo 3.2’de veri kümesindeki görüntü sayıları gözükmektedir.

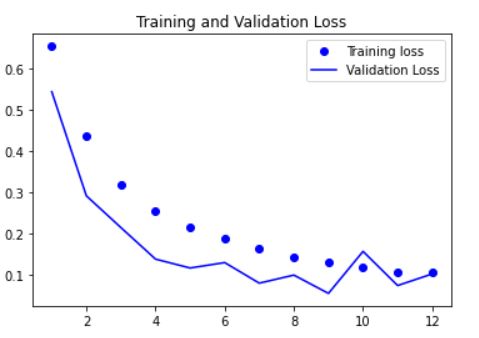
**Tablo 3.2.** Veri Kümesi Sınıf Dağılımı

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sahte | Gerçek | Toplam |
| Train | 8397 | 6726 | 15123 |
| Validation | 1200 | 962 | 2162 |
| Test | 1537 | 1560 | 3097 |

**Modelin eğitilmesi.** Bu aşamada başlık 3.3’te tanıttığımız model, bir önceki başlıkta gösterilen veri kümesi ile eğitilmiştir. Eğitim başlamadan önce veri kümesi üzerinde Image Augmentation (Veri Arttırımı) yöntemlerinden normalize etme (0-1 aralığına indirgeme), döndürme, kaydırma vb. pek çok yöntem kullanılarak veriler rastgele dönüşüme tabi tutulmuştur. Ardından modelin eğitimi için bir erken durdurma (EarlyStopping) fonksiyonu ayarlanmış, val\_loss parametresi izlenerek eğitim 20 epoch için başlatılmıştır. Eğitim toplam 12 epoch sürmüş, en iyi sonuç 9. epoch’da elde edilmiştir. 9. epoch’tan sonra ‘val\_loss’ parametresinde bir iyileşme olmamasından dolayı eğitim 12. epoch’tan sonra durdurulmuştur. Şekil 3.3 modelin 12 epoch için ‘accuracy’ değişimini, şekil 3.4 ise ‘loss’ değişimini göstermektedir.



**Şekil 3.3.** Modelin Training ve Validation Accuracy Grafiği



**Şekil 3.4.** Modelin Training ve Validation Loss Grafiği

**Modelin test edilmesi.** Son olarak model test verileri kullanılarak değerlendirilmiş (evaluate), ve model başarısı 0.967 accuracy ve 0.09 loss olarak elde edilmiştir. Test verilerinin her biri ayrı ayrı tahmin (predict) işlemine girdiğinde ise tablo 3.3’de görüldüğü üzere yüksek başarı elde edilerek modelin tahmin doğruluğu kanıtlanmıştır. Sonuç olarak tahminlerin yaklaşık %93’ü başarılı gerçekleşmiştir. Şekil 3.5’de test verileri üzerinde gerçekleştirilen tahmin işlemi için bir örnek görülmektedir. Şekilde de anlaşılacağı üzere Celeb-DFv2 veri kümesi oldukça gerçekçi sahte görüntüler içermesine rağmen sınıflandırma başarılı bir şekilde gerçekleştirilmektedir.

**Tablo 3.3.** Test Verileri ile Tahmin(Predict) İşlemi

|  |  |
| --- | --- |
| Doğru Tahmin Edilen Gerçek Görüntü | 1519 |
| Yanlış Tahmin Edilen Gerçek Görüntü | 41 |
| Doğru Tahmin Edilen Sahte Görüntü | 1343 |
| Yanlış Tahmin Edilen Sahte Görüntü | 194 |



**Şekil 3.5.** Eğitilen Model ile Tahmin İşlemi

**Modelin benzer modellerle kıyaslanması.** Önerilen model, incelenen benzer modeller arasında Celeb-DF veri kümesini kullanan modellere göre üstünlüğünü açıkça göstermektedir. Bunu yanı sıra geri kaldığı modellerden de bu sebeple geri kaldığı düşünülmektedir. Celeb-DF veri kümesi daha az artefakt içeren çok daha iyi derin sahte videolar içerdiğinden daha zorlu bir veri kümesidir. Veri kümeleri gözle incelendiğinde de sahteliği en anlaşılmayan veri kümesi Celeb-DF veri kümesi olarak görülmektedir. Ayrıca [4][10][14][16] kaynaklardan da bu veri kümesi ile derin sahte algılamada diğerlerinden düşük başarıları elde edildiği görülmektedir. Tablo 3.4’de kıyaslama görülmektedir.

**Tablo 3.4.** Başarım Kıyaslaması

|  |  |  |  |
| --- | --- | --- | --- |
| **YÖNTEM** | **SINIFLANDIRICI** | **VERİ KÜMESİ** | **ACC-AUC** |
| EfficientNetB4**[6]** | CNN | DFDC | 0.918077 |
| EfficientNetV2**[15]** | CNN | FF++ ve FFIW10K | 0.97 |
| YIX**[7]** | CNN - XGBoost | Celeb-DF ve FF++ (2848 Video) | 0.9062 |
| DeepfakeUCL**[10]** | CNN | Celeb-DF | 0.90 |
| DeepfakeNet**[11]** | CNN | FF++ - Kaggle - TIMIT | 0.96 |
| NA-VGG**[16]** | CNN | Celeb-DF | 0.85 |
| **NASNetLarge(Önerilen Model)** | **CNN** | **Celeb-DFv2** | **0.967** |

