
GAZİ ÜNİVERSİTESİ
Gazi
1926

ResNeXt50 VE LSTM MİMARİLERİYLE DERİN SAHTE TESPİTİ

Leili ESLAM JAMALGOLZARI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

EYLÜL 2025

Leili ESLAM JAMALGOLZARI tarafından hazırlanan “RESNEXT50 VE LSTM MİMARİLERİYLE DERİN SAHTE TESPİTİ” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Aydın ÇETİN

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ...……

Başkan: Prof. Dr. Aysun ÇOŞKUN

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Doç. Dr. Cevat RAHEBİ

Yazılım Mühendisliği Ana Bilim Dalı, İstanbul Topkapı Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 26/09/2025

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……

Prof. Dr. Uğur GÖKMEN

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Leili ESLAM JAMALGOLZARI

26/09/2025

ResNeXt50 VE LSTM MİMARİLERİYLE DERİN SAHTE TESPİTİ

(Yüksek Lisans Tezi)

Leili ESLAM JAMALGOLZARI

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Eylül 2025

ÖZET

Derin sahte (deepfake) teknolojileri giderek daha sofistike hale gelmekte ve dijital güvenlik, mahremiyet ile yanlış bilgilendirme alanlarında kritik zorluklar oluşturmaktadır. Derin sahte tespiti, hem mekânsal hem de zamansal tutarsızlıkları yakalayabilen derin öğrenme modellerini gerektirmektedir. Bu tezde, üç farklı derin öğrenme mimarisinin—ResNeXt50, LSTM ve önerilen Hibrit LSTM/ResNeXt50 mimarisinin—DeepFake Detection Challenge (DFDC) veri kümesi üzerinde karşılaştırmalı analizi sunulmaktadır. Deneysel sonuçlar, saf ResNeXt50 modelinin %44 kesinlik ve %44 F_1-skoru elde ettiğini, LSTM modelinin ise %62 kesinlik ve %70 F_1-skoru ile daha iyi performans sergilediğini, ancak her iki modelin de tek başına uygulandığında sınırlı kaldığını göstermektedir. Buna karşın, Hibrit LSTM/ResNeXt50 mimarisi %91,85 kesinlik, %90,68 geri duyarlılık, %85,26 doğruluk ve %91 F_1-skoru ile bu modelleri önemli ölçüde geride bırakmıştır. Bulgular, LSTM mimarilerinin zamansal bağımlılıkları ResNeXt50'ye göre daha etkili yakaladığı, ancak mekânsal ve zamansal özellik çıkarımının bütünleştirilmesiyle çok daha dengeli ve sağlam bir tespit performansına ulaşıldığını, hibrit mimarinin daha güvenilir derin sahte tespiti potansiyeli olduğunu ortaya koymaktadır.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Derin sahte, derin sahte tespiti, lstm, resnext50, hibrit
Sayfa Adedi	:	72
Danışman	:	Prof. Dr. Aydın ÇETİN

DEEPFAKE DETECTION WITH ResNeXt50 AND LSTM ARCHITECTURES

(M.Sc. Thesis)

Leili ESLAM JAMALGOLZARI

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

September 2025

ABSTRACT

Deepfake technologies have become increasingly sophisticated, posing critical challenges to digital security, privacy, and the spread of misinformation. Detecting such manipulations requires deep learning models capable of capturing both spatial and temporal inconsistencies. This thesis presents a comparative study of three deep learning architectures—ResNeXt50, LSTM, and a proposed Hybrid LSTM/ResNeXt50—evaluated on the DeepFake Detection Challenge (DFDC) dataset. Experimental results show that the pure ResNeXt50 model achieves a Precision and F_1-score of 44%, while the LSTM model performs better with 62% Precision and 70% F_1-score, yet both remain limited when applied independently. In contrast, the Hybrid LSTM/ResNeXt50 architecture significantly outperforms these baselines, achieving 91.85% Precision, 90.68% Recall, 85.26% Accuracy, and an F_1-score of 91%. These findings demonstrate that while LSTMs capture temporal dependencies more effectively than ResNeXt50, integrating spatial and temporal feature extraction yields a more balanced and robust detection performance, underscoring the potential of hybrid architectures for reliable deepfake detection.

