GAZİ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

BM496 BİLGİSAYAR PROJESİ II

LİTERATÜR TARAMASI

Sahte Fotoğraf Analizi

Dr. Öğr. Üyesi Çağrı ŞAHİN

181180030 - İsmail ERTAYLAN
181180006 - Büşra ARIK

2023

İÇİNDEKİLER

KISALTMALAR

Kısaltmalar	Açıklamalar
AI	Artificial Intelligence
AUC	Area Under the Curve
CASIA	Chinese Academy of Sciences
CEW	Closed Eyes in the Wild
CFF	Common Fake Feature
CNN	Convolutional Neural Networks
ELA	Error Level Analysis
EAR	Eye Aspect Ratio
EXIF	Exchangeable Image File
FC	Fully Connected
FFHQ	Flickr Faces High Quality
GAN	Generative Adversarial Networks
IJACSA	International Journal of Advanced Computer Science and Aplications
JPEG	Joint Photographic Experts Group
LRCN	Long-term Recurrent Convolutional Networks
MLP	Multi-Layer Perceptrons
RELU	Rectified Linear Unit
RNN	Reccurrent Neural Network
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TL	Transfer Learning

1. GİRİŞ

Günümüzde yapay zekanın yaygınlaşması ile teknoloji ciddi boyutlarda gelişmiştir. Yapay zeka konularından biri de derin sahte fotoğraf ve videolardır. Bu medya içerikleri için gerekli modeller yapay zeka ile yapılmaktadır. Fonksiyonel açıdan derin sahte, bireyin yüzü veya bedeninin tamamı ile yapay zeka teknolojisinden faydalanılarak hareket ve konuşmaların değiştirilmesidir.

Yapay zekanın ve makine öğrenmesi algoritmalarının gelişmesi ile daha gerçekçi hale gelen derin sahte içerikler tehlikeli hale gelmektedir. Örneğin ünlü bir siyasinin sahte bir suç videosu yayınlanabilir. Bu teknoloji ileride ciddi sorunlar oluşturabilir hale gelmektedir. Bu durumda ise teknolojinin kötüye kullanımını yine bir başka teknoloji tarafından durdurulup, derin sahte içeriklerin tespiti yapılabilmektedir. Büyük öneme sahip olan bu teknoloji aynı zamanda hukuki anlamda siber güvenlik sorunlarında da ciddi bir role sahiptir. Derin sahte içeriklerin tespiti de oldukça günceldir ve gelişmeye devam etmektedir. Bu içerikler üretilirken çeşitli algoritmalardan yararlanılmaktadır ve tespitinde de benzer durum söz konusudur.

Derin sahte içeriklerin tespitinde yüz ifadeleri ve kafa hareketleri önemli ipuçları vermektedir. Renk tutarsızlıkları, bozuk dokular, optik hatalar, artefaktlar ve izler de derin sahte tespitinde öne çıkmaktadır. Yüz değişimleri için dişler ve göz yansımaları, tutarsız ağız ve dudak hareketleri analiz edilmelidir. Göz kırpma anormallikleri algoritmalar tarafından yeterince geliştirilmemiş ve tespiti kolaylaştıran unsurlardandır. Göz kırpmanın frekansı olağandışı olabilmektedir. Araştırmacılar yöntemlerde artefaktlara, tutarsız renklere, doku bozulması ve parmak izlerine odaklanmışlardır. Baş pozisyonundaki farklılıklar, bahsedilen yapay göz kırpmaları ve yüz çarpıklıkları da araştırmacıların konularından olmuştur.

Yüz artefaktları da tespit yöntemlerinde kullanılır ve bunlar yüzde bulanıklık ve ışık farkları, ton değişiklikleri, kaydırma sonucu çift bölgelerin oluşması, titreme olmasıdır. En sık kullanılan yöntemlerden biri yüz manipülasyonlarıdır. Görüntü işlemleri yapılırken yeniden ölçeklendirme, döndürme ve sentezleme işlemleri yapılmaktadır. Bu işlemler yapılırken bazı çarpıtmalar oluşur ve bu manipülasyonlar tespit edilebilmektedir.

