Araştırma Makalesi
Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi
8 (2021) 47-60

DERİN SAHTE VİDEOLARIN TESPİTİ VE UYGULAMALARI İÇİN BİR KARŞILAŞTIRMA ÇALIŞMASI

İsmail İLHAN[1*], Mehmet KARAKÖSE[2]

[1]Adıyaman Üniversitesi, Teknik Bilimler Meslek Yüksekokulu,Bilgisayar Teknolojileri Bölümü , Adıyaman, 02040, Türkiye
[2]Fırat Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Elazığ, 23000, Türkiye
Geliş tarihi: 29.03.2021 Kabul tarihi: 21.06.2021

ÖZET

Yapay zekânın birçok alanda kullanılması teknolojinin daha da ilerlemesini sağlamıştır. Bu alanlardan biri de derin sahte videoların oluşturulmasıdır. Derin sahte videolar için gerekli modeller yapay zekâ ile oluşturulmaktadır. Sosyal medyanın etkin bir şekilde kullanılması manipüle edilmiş videolara olan ilgiyi arttırmıştır. Derin sahte tespiti hala tam olarak çözülemeyen bir problem olduğu için Google, Youtube, Facebook, Microsoft, AWS ve AI gibi sosyal medya ve teknoloji geliştircileri, araştırmalara destek sağlamakta ve Kaggle ve Github gibi platformlarda önerilen çözümler açık kaynak olarak sunulmuştur. Derin sahte videoların oluşturulmasında ve tespit edilemsinde kullanılan yöntemler ve mimariler benzerdir. Ayrıca bu ikili mücadelede yeni önerilen yöntemleri kendilerini iyileştirmek için kullanmaktadırlar. Bu da her zaman için yeni bir tespit yöntemine ihtiyaç oluşturacaktır. Bu çalışmada derin sahte videoların tespit edilmesinde kullanılan yöntemler incelenmiştir. Uygulamaların performans etki analizleri yapılmıştır. Farklı özellikteki veri setleri, farklı yöntemlere sahip tespit uygulamaları ve özellikleri tablolar halinde verilmiştir. Uygulamalar karşılaştırılarak, zorlukları ve eğilimleri değerlendirilerek araştırmacılara kaynak olarak sunulmuştur.

Anahtar Kelimeler: *Derin Sahte, Manipülasyon, Sahte Video Tespiti, Yüz Değiştirme, Yüz Canlandırma*

A COMPARISON STUDY FOR THE DETECTION AND APPLICATIONS OF DEEPFAKE VIDEOS

ABSTRACT

The use of artificial intelligence in many areas has led to further advancement of technology. One of these areas is the creation of deep fake videos. The necessary models for deep fake videos are created with artificial intelligence. The effective use of social media has increased the interest in manipulated videos. As deepfake video detection is still an unresolved problem, social media such as Google, Youtube, Facebook, Microsoft, AWS and AI and technology developers provide support for research and proposed solutions are open sourced on platforms such as Kaggle and Github. The methods and architectures used to create and detect deep fake videos are similar. In addition, they use the newly proposed methods in this dual struggle to improve themselves. This will always create a need for a new detection method. In this study, the methods used to detect deep fake videos were examined. Performance impact analyzes of the applications were made. Data sets with different properties, detection applications with different methods and their properties are given in tables. The applications were compared, and their difficulties and tendencies were evaluated and presented to researchers as a resource.

Keywords: *Deepfake, manipulation, deepfake video detection, faceswap, face reenactment*

[*] e-posta1: iilhan@adiyaman.edu.tr ORCID ID https://orcid.org/0000-0002-5972-4295 (Sorumlu Yazar)
e-posta2: mkarakose@firat.edu.tr ORCID ID https://orcid.org/0000-0002-3276-3788

1. Giriş

Özellikle görüntü işleme ve yapay zekâ algoritmaların son zamanlarında ilerlemesiyle, sahte medya üretimi daha kolay hale gelmiştir [1]. Deepfakes ve FaceSwap gibi derin öğrenme tekniklerini kullanan uygulamalarda, videodaki kişinin yüzünü/sesini değiştirerek sahte videolar oluşturulmuştur. Deepfake (Derin Sahte) olarak da bilinen bu yapay zekâ ile sentezlenmiş yüz değiştirme videoları, çevrimiçi bilgilerin gerçekliği ve güvenirliği için büyük bir tehdit oluşturabilir ve dahası kötü amaçlar için kullanılabilir [2]. Bu gerçekçi sahte videoların tespiti için geleneksel adli tıp teknikleri yetersiz ve güçsüz kalmaktadır. Derin sahte uygulamalarıyla oluşturulan sahte haberlerin yayılması, Google, Facebook ve Microsoft gibi küresel teknoloji liderleri grubunu bu sorunu çözmeleri için bir araya getirdi. AI's Media Integrity Steering Committee, Oxford, Berkeley ve MIT dahil olmak üzere çeşitli üniversiteler ile sahte yapay zekâ videolarını tespit etme konusunda büyük bir yarışma düzenlenmiştir. Deepfake Detection Challenge olarak adlandırılan bu yarışma ve etkinlik Kaggle platformunda yer alır [3]. Derin sahte yöntemlerinin platformlarda paylaşılması ayrıca uygulamalarının çevrim içi olarak kullanılması kullanıcılar için erişilebilir ve kullanışlı olmasını sağlamıştır.

Mırsky ve ark. derin sahtelerin oluşturulmasını ve tespit edilmesini araştırmış ve mimarilerini ayrıntılı olarak sunmuşlardır. Derin sahte oluşturma yöntemlerini sınıflandırarak blok şemalarını göstermiş ve kısaca açıklamışlardır. Uygulanan oluşturma yöntemlerin karşılaştırmalarını tablolar halinde göstermişlerdir. Ayrıca tespit yöntemlerinden birkaçını tablo ile açıklamışlardır [3]. Nguyen ve ark. derin yüzeylerin algılanması ve derin sahtelerin tespit edilmesini incelemişlerdir. Derin sahte algoritmaların prensiplerini ve derin öğrenmenin nasıl kullanıldığını açıklamışlardır. Sahte resim tespit çalışmalarını kısaca anlatmış, sahte video tespiti için güncel yöntemlerde kullanılan sınıflandırmaları incelemişlerdir [4]. Rossler ve ark. FaceForensics ++ veri kümesi ile Deep-Fakes, Face2Face, FaceSwap ve NeuralTextures uygulamalarını ele alarak otomatik tespit yöntemlerini incelemiş ve sıkıştırılmış video dosyalarındaki performanslarını karşılaştırmışlardır. Geniş bir veri seti kullanarak kapsamlı bir analiz yapmışlarıdır [5]. Albahar ve ark. derin sahte tekniği, kökeni ve tarihsel bağlamı hakkında bilgi vermişlerdir. Ayrıca derin sahte videoların veya fotoğrafların nasıl oluşturulduğu ve değiştirilen fotoğraf ve videoların özellikleri göstermişlerdir ve yöntemlerin sistematik bir incelemesini sunmuşlardır [6]. Tolosana ve ark. yüz manipülasyon tekniklerini ve derin sahte yöntemleri kapsamlı bir incelemesini vermişlerdir. Öznitelik manipülasyonu, yüz sentezi, kimlik değişimi, ifade değişimi olan dört farklı yüz manipülasyon tekniğini ele alarak kullanılan veri tabanları ve uygulanan yöntemlerin doğruluk performanslarını değerlendirmişlerdir. Araştırılması gereken konuları ve eğilimleri sunmuşlardır [7]. Verdoliva son yıllarda önerilen en etkili manipülasyon yöntemlerinin kısa bir analizi ile geleneksel yaklaşımlar ve derin öğrenme yaklaşımları ile yüz manipülasyonlarının tespitini açıklamıştır. Konvansiyonel tespit yöntemlerini yani multimedya adli tıp alanındaki ana araştırma hatlarını kısaca açıklamışladır [8]. Samuele ve ark. DFDC veri setini kullanarak SHAP, GradCAM, LTPA ve Bonettini tekniklerini incelemişler ve karşılaştırmak için dört farklı iç ölçüm (varyans, çerçeve içi tutarlılık, çerçeveler arası tutarlılık ve merkezlilik) tanımlamışlar ve deneysel olarak karşılaştırmışlardır [9].