1. **Sonuçlar**

Bu çalışmada günümüzün önemli problemlerinden olan derin sahta içeriklerin algılanması için eğitimi kolay derin sahte algılama oranı yani başarısı yüksek bir model önerilmiştir. Önerilen model, CNN mimarisi ile özellik çıkarımı yaparak derin sahte algılama için oluşturulan veri kümeleri arasında en zorlu veri kümesi olabileceği gösterilmiş Celeb-DFv2 veri kümesinde en yüksek başarıyı elde etmiştir. Ayrıca derin sahte algılama probleminin çözümünde genellikle çok büyük veri kümeleri kullanılırken bu çalışma daha küçük veri kümesi ile başarı elde edilebileceğini göstermiştir. Önerilen yöntem CNN ile özellik çıkarımı dışındaki yöntemlerle kıyaslandığında da rekabetçi ve yüksek performanslı bir sonuç elde etmiştir. Sonuç olarak model test edildiğinde 0.967 ‘accuracy’ ve 0.09 ‘loss’ elde edilmiş, ardından test verileri üzerinde yapılan tahmin işlemi ile %93 doğru sınıflandırma yapılmıştır. CNN mimarileri ile özellik çıkarımı ve sınıflandırma yöntemleri arasında Celeb-DF veri kümesi ile en yüksek başarı elde edilmiştir. Gelecek çalışmalarda, elde edilen bu başarının farklı sınıflandırıcılar ile denenerek daha da yükseltilmesi ve projenin web ortamına aktarılması ile insanların yararlanacağı bir çıktı üretilmesi amaçlanmaktadır.

**KAYNAKLAR**

[1]N. Aalami, Derin öğrenme yöntemlerini kullanarak görüntülerin analizi, Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, no. 1, pp. 17-20, 2020.

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

[3] N. Guhagarkar, S. Desai, S. Vaishyampayan, A. Save , Deepfake Detection Techniques: A Review, VIVA Institute of Technology 9 th National Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021), ISSN (Online): 2581-7280

[4] İ. İlhan ve M. Karaköse , "DERİN SAHTE VİDEOLARIN TESPİTİ VE UYGULAMALARI İÇİN BİR KARŞILAŞTIRMA ÇALIŞMASI", Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi, c. 8, sayı. 14, ss. 47-60, Haz. 2021

[5] Zoph, Barret & Vasudevan, Vijay & Shlens, Jonathon & Le, Quoc. (2018). Learning Transferable Architectures for Scalable Image Recognition. 8697-8710. 10.1109/CVPR.2018.00907.

[6] Ş. Korkmaz and M. Alkan , "Derin Öğrenme Algoritmalarını Kullanarak Deepfake Video Tespiti", Politeknik Dergisi, pp. 1-1, Dec. 2022, doi:10.2339/politeknik.1063104

[7] Ismail, A.; Elpeltagy, M.; S. Zaki, M.; Eldahshan, K. A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost. Sensors 2021, 21, 5413.

[8] Giudice, O.; Guarnera, L.; Battiato, S. Fighting Deepfakes by Detecting GAN DCT Anomalies. J. Imaging 2021, 7, 128.

[9] T. Zhao, et al., "Learning Self-Consistency for Deepfake Detection," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021 pp. 15003-15013.

[10] S. Fung, X. Lu, C. Zhang and C. -T. Li, "DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning,"2021 International Joint Conference on Neural Networks (IJCNN)*,* 2021, pp. 1-8

[11] Dafeng Gong, Yogan Jaya Kumar, Ong Sing Goh, Zi Ye and Wanle Chi, “DeepfakeNet, an Efficient Deepfake Detection Method” International Journal of Advanced Computer Science and Applications(IJACSA), 12(6), 2021.

[12] Jin, Xinlei & Dengpan, Ye & Chen, Chuanxi. (2021). Countering Spoof: Towards Detecting Deepfake with Multidimensional Biological Signals. Security and Communication Networks. 2021. 1-8.

[13] B. Xu, J. Liu, J. Liang, W. Lu and Y. Zhang, "Deepfake videos detection based on texture features," Computers, Materials & Continua, vol. 68, no.1, pp. 1375–1388, 2021.

[14] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,* 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327.

[15]Liwei Deng, Hongfei Suo, Dongjie Li, "Deepfake Video Detection Based on EfficientNet-V2 Network", Computational Intelligence and Neuroscience, vol. 2022, Article ID 3441549, 13 pages, 2022.

[16] X. Chang, J. Wu, T. Yang and G. Feng, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network," 2020 39th Chinese Control Conference (CCC), 2020, pp. 7252-7256, doi: 10.23919/CCC50068.2020.9189596.

[17] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7\_12.

[18] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[19] H. Agarwal, A. Singh and R. D, "Deepfake Detection Using SVM," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1245-1249, doi: 10.1109/ICESC51422.2021.9532627.

[20] Y. Al-Dhabi and S. Zhang, "Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)," 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), 2021, pp. 236-241, doi: 10.1109/CSAIEE54046.2021.9543264.

[21] Tariq, Shahroz & Lee, Sangyup & Woo, Simon. (2020). A Convolutional LSTM based Residual Network for Deepfake Video Detection.

[22] Patel, D., Motiani, J., Patel, A., Bohara, M.H. (2022). DeepFake Creation and Detection Using LSTM, ResNext. In: Hemanth, D.J., Pelusi, D., Vuppalapati, C. (eds) Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies, vol 101. Springer, Singapore. https://doi.org/10.1007/978-981-16-7610-9\_75

[23] Taeb, M.; Chi, H. Comparison of Deepfake Detection Techniques through Deep Learning. J. Cybersecur. Priv. 2022, 2, 89–106. https://doi.org/10.3390/jcp2010007

[24] Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8261–8265.

[25] Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. arXiv 2018, arXiv:1812.08685.

[26] Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.

[27] Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The DeepFake Detection Challenge Dataset. arXiv 2020, arXiv:2006.07397.