Sahte video oluşturma yöntemleriyle derin sahte tespit yöntemleri benzerdir. Sebebi tespit modülünün eğitim sürecinin bir parçası olmasıdır. Derin sahte tespit yöntemleri de çeşitli yöntemleri barındırmaktadır. Bu dökümanda derin sahte tespiti ile ilgili literatür taranmış ve makaleler incelenmiştir.

2. İLGİLİ ÇALIŞMALAR

2.1. Forged Face Detection using ELA and Deep Learning Techniques

Qurat-ul-ain, sahte fotoğraf analizi için CNN'i kullanan bir teknik önermiştir. Başlangıçta tüm veri kümesinin (128*128) piksel olarak yeniden boyutlandırıldığı ve normalleştirildiği pre-processing işlemi yapılmaktadır. Daha sonra ELA kullanılarak extraction işlemi yapılmıştır. İşlemden geçen görüntüler training ve test setlerine bölünüp, gerçek ve sahte görüntüleri tanımak için Deep-CNN modellerine iletilir. Bu modeller VGG-16, ResNet-50, InceptionV3 ve VGG-19'dir. VGG-16 ve 19 modelleri %91,97 ve %92,09 oranında training doğruluğu verirken, VGG-16'nın aynı veri setlerinde diğer önceden eğitilmiş modellere kıyasla daha iyi olan %64,49 test seti doğruluğu verdiği gözlemlenmiştir [1].

2.2. Methods of Deepfake Detection Based on Machine Learning

Makalede face swapping AI tabanlı algoritmalarla videonun/fotoğrafın değiştirilip değiştirilmediğine karar vermek için kullanılabilecek indikatörler açıklanmıştır. Bunlar; çok pürüzsüz bir cilt, sentezlenen yüz ile orijinal yüz arasındaki renk uyumsuzluğu, baş pozisyonu, göz kırpma oranı, küçük hareketli parçalardaki artefaktlar ve yüz çarpıtma artefaktlarıdır. Yüz çarpıtma artefaktları, düşük çözünürlüklü yüz çıktısına sahip algoritmalar (64x64 veya 128x128) tarafından oluşturulan sahte videoların en iyi indikatörlerindendir. Model olarak DenseNet169 ile yüz çarpıtma artefakt indikatörü kullanılmıştır. Modeli değerlendirmek için Celeb-DF veri seti kullanılmıştır. Bu veri setindeki içeriklerin test sonucunda doğru çıktılar verip vermediği anlaşılması için üzerinde değişiklikler yapılmıştır. İçeriklere gürültü eklenip Gauss bulanıklığı, exponential bulanıklığı ve Rayleigh bulanıklığı test edilmiştir. En yüksek AUC değerine sahip model %60.1 ile DenseNet169 + Rayleigh blur modeli olmuştur [2].

2.3. Exposing AI Generated Fake Face Videos by Detecting Eye Blinking

Bu çalışmada, sinir ağları ile oluşturulan sahte yüz videolarının tespiti için göz kırpmaya dayalı bir yöntem anlatılmıştır. Spontane göz kırpma, refleks olarak göz kırpma ve istemli göz kırpma olmak üzere 3 göz kırpma türü vardır. Burada kullanılan yöntem, göz kırpma sürecindeki fenomenolojik ve zamansal düzenlilikleri yakalamak için CNN'i, RNN ile birleştiren bir derin öğrenme modeline dayanmaktadır. Deneyde yapılan işlemler pre-processing, LRCN ve model eğitimi olmak üzere 3 başlıkta toplanmaktadır. Pre-processingte, yer işareti tabanlı yüz hizalama algoritmaları kullanılarak yüz bölgeleri hizalanır ve yüz dedektörü kullanılarak yüz işaretleri çıkarılmaktadır. LRCN aşamasında model özellik çıkartma, dizi öğrenme ve durum tahmini işlemlerinden geçirilip eğitim aşamasına gönderilmektedir. LRCN modeli, gözün açık halinin görüntü veri kümelerine göre eğitilmiştir. Daha sonra Deep Fake algoritması ile oluşturulan gerçek ve sahte videolarda göz kırpmayı algılayan algoritma test edilmiştir. Deneyde CEW veri seti kullanılmıştır. LRCN metodu, EAR ve CNN metotları ile karşılaştırılarak değerlendirildiğinde CNN görüntü sınıflandırıcısı, farklı sınıfları ayırt etmek için görüntü alanında eğitilmiştir. Göz durumunu ayırt etmek için CNN modeli olarak VGG16 kullanılmıştır. EAR metodu üst ve alt kapak mesafesi ile sol ve sağ köşe noktası arasındaki mesafe arasındaki oran açısından göz durumunu analiz etmek için göz işaretlerine yanıt vermektedir. En büyük dezavantajı tamamen göz işaretlerine bağlı olmasıdır. Hesaplamalara bakıldığında LRCN %99 başarı oranıyla en iyi sonucu verirken CNN %98 ve EAR %79 oranında başarılıdır. Ama göz durumuna göre tespit konusunda CNN olağanüstü başarılı bir sonuç vermektedir. Mevcut çalışmada eksiklik olarak dinamik göz kırpma modeli dikkate alınmamaktadır. Göz kırpma, sahte yüz videolarını tespit etmede kolay bir ipucudur ve