Bu çalışmada derin sahte videoların tespitinde kullanılan yöntemler sınıflandırılarak 3. Bölümde uygulamalardaki performansı etkileyen unsurlar başlıklar şeklinde açıklanmıştır. Bölüm 4'de ise derin sahte video tespit yöntemlerinin şekilleri verilerek açıklanmıştır ve yöntemlerin özellikleri ve doğruluk performansı Tablo 2'de verilmiştir. Tespit uygulamalarının zorlukları genel olarak verildikten sonra son bölümde genel bir değerlendirme yapılmıştır.

2. Derin Sahte Videoların Tespitinde Kullanılan Yöntemler

Geçtiğimiz birkaç ay içinde, derin sahte teknolojisinde çok daha yüksek bir gerçeklik seviyesi sunabilen veya neredeyse gerçek zamanlı olarak çalışabilen yeni algoritmalar görülmektedir. Derin

sahte videolarının en son biçimi, basit yüz değiştirmenin ötesine, tüm kafa sentezine (kafa kuklası), ortak görsel-işitsel sentezine (konuşan kafalar) ve hatta tüm vücut sentezine kadar gitmektedir [3, 4].

Derin sahte ilk anda fark etmesi çok zor olmasa da uzmanlar tarafından sahteliği tespit edilebilen bir teknoloji. Yinede, kendini geliştirmeye devam eden teknolojinin ileride sahte olduğu anlaşılamayacak kadar gerçek olmasından korkulmaktadır. Yapay zekânın kullanılmadığı sahte video tespit yöntemlerinde video özelliklerinin istatistiksel korelasyonu, istatistiksel anormalliklerin analizi ve cihaz tutarsızlıkları bilgileri kullanılmaktadır [3]. Yapay zekânın daha etkin kullanımı ile hem derin sahte üretmek hem de derin sahte tespit etmek daha başarılı ve daha kolay olmaktadır.

Tespit yöntemlerinde, temeli piksel seviyesine dayanan yöntemleri geliştiren araştırmacılar; artefaktlara, parmak izlerine, renk tutarsızlıklarına, doku bozulmalarına ve optik akış analizine ve hatta kameranın fiziksel özelliklerine özel önem verdiler. Biyolojik yöntemleri kullananlar, insan özelliklerini veya canlılık özelliklerini analiz eden görüntünün fiziksel / fizyolojik yönlerine dikkat etmektedir. Araştırmacılar, derin sahte videolarda baş pozlarında tutarsızlıklar olduğunu [10], doğal olmayan göz kırpmalarının [11], biyolojik sinyallerin korunmadığını ve bazı yüz çarpıklıklarının fark edilebildiğini keşfetmişlerdir [12]. Manipüle edilmiş görüntüleri eşzamanlı olarak algılamak ve manipüle edilmiş bölgeleri tahmin etmek için özellik haritaları işlenmiş ve geliştirmek için bir dikkat mekanizması kullanılmıştır. Ayrıca manipüle edilen bölgeler görselleştirilmiştir [13].

Oluşturma yöntemlerinden dolayı derin sahte videolarda, çerçeveler arası tutarsızlıklar ve kareler arasındaki zamansal tutarsızlıklar oluşur. Bu tutarsızlıklar tespit yöntemlerinde kullanılmaktadır. Tespit yöntemlerinde yüz artefaktları incelenebilir [5]. Bunlar; videodaki yüz bölgesinin diğer alandakilere göre daha bulanık olması, parlıyor olması, yüz kenarlarında cilt tonunda değişiklik olması; yüz bölgesinde çift çene, çift kaş, çift kenarın olması; el veya başka şeyler tarafından yüz bölgesi engellendiğinde titreme ve bulanıklaşma olmasıdır. Derin sahte videolarda en çok kullanılan yöntem yüz manipülasyonları yani yüz değiştirme yöntemleridir. Sabit boyutlarda görüntü işlemleri yapmak kaynak ve çeşitlilik için zordur. Bu nedenle görüntülerde yüzün pozlarına ve koordinatlarına uyması için yeniden ölçeklendirme, döndürme ve sentezleme işlemleri yapılmaktadır. Bu tür işlemlerde görüntülerde çözünürlük ve renk tonu gibi bazı çarpıtmalar oluşacaktır. Özellik çıkarım yöntemleri ile bu kusurlara bağlı manipülasyonlar tespit edilebilir.

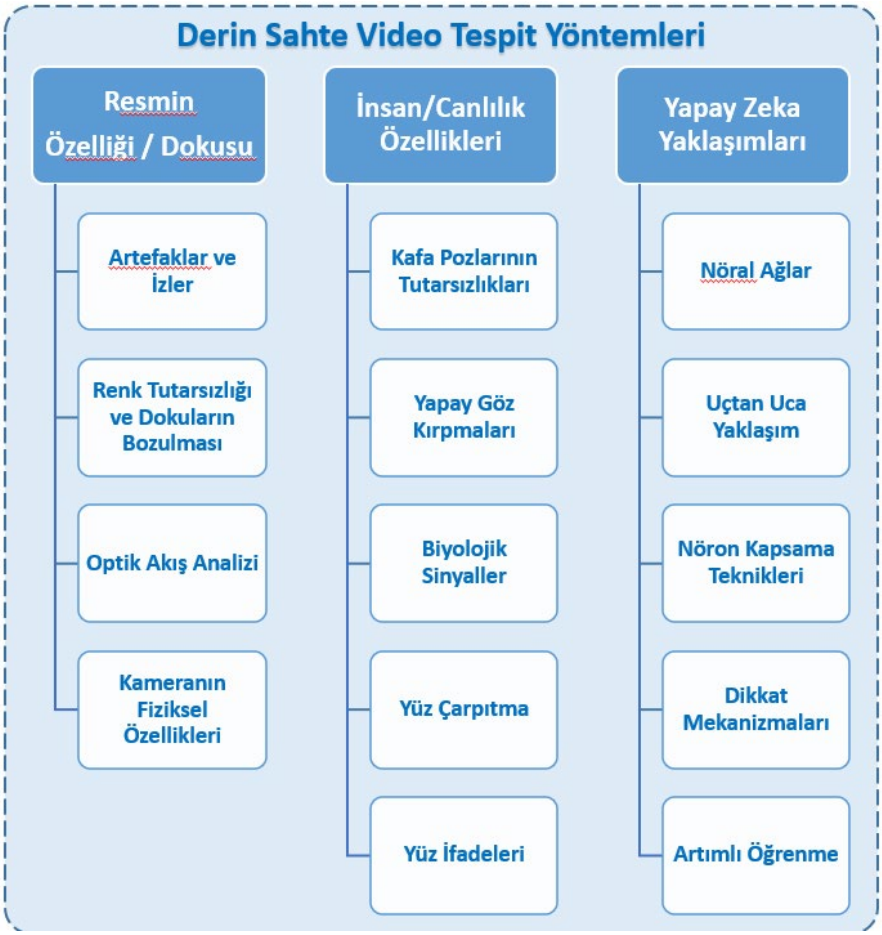
Yapay zekâ yöntemlerinde, el yapımı veya gözlemlenen özellikler hesaba katılmaz [12]. Yapay zekâ yaklaşımlarının çoğunun temelinde uygun bir sinir ağı kullanılır. Evrişimli Sinir Ağlarının bir araya getirilmesi, Evrişimli Sinir Ağlarının ve Tekrarlayan Sinir Ağlarının kombinasyonu, Uzamsal Evrişim Ağları, 3B Evrişimli Sinir Ağları veya MesoNet, kapsül ağları, uçtan uca yaklaşımlar, nöron kapsama teknikleri, dikkat mekanizmaları, artımlı öğrenme veya çok görevli öğrenme geliştirilen yöntemlerdir. Bahsedilen tüm tespit tekniklerinin karşılaştırılması oldukça zordur çünkü bu yöntemlerin yazarları farklı değerlendirme ölçütleri kullanmış ve ayrıca performansları farklı veri setlerinde ölçülmüştür [3, 4, 7].