Science Code	:	92432
Key Words	:	Deep learning, deepfake, deepfake detection, lstm, resnext50, hybrid
Page Number	:	72
Supervisor	:	Prof. Dr. Aydın ÇETİN

TEŞEKKÜR

Gazi Üniversitesi'nde yüksek lisans eğitimim süresince bana yol gösteren, teşvik eden ve destek olan Sayın Prof. Dr. Aydın ÇETİN'e en derin şükranlarımı sunmak isterim. Kendisi, bu tezin tamamlanmasında olduğu kadar, araştırmacı olarak kişisel gelişimimde de büyük rol oynamıştır. Değerli geri bildirimleri, akademik rehberliği ve sürekli motivasyonu sayesinde bu süreci başarıyla tamamlayabildim. Sağladığı fırsatlar ve onun danışmanlığında kazandığım bilgi ve deneyim için minnettarım.

Ayrıca, araştırmamı yürütmem için gerekli akademik ortamı ve kaynakları sağlayan Gazi Üniversitesi'ne ve özellikle Bilgisayar Mühendisliği Bölümü'ne içten teşekkürlerimi sunarım.

Son olarak, bu yolculuk boyunca en büyük güç kaynağım olan, sabrı, anlayışı ve sürekli desteğiyle yanımda olan sevgili eşim Tohid HARİGHİ'ye derin minnettarlığımı ifade etmek isterim.

İÇİNDEKİLER

Sayfa

ÖZET iv

ABSTRACT v

TEŞEKKÜR vi

İÇİNDEKİLER vii

ÇİZELGELERİN LİSTESİ ix

ŞEKİLLERİN LİSTESİ x

SİMGELER VE KISALTMALAR xi

1. GİRİŞ 1

2. DERİN SAHTE VE ETKİ DEĞERLENDİRMESİ 7

2.1. Derin sahte 7

2.1.1. Derin sahte tespiti 10

2.1.2. Derin sahte tespiti zorlukları 12

2.2. Derin sahte'lerin Sosyolojik Etkisi 13

2.2.1. Derin sahte anlamada kuramsal yaklaşımlar 13

2.2.2. Medya ve politika üzerindeki etkileri 14

2.2.3. Demografi ve güç dinamikleri 15

2.2.4. Sosyal bilim ve deneysel uygulamalar 16

2.2.5. Güven aşınması ve epistemik krizler 17

2.2.6. Toplumsal tepkiler ve uyum 18

2.2.7. Derin sahte paylaşımını etkileyen faktörler 19

3. MATERYAL VE METOT 23

3.1. Derin Sahte Veri Kümeleri 23

3.2. Üretken Çekişmeli Ağlar 26

Sayfa

3.2. Üretken Çekişmeli Ağlar 26
3.3. Evrişimli Sinir Ağı 29
3.4. RESNET 30
3.5. ResNeXt50 30
3.6. Uzun Kısa Süreli Bellek 33
3.7. Hibrit Model 35
4. UYGULAMA 39
5. BULGULAR VE DEĞERLENDİRME 45
5.1. Değerlendirme Ölçütleri 45
5.2. Deneysel Bulgular 47
6. SONUÇ 59
KAYNAKLAR 61
EKLER 67
EK-1. DFDC doğruluk veri kümeleri 68
EK-2. DFDC kayıp veri kümeleri 69
EK-3. DFDC eğitim veri kümeleri 70
ÖZGEÇMİŞ 71

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 2.1. Derin sahte tespit yöntemlerinin karşılaştırılması 12

Çizelge 3.1. Derin sahte çalışmalarında kullanılan veri kümeleri 25

Çizelge 5.1. Farklı modellerin performans değerleri 56

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 2.1 Derin sahte videolarının yıllar içindeki artışı 8