daha gelişmiş modeller, daha fazla eğitim verisi ile hala gerçekçi yanıp sönme efektleri oluşturabilmektedir. Bu nedenle, önerilen metot eksiklikler barındırmaktadır [3].

2.4. A Detection Method of Operated Fake-Images Using Robust Hashing

Bu makalede, görüntü işlemlerinden kaynaklanan bozulmalar da dahil olmak üzere sahte görüntüleri tespit etmek için bir yöntem önerilmiştir. Makalede, Robust hash yöntemi kullanılarak referans görüntülerden robust hash değerleri hesaplanır ve değerler veri tabanında saklanır. Referans görüntülere benzer şekilde robust hash yöntemi kullanılarak bir sorgu görüntüsünden robust hash değeri hesaplanmaktadır. Sorgunun hash değeri, veritabanında depolananlarla karşılaştırılıp, hash değerleri arasındaki mesafeye göre sorgu görüntüsünün gerçekliğine karar verilmektedir. Sahte görüntü algılamaya yönelik hash değerlerinin, sıkıştırma ve yeniden boyutlandırma gibi bir dizi görüntü işlemi türüne karşı yeterince sağlam olması gerekir çünkü bu tür bir işlem, görüntülerin kalitesini düşürmesine rağmen görüntülerin içeriğini değiştirmez. Bu nedenle, sorgu görüntülerine benzer görüntüleri sağlam bir şekilde almayı amaçlayan robust hashing yöntemi kullanılmıştır. Buna karşılık, robust hash yöntemi kullanılarak oluşturulan hash değerlerin, kopyala-taşı ve GAN'lar gibi sahte görüntüler oluşturmak için kullanılan manipülasyonun etkisine duyarlı olması gerekmektedir. Bu gereksinimler altında, Li et al.'s yönteminin sahte görüntü tespiti için uygun bir performansa sahip olduğu anlaşılmıştır. Deneyde Görüntü Manipülasyon Veri Kümesi, UADFV, CycleGAN ve StarGAN veri setleri kullanılmıştır. Orijinal görüntüler referans olarak kullanılmış, her deney için farklı sahte görüntü veri seti ile oluşturulan ayrı bir referans veri seti hazırlanmıştır. Sorgu görüntüleri olarak hem orijinal görüntüler hem de sahte görüntüler kullanılmıştır. Bu deneyde, önerilen yöntem, gerçek sorgu görüntüleri olmasına rağmen, veri kümelerinden gelen sorgu görüntülerinin herhangi bir ek işlem yapılmadan doğrudan kullanıldığı Wang'ın yöntemi ve Xu'nun yöntemiyle karşılaştırılmıştır. Wang'ın yöntemi, sınıflandırıcının ProGAN kullanılarak eğitildiği GAN modelleriyle birlikte çeşitli CNN'ler tarafından oluşturulan görüntüleri tespit etmek için önerilmiştir. Önerilen yöntemin neredeyse tüm kriterler açısından daha yüksek bir doğruluğa sahip olduğu gösterilmektedir. Ayrıca, Görüntü Manipülasyonu ve UADFV veri kümeleri kullanıldığında geleneksel yöntemlerin doğruluğu oldukça azalmıştır. Bunun nedeni, geleneksel olanların CNN'ler kullanılarak oluşturulan sahte görüntüleri tespit etmeye odaklanmasıdır. Görüntü Manipülasyonu Veri Kümesi, GAN'larla oluşturulan görüntülerden oluşmaz. Ayrıca UADFV derin sahte videolardan oluşsa da veri setindeki videolar zaten video sıkıştırma etkisine sahiptir. Sıkıştırma için orijinal bir hash kod olduğunda, önerilen yöntem, geleneksel yöntemlere göre sahte görüntüleri daha iyi bir şekilde tespit edebilmektedir, eğer bir hash kodu yoksa görüntü tespiti yapılamamaktadır. Deneyde, önerilen yöntemin diğerlerinden daha iyi performans gösterdiği ve aynı zamanda birden fazla işlemi birleştirirken de iyi bir sonuç verdiği gözlemlenmiştir [4].