Derin sahte tespit yöntemleri, sahte video oluşturma yöntemlerine benzemektedir. Çünkü tespit modülü derin sahte oluşturma yönteminin eğitim sürecinin bir parçası olarak kullanılmaktadır. Derin sahte tespit yöntemlerinde de farklı alanlarda birçok yöntemler tasarlanmıştır. Bu yöntemlerin özetleri Şekil 1'de gösterilmiştir.

Otomatik kodlayıcı kare kare kullanıldığından daha önce oluşturulmuş yüzden tamamen habersizdir. Bu da zamansal tutarsızlığı oluşturur. En göze çarpanı kısıtlı veriler ile eğitilen derin sahte oluşturuculardaki yüz bölgesinde oluşan aydınlatma tutarsızlığı nedeniyle titreyen bir fenomen oluşturmasıdır. Bu fenomenler CNN (Evrişimsel Sinir Ağları, Convolutional Neural Network) özellikli derin sahte detektörler ile tespit edilebilir [7, 12, 13].

Derin sahte çıktıları ile objektif yüz pozu arasındaki sistematik farklılıkların, dönüm noktası odakları kullanılarak ayırt edilebildiğini ve vektör makine modellerinin temel yardımı kullanılarak sınıflandırılabileceğini göstermiştir. Bu yöntemler ile kafa ve yüz pozlarındaki tutarsızlıklar yakalanarak manipülasyonlar tespit edilebilir [10].

İlişkili eylemlere sahip karmaşık videolarda zamansal etkileşimi öğrenmek için başka bir tasarıma ihtiyaç vardır. Dahası, gürültünün solması ve degradeleri nedeniyle uzun video sahnelerinin öğrenilmesinde tekrarlayan ağlar tarafından ciddi zorluklar yaşanmıştır. LSTM (Uzun-Kısa Süreli Bellek, Long Short-Term Memory) olarak bilinen bir tür RNN (Tekrarlayan Sinir Ağları, Recurrent Neural Network), bu sorunu çözmek için olağanüstü bir benzersiz ağdır [11].



Şekil 1. Derin sahte video tespit yöntemleri [14]

Çok akışlı öğrenme ağlarında akışların dengeli öğrenilmesi için, bir kayıp fonksiyonu kullanarak ve bölgesel yapay nesneleri öğrenerek farklı sıkıştırma yöntemleri uygulanmış videolarda manipülasyon tespitini yüksek performanslarda yapılmıştır [13].

3. Derin Sahte Video Tespiti Uygulamalarında Performans Etki Analizi

Derin sahte video tespit uygulamalarındaki doğruluk, hız, işlem gücü gibi performansa etki eden unsurlar aşağıda açıklanmıştır.

3.1. Eğitim Verisi

Derin sahte uygulamalarında hem oluşturma yöntemlerinde hem de tespit yöntemlerinde veri setlerine ihtiyaç duyulmaktadır. Özellikle derin sinir ağlarının eğitim süreçlerinde mimariye uygun büyük veri setleri kullanılmaktadır. Tablo 1'de gösterildiği gibi veri setlerindeki görüntü sayısı,

görüntülerin süresi, görüntülerdeki çerçeve sayısı, çözünürlüğü, görüntünün dosya formatı yani sıkıştırılma algoritması, uygulamanın eğitim performansında etkili olduğu gibi uygulamanın hız, doğruluk ve üretim kalitesini de doğrudan etkilemektedir. Ayrıca veri setinin tespit yöntemindeki odaklanan algoritmaya uygun olması gerekmektedir. Şekil 1'de gösterilen yöntemlerin odaklandığı algoritmalar ile eğitim sürecinde kullandıkları veri setlerinin oluşturulmasında kullanılan algoritmaların uyumlu olmaları gerekmektedir. Örneğin kafa pozlarının tutarsızlıklarını tespit eden bir uygulamada yapay göz kırpmaları için oluşturulmuş bir veri seti eğitim sürecinde kullanılmamalıdır. Veri setlerinin etkisini uygulamaların doğruluk performanslarında da açıkça görülmektedir. Eğitim setinde kullanılan veri setinin test doğruluk oranı başka bir veri seti ile karşılaştırıldığında farklı olduğu Tablo 2'de görülmektedir.

Tablo 1. Derin sahte tespit yöntemlerinde kullanılan veri setleri ve özellikleri

Veri Seti	Gerçek		Sahte		Tarih	Çözünürlük	Yöntem
	Video	Çerçeve	Video	Çerçeve			
UADFV	49	17,3k	49	17.3k	2018		FakeAPP
DeepFake-TIMIT-LQ (DF-TIMIT)	320	34,0k	320	34.0k	2018	64	faceswap-GAN
DeepFake-TIMIT-HQ	320	34,0k	320	34.0k	2018	128	faceswap-GAN
FaceForensics++(FF-DF)	1,000	509,9k	1,000	509.9k	2019	480, 720, 1080	Deepfakes, Face2Face, FaceSwap,NeuralTextures
Google/Jigsaw DeepFake (DFD)	363	315.4k	3,068	2,242.7k	2019	480, 720, 1080	DeepFake maker
Facebook DeepFake detection challenge Dataset (DFDC)	1,131	488.4k	4,113	1,783.3k	2019	480, 720, 1080	2 farklı faceswap algoritması
Celeb-DF	590	225.4k	5,639	2,116.8k	2019	256	DeepFake synthesis algorithm

Tablo 2. Öne çıkan derin sahte tespit yöntemlerin mimarisi ve performansı

Yöntemler	Sınıflandırma/ Teknikler	İçerik			Kaynak Türü			Uygulanan Veri Seti / ACC – EER - AUC					
		Özellik	Yüz	Görüntü	Görüntü	Video	Ses	FF-DF	UADFV	Celeb-DF	DF-TIMIT	DFD	DFDC	Diğer
Göz kırpması [11]	LRCN (LSTM+CNN)	●	●			●								0.99
Uzay-zamansal özellikleri kullanma [15]	RCN (CNN+RNN)	●		●		●		96,9						
Çerçeve içi ve zamansal tutarsızlıklar [16]	CNN+RNN	●		●										97,1
Yüz çarpıtma yapılarını kullanma [12]	CNN	●			●	●		93,0	97,7	64,6	99,9	93,0	75,5	
MesoNet [17]	CNN			●		●		95,23	82,1	53,6	87,8			
Kapsül adli tıp [18]	Capsule CNN		●		●	●		99,33						
Baş pozları [10]	SVM	●	●		●	●		47,3	89,0	54,6	53,2		55,9	
Yüz manipülasyonu [13]	CNN	●	●		●	●								100-0,1