2.5. Detecting Fake Images on Social Media using Machine Learning

Bu makalede sosyal medya üzerindeki sahte fotoğrafların makine öğrenmesiyle tespiti incelenmiştir. Araştırmacı, makine öğrenimi algoritmalarını kullanan ve CNN aracılığıyla bu tespiti sağlayan sınıflandırıcı bir model önermiştir. İlgili modelde normal görsel ve sahte görsel olmak üzere iki sınıf vardır. Araştırmacı, CNN aracılığıyla derin öğrenme tekniğini kullanmıştır. Yönteme göre önce Instagram'daki IJACSA veri setinden tespiti yapılacak görüntüler elde edilir. CNN aracılığıyla geleneksel matematiksel işlemler kullanılır ve görüntü özellikleri çıkartılıp aktivasyon fonksiyonu oluşturulur. Görüntü verilerinde doğrusallık olmadığından RELU işlevi kullanılmaktadır. Dizi boyutunu küçültmek amacıyla max pooling algoritması kullanılmaktadır. Bu aşamalardan sonra tahmin gerçekleştirilir ve bir sinir ağı ile görüntünün eşleşip eşleşmediğine karar verilir. SoftMax ile çıktının olasılıklar halinde

görünmesi sağlanır. Sinir ağı eğitimi tamamlandığında da veri seti test edilir ve doğruluğun hesaplandığı değişkenleri içeren karışıklık matrisi çıkarılır. Araştırmada performans metrikleri 3 ağa göre incelenir: Alexnet, klasik CNN ve AlexnetTL. Eğitim verilerine göre ortalama sonuçlar değişse de her seferinde sıralama Alexnet-AlexnetTL-klasik CNN şeklinde olmuştur. Eğitim verilerinden yola çıkıldığında Alexnet %99.3, AlexnetTL %94 ve klasik CNN ise %83.9 oranında başarı sağlamıştır. Sonuçlardan da anlaşıldığı üzere klasik CNN modeline göre Alexnet/AlexnetTL'in kullanımı, daha başarılı çıktılar elde etmektedir [5].

2.6. Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network

Bu çalışmanın amacı, derin sahte görüntüleri tespit etmenin güvenilir bir yolunu bulmak ve CNN mimarisiyle başarılı sonuçlar elde etmektir. Çalışmada, büyük bir veri setinden derin sahte görüntüleri tespit etmek için 8 CNN mimarisi kullanılmaktadır. Bunlardan üçü DenseNet mimarisi (DenseNet121, DenseNet169 ve DenseNet201), ikisi VGGNet mimarisi (VGG16, VGG19), biri ResNet50 mimarisi, biri VGGFace mimarisi ve sonuncusu ise özel bir CNN mimarisidir. CNN, özellik çıkartma ve sınıflandırma kısımlarından oluşmaktadır. Deneyde veriler, Kaggle üzerinden toplanan bir veri kümesinden elde edilir ve daha sonra evrişim katmanına gönderilir. Bu katman, girdi olarak alınan fotoğraflardan çok sayıda özellik çıkartır. Daha sonra havuzlama katmanına geçilir. Bu katmanın amacı, evrişimli özellik katmanının boyutunu en aza indirmektir. Önceki seviyelerden gelen girdiler düzleştirilir ve girdi FC katmanına gönderilir. Düzleştirilmiş vektör üzerinde matematiksel fonksiyonel işlemleri yürütmek için başka FC katmanları kullanılır. Bu aşama fotoğrafların sınıflandırma sürecini başlatmaktadır. Deney sonunda VGGFace, doğruluk, kesinlik, F1 puanı ve ROC eğrisi altındaki alan gibi ölçümlerde en iyi performansı göstermiştir. En kötü performansı ise VGG16 göstermiştir, %92 doğruluk elde etmiştir. ResNet50 de %97 doğruluk elde etmiştir. DenseNet201 ve DenseNet169, sırasıyla %96 ve %95 doğruluk elde etmiştir. En yüksek AUC puanı, %99.8 ile VGGFace mimarisi tarafından elde edilirken, en düşük AUC puan DenseNet121 mimarisi tarafından elde edilmiştir. Yazarlar tarafından önerilen model,%90 doğruluk elde etmiştir.Genele bakıldığında VGGFace en iyi performansı göstermiştir [6].

2.7. Fake Image Detection Using Machine Learning

Çoğu görüntü dosyası resim hakkında bilgi veren meta verilerini de barındırmaktadır. Meta veriler, dosyanın nasıl oluşturulduğu ve işlendiği ile ilgili bilgiler vermektedir. Meta veride aranması gereken bilgiler şunlardır;

1. Model ve yazılım: Bunlar, resmi oluşturan cihazı veya uygulamayı tanımlar. Kameralar, EXIF bilgisi olarak marka ve model içermektedir.
2. Görüntü boyutu: Meta veriler genellikle resmin boyutlarını kaydeder. İşlenen görüntü boyutu, meta verilerdeki diğer boyutlarla eşleşiyor mu diye kontrol edilir.
3. Zaman bilgisi: Bunlar genellikle fotoğrafın çekim ve değişim tarihlerini içerir. Zaman bilgilerinin beklenen zaman dilimine uyumu kontrol edilir.
4. Meta veri türü: Meta veri türlerinin bazıları sadece kameralar tarafından üretilirken, diğerleri yalnızca uygulamalar tarafından üretilir.
5. Açıklamalar: Gömülü ek açıklamalar içerir.
6. Eksik meta veri: Belirli meta verilerin olmaması genellikle orijinal bir fotoğrafın değil, resmin kaydedildiğini göstermektedir.
7. Değiştirilmiş meta veriler: Kasıtlı olarak meta verileri değiştirilebilmektedir.

Makalede ELA ve Meta veri analizi yöntemleri birlikte kullanılmıştır. Bir JPEG'in kalitesi kayıt edildikçe düşmektedir. ELA, fotoğraftaki görüntü kalite farklarına bakarak manipülasyonu anlayabilmektedir. ELA, ImageJ kütüphanesi aracılığıyla yapılmaktadır.

ImageJ, görüntüyü belirli bir sıkıştırma yüzdesiyle JPEG formatında kaydetme seçeneği sunmaktadır. Sistem önce görüntüyü kayıpsız kaydeder. Daha sonra aynı görüntü ImageJ kullanılarak %90 kaliteli görüntüye dönüştürülür. Aradaki fark, fark yöntemiyle bulunmaktadır. Elde edilen görüntü, giriş görüntüsünün gerekli ELA görüntüsüdür. Bu görüntü, arabelleğe alınmış bir görüntü olarak kaydedilir ve daha sonraki işlemler için sinir ağına gönderilir. Eğitim sırasında dizi, çok katmanlı algılayıcı ağına girdi olarak verilir ve çıktı nöronları ayarlanır. MLP, tamamen bağlı bir sinir ağıdır ve 2 çıkış nöronu vardır. İlki sahte, ikincisi gerçek görüntüyü temsil etmektedir. Verilen görüntü sahte ise, sahte nöron bire, gerçek ise sıfıra ayarlanır. Testte, görüntü dizisi giriş nöronlarına beslenir ve çıkış nöronlarının değerleri alınır. Meta veri analizinde ise önce meta verilerin çıkarılma işlemi yapılmaktadır. Sonra meta veri metni, meta veri analizi modülüne gönderilir. Bu analiz temelde bir etiket arama algoritmasıdır. Metinde Photoshop, Gimp, Adobe vb. kelimeleri arar. Sahtelik ve gerçeklik olarak adlandırılan ve gerçek ve sahteyi temsil eden iki değişken oluşturulur. Bir etiket alındığında, analiz edilir ve karşılık gelen değişken önceden tanımlanmış belirli bir ağırlıkla artırılır ve buradan alınan sonuçlarla ELA yönteminden alınan sonuçlar birleştirilir. Bu analiz, çok küçük bir işlem altında dahi tüm 'photoshopped' veya 'gimped' görüntülerde sahteliği tespit edebilmektedir ama WhatsApp, Google+ vb. üzerinden paylaşılan görsellerde hata vermektedir. Sinir ağı CASIA veri seti ile eğitilmiştir. Eğitimli sinir ağı, görüntüyü %83 başarı oranıyla tanıyabilmiştir [7].