Multi-task öğrenme [19]	CNN+DE		●		●	●				36,5	62,2			
Ses tutarsızlığı [20]	LSTM					●	●							24,74
Ses tutarsızlığı [21]	LSTM+DNN	●				●	●							17,6
Ses özellikleri [22]	DNN	●				●	●							1,26
Video ve Ses özellikleri [23]	LSTM	●		●		●	●	100		99,6				
XceptionNet [24]	CNN							99,26		38,7	56,7			

Tablo 2. Devamı

3.2. Görüntü Kalitesi ve Çözünürlüğü

Tespit uygulamalarının performansı düşük ve yüksek kalitedeki veri setlerinde ayrı ayrı değerlendirilmektedir. Uygulamalarda özellikle resmin özelliği ve dokusunu kullanan yöntemlere göre görüntünün çözünürlüğü bir kısıt olarak ortaya çıkmaktadır. Tablo 1'de veri setlerinin kullandıkları çözünürlük örnekleri verilmiştir. Birçok uygulamada ön işlemlerde kullanılan yüz tanıma ve çıkarma işlemlerinde görüntü kalitesinin ve çözünürlüğünün performansa etkisi görülmektedir. Ayrıca videolardaki çerçeve sayısı mekânsal ve zamansal tutarsızlıkların performansını da etkilemektedir. Aynı şekilde videoların sıkıştırılma formatları da videodaki çerçeve özelliklerini sıkıştırdığından dolayı performansa etki eden bir diğer unsurdur. Görüntülerin fiziksel özelliklerinin yanı sıra çekim ortamları ve portre çekimleri de uygulamalar için önemli etkenlerdir. Işık yansımaları, aydınlatma ve gölge oluşumu görüntü özellikleri için ayırt edici bir etkendir. Tespit uygulamaların bazıları mimarilerinde bu ayırt edici özellikleri kullanmışlardır. Veri setlerindeki portre çekim videolarında uzaklık ve bakış açısı performans etkilerinden biridir. Ayrıca portre çekimlerinde kullanılan gözlük, pirsin, şapka, atkı gibi aksesuarlar, yapılan makyajlar, saç, sakal, bıyık, kaş gibi ayırt edici özellikler veri setini şekillendiren diğer etkenlerdir.

3.3. Ön İşlemler

Ön işlemler uygulamaların giriş katmanı olarak değerlendirilir. Uygulamaya giriş olarak alınan kaynak verinin uygun formata dönüştürülmesi gerekmektedir. Uygulamaların genelinin resim ile çalıştıklarından, kaynak olarak verilen video çerçevelere yani resme dönüştürülür. Resim üzerinde uygulanacak işlemler için resim boyutlandırma ve tonlama, kontrast gibi filtreler uygulanır. Uygulamanın gerekli olan çalışma bölgesi için resim kırpılır. Çoğu uygulamada kullanılan yüz bölgesi için bu işlem yüz tespiti ve yüz çıkarımı olarak yapılır. Farklı uygulamalar için bunlar kafa, göz veya ağız bölgeleri olabilir. Bölge seçimi tamamen uygulamadaki algoritmaya bağlıdır. Çerçeve seçimi, bölge seçimi, bölge çıkarımı ve sonrası için uygulanan görüntü filtreleri, iki boyutlu veya üç boyutlu dönüşümler için seçilen yöntemler hız ve doğruluk performanslarına etki eden unsurlardır. İşleme alınan görüntüler ön işlemler ve bir sonraki katmanda uygulanacak adımlar için boyutlandırma ve döndürme işlemleri uygulanmaktadır. Yapılan bu tür işlemler doğruluk performansı için doğru, hız performansı için ters orantılı olabilmektedir veya tam tersi de olabilir. Bu tamamen sonraki katmanlarda yapılan işlemlere bağlıdır. Performans etkisini iyileştirmek için ön işlemlerde kullanılan algoritmada doğru ve hızlı yöntemlerin seçilmesi gereklidir.

3.4. Yapay Sinir Mimarisi

Ön işlemlerden sonra gelen bu orta katman, birkaç katmanın birleşiminden oluşabilir. Bu katmanda resimler bir vektör olarak işlenir. Tasarlanan yapay sinir ağı ile vektörlerdeki mekânsal ve

zamansal tutarsızlıklar tespit edilmeye çalışılır. Kullanılan sinir ağı modülleri, sinir ağındaki katman sayısı, kayıp fonksiyonu ve döngü sayısı performansa etki eden unsurlardır. Bazı uygulamalarda sinir ağı ile özellik çıkarımı veya çerçeveler arasındaki zamansal tutarsızlıklar ele alınırken bazı uygulamalar her ikisini de kullanmaktadır [23, 16, 15]. Evrişimli sinir ağlarındaki vektör işlemlerinde kullanılan yöntemler sinir ağının performansını etkilemektedir. Yapay sinir mimarisinde kullanılan CNN, RNN, LSTM, RCN, LCRN, DNN, Capsule CNN ve AE gibi yapıların her birine özgü efektif özelliklerinin bulunması nedeniyle uygulamalarda birçok karma mimarilerin oluşmasını sağlamıştır. Bunun sonucunda da uygulamaların kendine özgü özelliklerinin ön plana çıkmasına sebep olmuştur. Karşılaştırmalara bakıldığında bazı uygulamaların daha doğru, bazı uygulamaların daha hızlı, bazı uygularının ise genelinin dışında alanına göre daha iyi olduğu görülür. Odaklandığı çözüm yöntemine göre bir mimari oluşturur ve o alanda daha başarılı bir performans gösterir. Bu durum hastalık türüne göre verilen tedaviye benzemektedir.

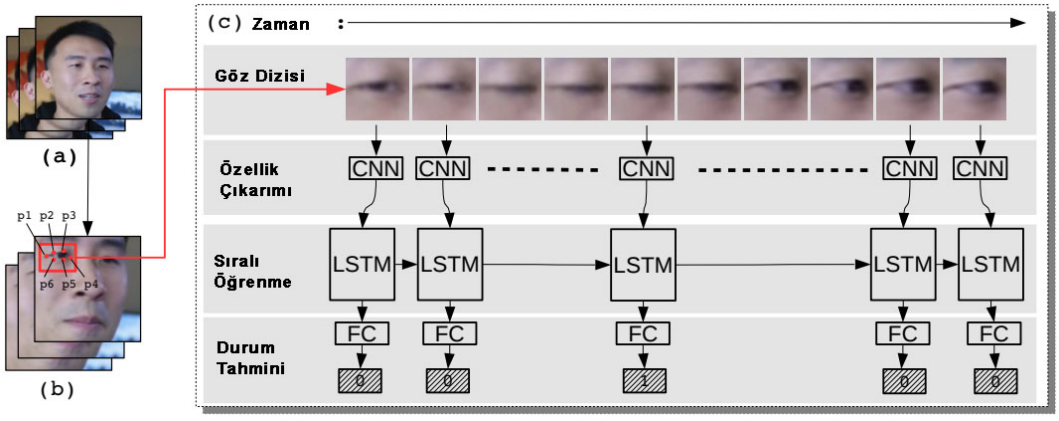
3.5. Sınıflandırma yöntemi

Tespit uygulamalarının mimarisinde yer alan son katmandır, son işlem katmanı olarak da adlandırılır. Bu katmanda tek satır/sütun veri vektörleri kullanılarak sınıflandırma işlemi yapılır. Algoritmalardan elde edilen özellik vektörünün sahte veya gerçek olduğu etiketlenir. Sınıflandırma yönteminde genellikle sinir ağları veya sınıflandırma algoritmaları yer alır. Sınıflandırma ağında Softmax, Gated Recurrent Unit (Kapılı Tekrarlayan Birim, GRU), CNN, RNN, Capsule gibi sinir ağları kullanılır. Ayrıca Logistic Regression, SVM (Destek Vektör Makineleri,Support Vector Machine) gibi meta sezgisel yöntemler de kullanılıyor.