2.8. Image Forgery Detection Using Deep Learning by Recompressing Images

Bu yazıda, çift görüntü sıkıştırma bağlamında görüntü sahteliğini belirlemek için robust derin öğrenme tabanlı bir sistem anlatılmıştır. Bir görüntünün orijinal ve yeniden sıkıştırılmış sürümleri arasındaki fark, modeli eğitmek için kullanılmıştır. Görüntü yeniden sıkıştırıldığında, sahtelik içeriyorsa, orijinal görüntünün kaynağı ile sahte bölümün kaynağı arasındaki fark nedeniyle görüntünün sahte kısmı görüntünün geri kalanından farklı şekilde sıkıştırılmaktadır. Orijinal görüntü ile yeniden sıkıştırılmış versiyonu analiz edilir. Makalede, CNN mimarisi yaklaşımını vurgulayan, sinir ağları ve derin öğrenmeye dayalı bir görüntü sahteciliği tespit sistemi sunulmuştur. Bu yöntem, görüntü sıkıştırmasındaki varyasyonları içeren CNN mimarisini kullanmaktadır. Modeli eğitmek için orijinal ve yeniden sıkıştırılmış görüntüler arasındaki fark kullanılmıştır. Önerilen teknik, birleştirme ve kopyala-taşı ile değişiklik yapılmış fotoğraflardaki sahteliği saptayabilmektedir. Deney sonuçları, %92,23 genel doğrulama oranı göstermektedir. Mevcut teknik, minimum 128x128 çözünürlük gerektirmektedir [8].

2.9. A Landscape View Of Deepfake Techniques And Detection Methods

Bu makalede derin sahte çalışma ve kavramları, teknikleri ve algoritmaları incelenmiştir. Derin sahte içerikler manipülasyon derecesine göre tüm yüzün sentezi, kimlik değişikliği, özellik manipülasyonu ve ifade değişimi olarak 4 sınıfta toplanmaktadır. Tüm yüz sentezi, StyleGAN kullanarak aslında var olmayan tam yüz görsellerini üretmektedir. Bu sentez video oyunları, 3 boyutlu modelleme, fotoğrafçılık gibi çeşitli alanlarda avantajlar sağlamaktadır. Kimlik değiştirme yöntemi, FaceSwap6 ve DeepFakes7 gibi yöntemler ile yüz değişimini sağlar. Özellik manipülasyonu, bir GAN ve StarGAN yöntemi kullanılarak yüzde saç/ten rengi, sakal, bıyık, yaş, cinsiyet değişiklikleri gibi rötuşlara olanak sağlamaktadır. İfade değişimine bakıldığında ise, standart GAN mimarileri aracılığıyla bir kişinin mimiklerinin değişimi sağlanmaktadır. Çalışmalarda odaklanılan çeşitli noktalarla farklı sonuçlar elde edilebilmektedir. NIST MFC2018 veri setinde renk farklılıklarına odaklanılarak yapılan bir çalışma ile %70 AUC elde edilmiştir. Sinirsel davranışı izleyen başka bir çalışmada ise FakeSpotter yöntemleri kullanarak bir SVM eğitilmiştir. Önerilen teknikler CelebA-HQ,

FFHQ veri setlerinden gerçek yüz görselleri; InterFaceGAN ve styleGAN'ın ürettiği sentetik yüz görselleri kullanarak test edilmiş ve %84.7 oranında doğruluk elde edilmiştir [9].