4. Derin Sahte Video Tespiti Uygulamalarının Karşılaştırılması

Tablo 2' de verilen derin sahte tespit yöntemlerinde çoğunlukla yüz içeriklerinin kullanıldığı görülmektedir [3]. Özellik çıkarımı, resim ve beden içeriklerini kullanan uygulamaların sayısı daha az ve veri setlerine bakıldığında doğrulukları düşüktür. Uygulamalar klasik makine öğreniminde SVM modelini, derin öğrenme teknolojisinde ise CNN modelini çoğunlukla kullanmışlardır.

Göz kırpması özelliği ile derin sahte tespiti yapılmaktadır. Kaynak videodan yüz tespiti yapılarak her bir çerçeve için göz bölgesi çıkarılır. CNN ağı ile çıkarılan özellikler LSTM ile zamansal olarak göz kırpma frekansı işlenir. Uygun olmayan göz kırpmalar bu şekilde tespit edilir. Tasarlanan model LRCN (Uzun Süreli Tekrarlayan Evrişimli Sinir Ağları, Long-term Recurrent Convolutional Neural Networks) olarak adlandırılır ve yapısı Şekil 2' de gösterilmiştir [11].

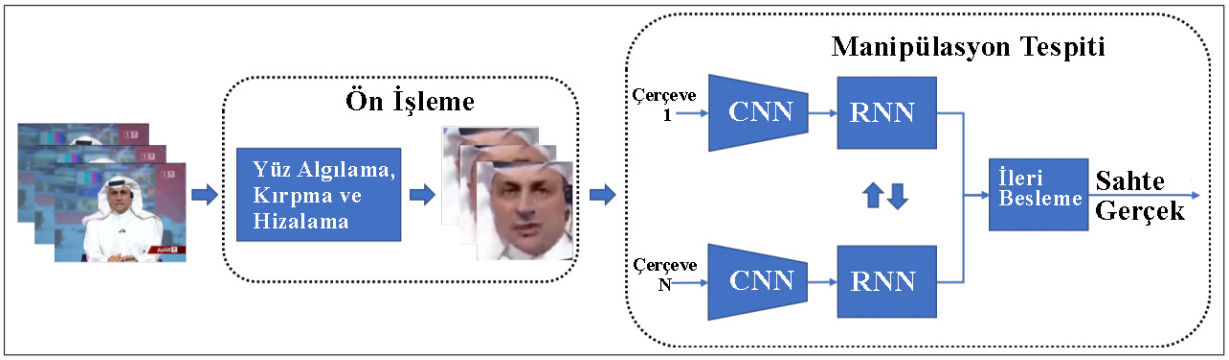


Şekil 2. Uzun süreli tekrarlayan evrişimli sinir ağları yöntemi [11]

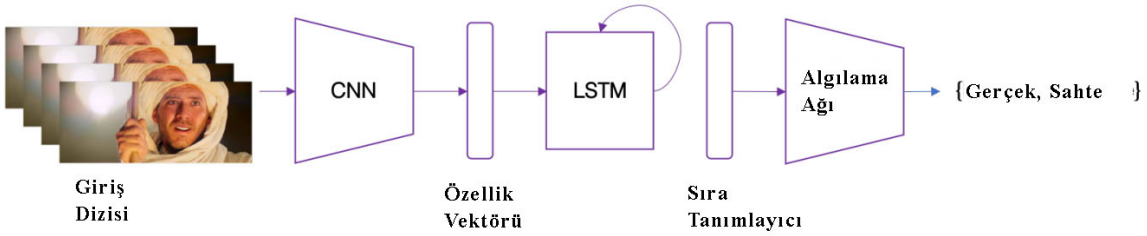
Tekrarlayan evrişimli sinir ağları (Recurrent Convolutional Neural Networks, RCN) yöntemi ile videolardaki sahte yüz tespit Şekil 3'de gösterilmiştir. Ön işleme aşamasında videodaki çerçevelerden yüz tespiti yapılarak yüz bölgesi kırpılır ve hizalanır. İkinci aşamada ön işlem yüz bölgesinin manipülasyon tespiti yapılır.

Belirli bir çerçeve sayısı için CNN ile özellik çıkarımı yapılır ve çerçevelerin birbirine zamansal olarak tutarlı olup olmadığı RNN ile tespit edilir [15].

Çerçeve içi ve zamansal tutarsızlıklar yönteminde ise CNN ve LSTM kullanılmıştır. Şekil 4'da gösterilen yöntemde her çerçevede CNN tarafından oluşturulan bir özellik vektörü elde edilir. Daha sonra, birden fazla ardışık çerçevenin özelliklerini birleştirip analiz için LSTM' ye aktarılır. Eğitim sırasında, LSTM modeli 2048 boyutlu ImageNet özellik vektörlerinin bir dizisini alır. LSTM'yi, 512 boyutlu tam bağlı katman izler. Son olarak da sahte video oluşup oluşmadığı tahmini için özellik vektörü softmax katmanına gönderilir [16].

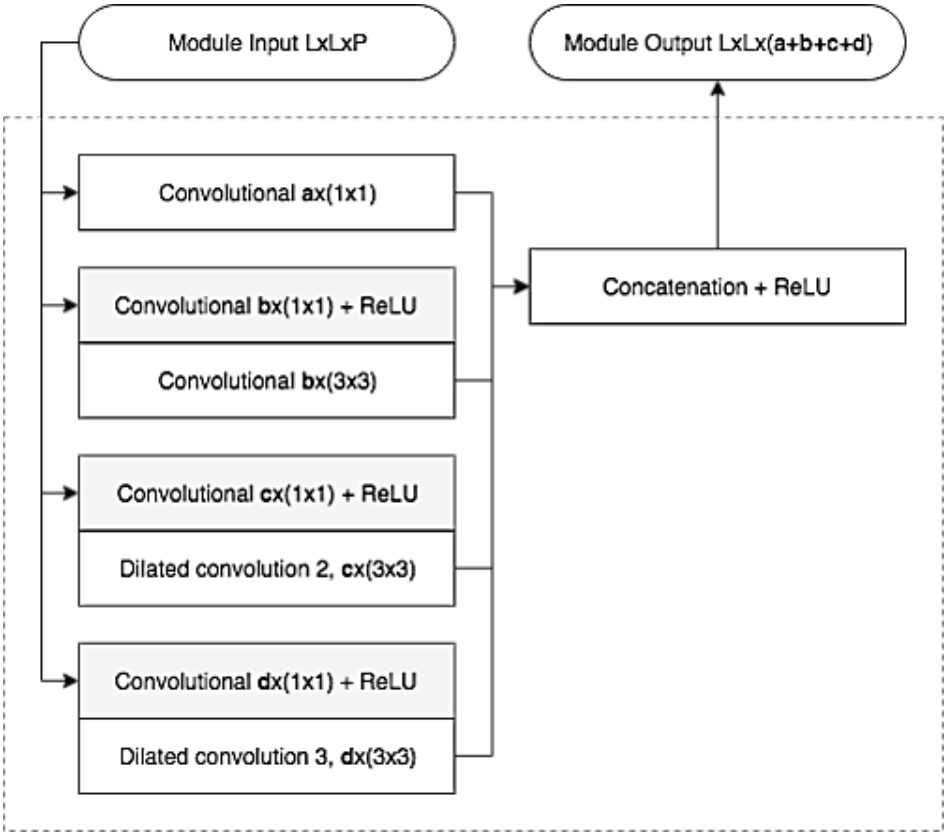


Şekil 3. RCN ile videodaki yüz manipülasyon tespiti [15]



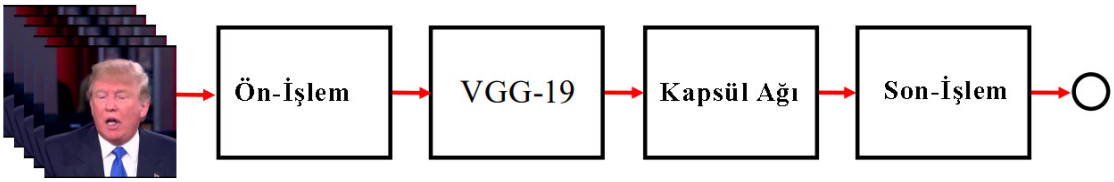
Şekil 4. Çerçeve içi ve zamansal tutarsızlıklar yöntemi [16]

Görüntü artefaktların tespiti için mikroskobik analizler yapılabilir. Fakat görüntüdeki artefaktları güçlü bir şekilde azaltan sıkıştırma algoritmaları kullanılmaktadır. Videolarda uygulanan farklı sıkıştırma formatları birçok derin sahte tespit yöntemini olumsuz yönde etkilemektedir. MaSonet yöntemi, az sayıda katmana sahip derin bir sinir ağı kullanarak orta düzey bir yaklaşım sunan ve Şekil 5'te gösterilen mimarisi ile sıkıştırılmış videolarda önemli bir başarı sağlamıştır [17].



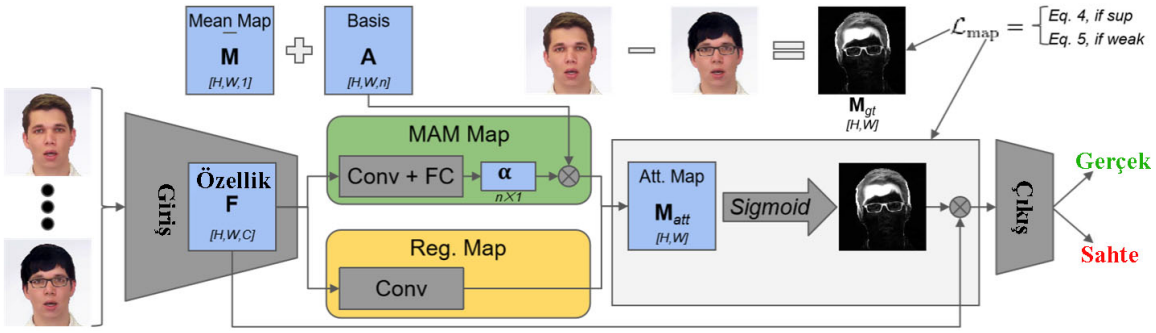
Şekil 5. MesoInception-4 yönteminin mimari yapısı [17]

Kapsül yöntemi ile yapılan uygulamalarda, dinamik yönlendirme algoritması tarafından hesaplanan kapsüller arasındaki anlaşma ile, nesne parçaları arasındaki hiyerarşik poz ilişkilerinin iyi tanımlanabileceğini göstermiştir. Dinamik yönlendirme algoritması kullanılarak elde edilen kapsüller arasındaki anlaşma, karmaşık ve neredeyse kusursuz olan sahte görüntüler ve videolar üzerinde algılama performansını artırabilir. Kapsül yöntemi Şekil 6'de gösterilmiştir. Giriş için çerçeveler kullanılır ve önişlemde yüz algılama yapılır. VGG-19 ağında kapsül ağı için gizli özellikler çıkarılır. Kapsül ağı ise 3 farklı kapsülden oluşmaktadır. Nihai sonucu elde etmek için işlem sonrası aşamasında, olasılıkların ortalaması alınır [18].



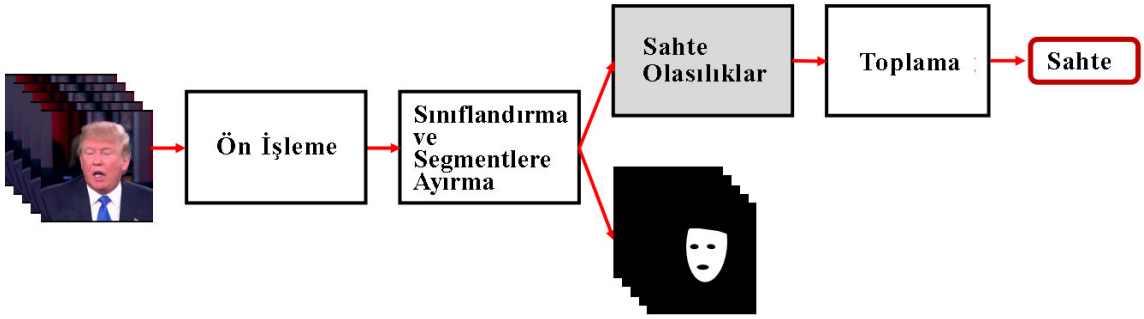
Şekil 6. Capsule CNN yöntemi genel gösterimi [18]

Şekil 7' de gösterilen mimaride, manipüle edilmiş yüz algılamayı CNN tabanlı bir ağ kullanarak ikili bir sınıflandırma problemi uygulanmıştır. Ayrıca, sınıflandırıcı modelin özellik haritalarını işlemek için dikkat mekanizmasını (attention mechanism) kullanılmıştır. Öğrenilen dikkat haritaları, bir görüntüdeki CNN'in kararını etkileyen bölgeleri vurgulayabilir ve ayrıca CNN'in daha ayırt edici özellikleri keşfetmesine rehberlik etmek için kullanılabilir [13].



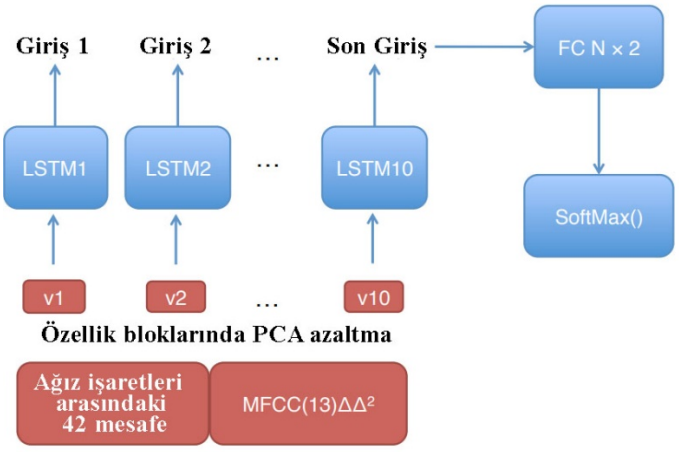
Şekil 7: Yüz manipülasyon algılama mimarisi [13]

CNN ve linear unit (ReLU) ağını kullanarak Y-shaped Autoencoder yöntemini oluşturan uygulama Şekil 8'de gösterilmiştir. Uygulama farklı boyutlardaki resim ve videodaki kaynaklardan yüz bölgesinde bölümleme ve sınıflandırma işlemi yapmaktadır [19].