2.10. Deep Fake Image Detection Based on Pairwise Learning

Bu makalede, kontrast kaybı kullanarak sahte görüntülerin tespiti için derin öğrenmeye dayalı bir yaklaşım önerilmektedir. Sahte-gerçek görüntü çiftlerini oluşturmak için son teknoloji GAN'lar kullanılmıştır. İndirgenmiş DenseNet, girdi olarak ikili bilgiye izin vermek için iki akışlı bir ağ yapısına dönüştürülmüştür. Ardından, önerilen ortak sahte özellik ağı, görüntüler arasındaki özellikleri ayırt etmek için ikili öğrenme kullanılarak eğitilmiştir. Son olarak, sahteliğini algılamak için önerilen ortak sahte özellik ağına bir sınıflandırma katmanı eklenmiştir. Yöntemi doğrulamak için, sahte yüz ve genel görüntüleri tanımlamak için önerilen DeepFD uygulanmıştır. Deneysel sonuçlar, yöntemin yöntemlerden daha iyi performans sağladığını göstermiştir. Önerilen ikili öğrenme stratejisi, eğitilmiş sahte görüntü dedektörünün, eğitim aşamasına dahil edilmemiş olsa bile, yeni bir GAN tarafından oluşturulan sahte görüntüyü algılama yeteneğine sahip olmasını sağlayan sahte özellik öğrenmesini sağlamaktadır. Deneysel sonuçlar, yöntemin kesinlik ve geri çağırma oranı açısından daha başarılı olduğunu göstermektedir. Yöntemin dezavantajı, eğitim örneklerinin toplanması ile ilgilidir. Bazı sahte görüntü oluşturucuların teknik detayları açıklanmadığından eğitim örneklerinin toplanması zor olabilmektedir. Bunu aşabilmek için CFF küçük bir eğitim setinden birkaç aşamalı olarak öğrenilmelidir [10].

3. TARTIŞMALAR VE SONUÇLAR

Bu dokümanda öncelikle derin sahtenin ne olduğu, teknolojideki yeri ve gelişimi, kötüye kullanımda ortaya çıkan sorunlar ve bunların önlenmesi konuları ele alınmıştır. Ayrıca derin sahte tespiti üzerine gerçekleştirilecek olan bir çalışmanın araştırmaları bulunmaktadır. Bu araştırmalardan elde edilen sonuçlar aşağıdaki gibi özetlenmiştir:

CNN temelli bir önermeye göre ara aşamalar ve ELA sonrası bazı Deep-CNN modellerinde training ve testler gerçekleştirilmiştir. Bu öneri %92'ye kadar başarı oranı sağlayabilmektedir. Yüz artefaktlarına odaklanan face swapping temalı bir başka öneri ise DenseNet169 indikatörü ve Celeb-DF veri setini kullanmıştır. Bunlara eklenen bazı bulanıklıklarla %60 gibi doğruluk oranları mümkün olmaktadır. Göz kırpma frekansının öne çıktığı bi öneri ise CNN-RNN birleşimi biri derin öğrenmeyi kullanmaktadır. Ara aşamalar ve LRCN süreci sonrası CEW veri seti aracılığıyla deep fake algoritması uygulanan bu öneri %99 gibi çok yüksek başarı oranlarına sahip olabilmektedir fakat dinamik göz kırpmanın dikkate alınmaması gibi bir dezavantaja sahiptir. Robust hash yöntemi ise görseller için orjinal bir hash kodu hesaplar ve kıyaslamalar sonrasında içeriği dönüştürmeden bazı GAN yöntemleri ile CNN'ler tarafından oluşturulan görüntüleri tespit eder. Sosyal medyadaki sahte fotoğraflara odaklanan başka bir önerme ise IJACSA veri setini kullanarak CNN aracılığıyla matematiksel işlemler ile sınıflandırma sağlar. RELU işlevi ve max pooling algoritması ile düzeltmeleri sağlar. Alexnet aracılığıyla %94 gibi yüksek başarı oranlarını yakalamaktadır. 8 farklı CNN mimarisini karşılaştırmalı analiz eden bir başka çalışmadaki sonuçlara bakıldığında VGGFace mimarisi %99.8 gibi zirve başarı oranlarını elde etmektedir. Resim dosyalarının Meta verilerine odaklanan bir önerme ise ELA aracılığıyla görselin her kaydedilişinde oluşan kalite farkına dayanır. Sinir ağı CASIA veri setiyle eğitilir ve %83 başarı oranı yakalar. Yeniden sıkıştırmaların kıyaslanmalarına odaklanan bir çalışmada ise robust derin öğrenmesi ve CNN mimarisini baz alıp, %92 başarı oranı sağlamaktadır ama minimum 128x128 çözünürlük gerektirmektedir. 4 farklı manipülasyon tarzını benimseyen bir başka önermede ise GAN mimarileri kullanılmaktadır. NIST MFC2018 veri setiyle renk odaklı çalışmada %70 başarı oranı sağlanırken, yüz değişimlerine bakıldığında %85'e yakın bir oran elde edilmiştir. Kontrast kaybına odaklanan bir çalışmada ise GAN mimarisi kullanılıp DenseNet aracılığıyla ağ yapıları oluşturulmuştur. Katmanlı aşamalardan sonra DeepFD uygulanıp çoğu yönteme kıyasla yüksek başarı oranları elde edilmiştir. Fakat eğitim örneklerinin toplanmasının zorluğu da bu yöntemi pratikte dezavantajlı hale getirmektedir.