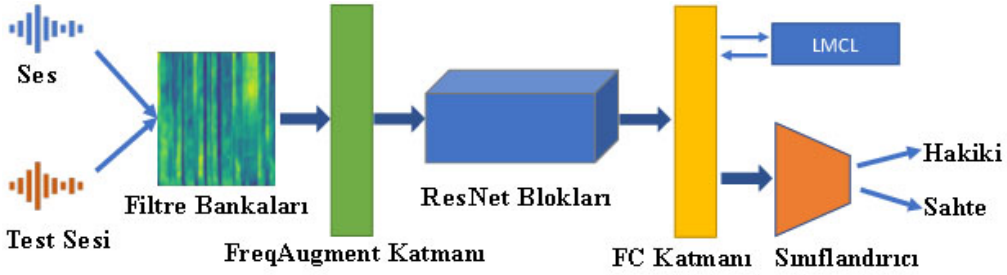
Şekil 8: CNN+DE modülleri ile Multi-task öğrenme yöntemi [19]

Dublaj ve dudak senkronizasyonu kullanılarak görsel-işitsel tutarsızlık tespiti yapılabilir. Yöntem Şekil 9'de gösterilmiştr. Ses için genellikle Mel-Scale Frequency Cepstral Coefficients (MFCC) kullanılırken, derin sinir ağları (DNN'ler) ile optik akıştan öğrenilen özelliklere kadar değişen farklı görsel özellikler yüzün ağız bölgesinden çıkarılır. Çıkarılan özellikler daha sonra, bir sınıflandırıcıya gönderilmeden önce, uzun kısa süreli bellek veya evrişimli sinir ağları gibi en iyi performans gösteren örnekler olarak bir miktar işlemden geçer. Algılama sistemi için, MFFC'leri ses özellikleri olarak ve ağızdaki önemli noktalar arasındaki mesafeleri görsel özellikler olarak bu yöntemde kullanıldı [20].



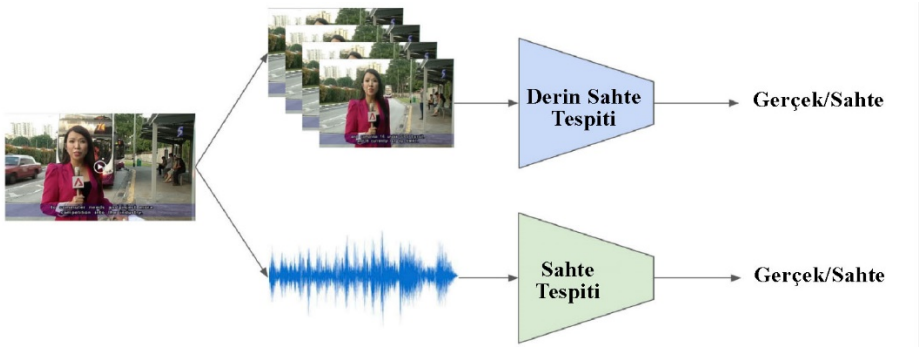
Şekil 9: Ses tutarsızlığı ile manipüle video tespit yöntemi [20]

Videolardaki sahte ses tespiti için Şekil 10' de önerilen yöntemde öğrenmeyi daha etkin kılmak için LMCL (Büyük Marj Kosinüs Kaybı Fonksiyonu, Large Margin Cosine Loss Function) ve çevrimiçi frekans maskeleme artırma işlemlerini önermiştir. DNN modelinin genelleme yeteneğini daha da artırmak için DNN eğitimi sırasında bitişik frekans kanallarını rastgele maskeleyen bir katman olan FreqAugment kullanılmıştır [22].



Şekil 10: Sahte ses tespit yöntemi [22]

Sahte seslerin oluşturulmasında iki ana yöntem olan Mantıksal Erişim (Logical Access:LA) ve Fiziksel Erişim (Physical Access: PA) vadır. LA durumunda, sahte ses, bir metinden konuşmaya veya bir ses dönüştürme yazılımı kullanılarak üretilir. Metinden konuşmaya sentez yazılımı ile, bir metin girişi, hedef konuşmacınınkine benzer bir sese sahip bir konuşma çıkışına dönüştürülebilir. Bir ses dönüştürme yazılımı, bir kaynak konuşmacı tarafından verilen bir konuşmanın, aynı dil içeriğine (yani metin bilgisi), ancak farklı bir konuşmacı kimliğine sahip olan farklı bir söyleyişe dönüştürülmesine izin verir. PA yönteminde güvenlik sistemini hedef konuşmacının gerçekten konuştuğuna inanması için hedef konuşmacının önceden kaydedilmiş bir konuşması yeniden oynatılır. Sahte seslerin tespiti için derin sinir ağları ile birlikte Constant-Q Cepstrum Coefficients (CQCC), MFCCs ve Gaussian Mixture Models (GMMs) gibi konuşma özellikleri kullanılmıştır. Şekil 11'daki yapıda hem ses hem de video tespiti için etkin yöntemler kullanılarak veri setlerinde test edilmiştir. XceptionNet, ConvLSTM ve FaceNetLSTM yapılarını kullanarak farklı kayıp fonksiyonları ile derin sahte tespitini yüksek doğruluklarda gerçekleştirmişlerdir [23].



Şekil 11: Derin sahte video ve ses algılama için tekrarlayan evrişimli yapılar [23]

4.1. Deeepfake Video Tespit Zorlukları

Ses ve video veri dosyalarını saklanması, taşınması ve işlenmesi önemli ölçüde teknolojik gereksinime ihtiyaç duyar. Bunu gidermek için çeşitli sıkıştırma işlemleri uygulanır. Sıkıştırma işlemleri ile verilerde birtakım kalite kayıpları olur bu da veri analizlerini olumsuz etkiler. Özellikle adli vakalardaki video analizlerinde ve sahte video tespitinde kalitedeki kayıplar önemli oranda etkilemektedir. Sahteciliğin tespit edilememesi için videoların sıkıştırma işlemi uygulanmaktadır. Mpeg

ve H.264 gibi sıkıştırma standartları uygulanarak videolardaki çerçeveler arasındaki fark bilgileri kullanılarak boyutu azaltılır.

Sahteciliğin fark edilemediği düşük kalitedeki videolar oluşturulmaktadır. Çoğu tespit yöntemlerinde düşük kalitedeki videolarda verim düşmektedir. Tespit yöntemlerinin hem düşük kalitede hem de yüksek kalitedeki videolarda yüksek ve kabul edilebilir doğrulukta sonuç vermeleri gerekir.

Orijinal videoları doğrulamak için gizli filigranlar ve blok zinciri teknolojileri kullanılabilir fakat bu teknikler internetteki paylaşımların önüne geçemiyor. Sahte haberlerde gördüğümüz gibi, bir içerik parçasının kolayca çürütülebilir olması, tıklanmayacağı ve çevrimiçi olarak okunmayacağı ve paylaşılmayacağı anlamına gelmez. Bu nedenle denetimlerin gerçek zamanlı olması veya paylaşımların yapılmadan önce denetilmesi gerekiyor.

Derin sahte tespit yöntemleri, derin sahte yapmak için kullanılan yöntemlere yeterince benzerdir ve tespit için tasarlanan modeller, sahte olanları iyileştirmek için kullanılabilir. Bu da derin sahte oluşturma modellerini daha da başarılı hale getirecektir. Anlaşılan budur ki ikisi arasındaki çekişme hiç bitmeyecek.

Organizatörlerin ve kuruluşların araştırmacılar için oluşturdukları derin sahte video veri setindeki performans ile gerçek veri setindeki performans arasında önemli bir tutarsızlık vardır. Yüksek doğruluğu olan modeller gerçek veri setinde test edildiğinde doğruluğu yüzde 65'e kadar düşmektedir. Bu nedenle paylaşılan veri setlerinin daha da geliştirilerek gerçek derin sahte videolara benzemesi gerekiyor ki tespit yöntemleri de daha iyi performans versin.

Derin sahte teknoloji ile oluşturulan videolar paylaşıma sunulmadan önce manuel olarak videodaki hatalar düzeltilmektedir. Bu tür sonradan yapılan düzeltmeler tespit oranını düşürür hatta tespiti boşa çıkarabilir.

Derin sahte eğitim ve uygulama süreci yüksek donanımlara rağmen uzun zaman sürebilmektedir. Veri setinin kapasitesine ve kalitesine bağlı olarak uzun süren eğitim süreci uygulamaların en büyük zorluklarından biridir. Sürenin kısaltılması için yüksek kapasite ve hızda GPU ve CPU donanımlar kullanılmaktadır bu da pahalıya mal olmaktadır.

Derin sahte video oluşturmanın kısıtlarından biri de kaynak verilerde kişiye ait görüntüleri aynı veya benzer olmasıdır. Benzer yüz ve ten tonları kullanılmaktadır. Kaynak görüntülerdeki açı çeşitliliğinin olması gereklidir.

5. Sonuçlar

Manipüle edilmiş içeriklerin oluşturulması hızla gelişen bir sorundur ve bu içerikleri tanımlamak teknik olarak zordur. Bu nedenle daha iyi algılama araçları oluşturmak gereklidir. Derin sahte için geliştirilen yapay zekâ algoritmaları derin sahte tespit için geliştirilen algoritmalardan bir kaç adım daha önde ilerliyor. Derin sahte oluşturma teknolojileri hızla geliştikleri ve kullanımların gittikçe kolaylaştığı için ayrıca sosyal medyada hızla paylaşılarak yayıldığı için kritik bir sorun oluşturmaktadır. Buna karşın çözüm araştırmalarının daha sağlam, ölçeklenebilir ve genelleştirilebilir yöntemler sunmaya odaklanması gerekmektedir. Mevcut tespit yöntemlerinin kısıtlarının azaltılarak daha da genelleştirilmesi gerekmektedir. Derin sahteler ile yapılan mücadelede resmi kurumların da gerekli önlemleri almaları ve gerekli hukuksal düzenlemeleri yapmaları gerekmektedir. Derin sahtelerin toplum üzerindeki etkileri ve olası kaosları önlemek için farkındalık eğitimleri verilmelidir. Bu mücadelede büyük bir paya sahip olan sosyal medya platformlarının da sistemlerini güncelleyerek ve yeni tespit sistemleri ile daha da geliştirerek çevrimiçi olarak müdahale etmeleri gerekmektedir.

Yeni bir konu ve güncel bir sorun olan derin sahte teknolojisi derin öğrenme teknolojisinin ve algoritmaların yeni modelleri ile gelişmeye devam etmektedir. Teknolojinin doğal bir sorunu olarak ortaya çıkan bu sorun güncelliğini her zaman koruyacaktır.

Gelecek çalışmalarda derin sahte videoların test edilmesi için çevrim içi bir web servis uygulaması yapılması planlanmaktadır. Araştırmacıların ve veri bilimcilerin ücretsiz olarak faydalandığı Google firmasının sunduğu Colaboratory (ya da kısaca "Colab"), TÜBİTAK ULAKBİM Yüksek Başarımlı ve Grid Hesaplama Merkezi olan Truba olarak adlandırılan sunucular ve üniversitelerin yüksek başarımlı hesaplama merkezlerinin sunucularının kullanılması planlanmaktadır. Python dili ile yazılacak olan bu tür uygulamaların test edilmesi için kullanılan cihazların 24 çekirdekli işlemcilerden daha fazla işlemciye, 32 GB'dan daha kapasiteli ön belleğe ve GPU özellikli ekran kartlarına (Nvidia P100 GPU) ihtiyaç vardır.

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] M. E. Berk, Dijital Çağın Yeni Tehlikesi "Deepfake", OPUS Uluslararası Toplum Araştırmaları Dergisi, no. 16 (28), pp. 1508-1523, 2020.

[3] Y. Mirsky ve W. Lee, The creation and detection of deepfakes: A survey, ACM Computing Surveys (CSUR), cilt 54, p. 1–41, 2021.

[4] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen ve S. Nahavandi, Deep learning for deepfakes creation and detection, arXiv preprint arXiv:1909.11573, cilt 1, 2019.

[5] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies ve M. Nießner, Faceforensics++: Learning to detect manipulated facial images,Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[6] M. Albahar ve J. Almalki, Deepfakes: Threats and countermeasures systematic review, Journal of Theoretical and Applied Information Technology, cilt 97, p. 3242–3250, 2019.

[7] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales ve J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion, cilt 64, p. 131–148, 2020.

[8] L. Verdoliva, "Media forensics and deepfakes: an overview.", IEEE Journal of Selected Topics in Signal Processing 14(5), pp. 910-932, 2020.

[9] S. Pino, M. J. Carman ve P. Bestagini, «What's wrong with this video? Comparing Explainers for Deepfake Detection,» arXiv preprint arXiv:2105.05902, 2021.

[10] X. Yang, Y. Li ve S. Lyu, Exposing deep fakes using inconsistent head poses, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[11] Y. Li, M.-C. Chang ve S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, 2018 IEEE International Workshop on Information Forensics and Security (WIFS).IEEE, pp. 1-7, 2018.

[12] Y. Li ve S. Lyu, Exposing deepfake videos by detecting face warping artifacts, arXiv preprint arXiv:1811.00656, 2018.

[13] H. Dang, F. Liu, J. Stehouwer, X. Liu ve A. K. Jain, On the detection of digital face manipulation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[14] J. Baciak, M. Żurawska, T. Czech ve B. Górny, Deepfake video detection using the ensemble of neural networks, comarch, 2020. Available: https://www.researchgate.net/publication/344554345_Deepfake_video_detection_using_the_ensemble_of_neural_networks.

[15] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi ve P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI), cilt 3, 2019.

[16] D. Güera ve E. J. Delp, Deepfake video detection using recurrent neural networks, 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).IEEE, pp. 1-6, 2018.

[17] D. Afchar, V. Nozick, J. Yamagishi ve I. Echizen, Mesonet: a compact facial video forgery detection network, 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[18] H. H. Nguyen, J. Yamagishi ve I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).IEEE, pp. 2307--2311, 2019.

[19] H. H. Nguyen, F. Fang, J. Yamagishi ve I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos, 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2019.

[20] P. Korshunov ve S. Marcel, Speaker inconsistency detection in tampered video, 2018 26th European Signal Processing Conference (EUSIPCO), 2018.

[21] P. Korshunov, M. Halstead, D. Castan, M. Graciarena, M. McLaren, B. Burns, A. Lawson ve S. Marcel, Tampered speaker inconsistency detection with phonetically aware audio-visual features, International Conference on Machine Learning, 2019.

[22] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman ve E. Khoury, Generalization of audio deepfake detection, Proceedings of the Odyssey Speaker and Language Recognition Workshop, Tokyo, Japan, 2020.

[23] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright ve R. Ptucha, Recurrent convolutional structures for audio spoof and video deepfake detection, IEEE Journal of Selected Topics in Signal Processing, cilt 14, p. 1024–1037, 2020.

[24] F. Chollet, Xception: Deep learning with depthwise separable convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251--1258, 2017.