Derin Sahte Tespit Yöntemi - Başarı Oranı Tablosu

Yöntem	Başarı Oranı % (AUC)	Yöntem	Başarı Oranı % (AUC)
VGG-16	%91.97	Alexnet(RELU)	%99.3
VGG-19	%92.09	Alexnet-TL(RELU)	%94
DenseNet169 + Rayleigh blur	%60.1	CNN(RELU)	%83.9
LRCN(Eye blanking)	%99	DenseNet121 & ResNet50	%97
CNN(Eye blanking)	%98	VGGFace	%99.8
EAR(Eye blanking)	%79	DenseNet201	%96

İlgili çalışmalara bakıldığında kullanım amacına, kullanılacak tekniğe, veri seti olanaklarına, mimari tercihlerine ve manipülasyon yöntemlerine göre her bir çalışmanın kendi avantaj ve dezavantajları bulunmaktadır. Bu da derin sahte analizi için tek bir en iyi metodun

olmadığını ve spesifik alanlara göre yöntemlerin tercih edilmesinin avantajlı sonuçlar oluşturacağını göstermektedir.

KAYNAKÇA

1. Qurat-ul-ain, Nida, N., Irtaza, A., & Ilyas, N. (2021). Forged Face Detection using ELA and Deep Learning Techniques. 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST). https://doi.org/10.1109/ibcast51254.2021.9393234
2. Maksutov, A. A., Morozov, V. O., Lavrenov, A. A., & Smirnov, A. S. (2020). Methods of Deepfake Detection Based on Machine Learning. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. https://doi.org/10.1109/eiconrus49466.2020.9039057
3. Yuezun Li, Ming-Ching Chang, & Siwei Lyu. (2018). In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *ArXiv: Computer Vision and Pattern Recognition*. http://export.arxiv.org/pdf/1806.02877
4. Tanaka, M., Shiota, S., & Kiya, H. (2021). A Detection Method of Operated Fake-Images Using Robust Hashing. *Journal of Imaging*, *7*(8), 134. https://doi.org/10.3390/jimaging7080134
5. AlShariah, N. M., & Khader, A. (2019). Detecting Fake Images on Social Media using Machine Learning. *International Journal of Advanced Computer Science and Applications*, *10*(12). https://doi.org/10.14569/ijacsa.2019.0101224
6. Shad, H. S., Rizvee, M. M., Roza, N. T., Hoq, S. M. A., Monirujjaman Khan, M., Singh, A., Zaguia, A., & Bourouis, S. (2021). Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network. *Computational Intelligence and Neuroscience*, *2021*, 1–18. https://doi.org/10.1155/2021/3111676
7. Villan, M., Kuruvilla, A., Paul, J., & Elias, E. (2017). Fake Image Detection Using Machine Learning. *IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS)*.
8. Ali, S. S., Ganapathi, I. I., Vu, N. S., Ali, S. D., Saxena, N., & Werghi, N. (2022). Image Forgery Detection Using Deep Learning by Recompressing Images. *Electronics*, *11*(3), 403. https://doi.org/10.3390/electronics11030403
9. Ahmed S Abdulreda, & Ahmed J. Obaid. (2022). A landscape view of deepfake techniques and detection methods. *International Journal of Nonlinear Analysis and Applications*, *13*(1), 745–755. https://doi.org/10.22075/ijnaa.2022.5580
10. Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, *10*(1), 370. https://doi.org/10.3390/app10010370