Şekil 2.2. Derin sahte olaylarının tiplerine göre dağılım 15

Şekil 2.3. Derin sahte kullanımını etkileyen faktörler 16

Şekil 2.4. Derin sahtelerin sosyolojik etkileri 18

Şekil 2.5. Derin sahte paylaşımını etkileyen faktörler 19

Şekil 3.1. GAN blok diyagramı 26

Şekil 3.2. Evrişimli sinir ağı 29

Şekil 3.3. ResNeXt 30

Şekil 3.4. ResNeXt50 yapısı 32

Şekil 3.5. ResNet50 ve ResNeXt50 blokları 33

Şekil 3.6. LSTM diyagramı 34

Şekil 4.1. Tez süreci akış diyagramı 39

Şekil 4.2. LSTM modeli eğitim sürecinden bir kesit 42

Şekil 4.3. ResNeXt50 modelin eğitim sürecinden bir kesit 43

Şekil 4.4. Hibrit model eğitim sürecinden bir kesit 43

Şekil 5.1. LSTM modeli (a) Doğruluk, (b) Kayıp grafiği 47

Şekil 5.2. ResNeXt50 Modeli (a) Doğruluk, (b) Kayıp grafiği 49

Şekil 5.3. Hybrid (ResNeXt50/LSTM) Modeli (a) Doğruluk, (b) Kayıp grafiği 51

Şekil 5.4. Hybrid (ResNeXt50/LSTM) Modelinin 30 ekoka kadar (a) Doğruluk, (b)... Kayıp grafiği 53

Şekil 5.5. LSTM ve ResNeXt50 ve Hibrit modellerin performans karşılaştırması 54

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
3D	Three Dimensional (Üç boyutlu)
AI	Artificial Intelligence (Yapay Zekâ)
CNN	Convolutional Neural Networks (Evrişimli Sinir Ağı)
Conv2D	Two-Dimensional Convolution Layer (İki Boyutlu Evrişim Katmanı)
CPU	Central Processing Unit (Merkezi İşlem Birimi)
DFDC	Deep Fake Detection Challenge
FF	FaceForensics
FF++	FaceForensics++
FN	False Negative (Yanlış Negatif)
FP	False Positive (Yanlış Pozitif)
GAN	Generative Adversarial Network (Çekişmeli Üretken Ağı)
GPU	Graphics Processing Unit (Grafik İşlemci Birimi)
Log	Logaritma
LSTM	Long Short Term Memories (Uzun Kısa Süreli Bellek)
Max	Maximum
Min	Minimum
NLP	Natural Language Processing (Doğal Dil İşleme)
OpenCV	Open Computer Vision Library (Açık Kaynak Bilgisayarlı Görü Kütüphanesi)
ResNet	Residual Network (Artık Ağlar)
ResNeXt	Residual Networks with Aggregated Transformations
RNN	Recurrent Neural Networks (Yinelenen Sinir Ağı)
TN	True Negative (Doğru Negatif)
Kısaltmalar	**Açıklamalar**

TP	True Positive (Doğru Pozitif)
TPU	Tensor Processing Unit (Tensör İşlem Birimi)
VGG	Visual Geometry Group

1. GİRİŞ

"Derin sahte" terimi, "derin öğrenme (deep learning)" ve "sahte (fake)" kelimelerinin birleşiminden oluşmaktadır. Derin sahteler, özellikle Çekişmeli Üretken Ağlar (GANs) ve otokodlayıcılar (autoencoders) gibi derin öğrenme modelleriyle oluşturulan, gerçekmiş gibi görünen sahte medya içerikleridir (video, görüntü ve ses). Bu içerikler gerçek ya da sahte olduklarını ayırt etmeyi güçleştirecek kadar gerçekçi şekilde üretilebilmektedir.

Tarihsel olarak bakıldığında, fotoğraf manipülasyonu 19. yüzyıldan bu yana var olup, daha sonraları diğer medya türlerine de uygulanmıştır. Medya manipülasyonunun kökeni ise 1990'lı yıllara, Adobe Photoshop'ın görüntüleri değiştirmek amacıyla kullanılmasına dayanmaktadır. Bu bağlamda, yapay zekânın görüntü işleme alanındaki uygulamaları ile birlikte 1990'larda derin sahte teknolojisinin akademik araştırma boyutu başlamıştır [1].

2014 yılında Çekişmeli Üretken Ağlar (GAN) tanıtılmıştır [2]. Bu gelişme, günümüzde kullanılan derin sahte'lerin başlangıcı olarak kabul edilebilir. "Derin sahte" terimi ilk olarak 2017 yılında bir Reddit kullanıcısı tarafından ortaya atılmıştır. Bu kullanıcı, topluluk üyeleriyle derin sahte tabanlı içerikler paylaşmak üzere bir topluluk kurmuştur. Bu içerikler, ünlülerin yüzlerinin genellikle pornografik içeriklerdeki oyuncuların vücutlarına yerleştirilmesi veya oyuncu Nicolas Cage'in yüzünün farklı filmlerdeki sahnelere yerleştirilmesi gibi örnekleri kapsamaktadır. Yapay zekâ kullanılarak hem gerçekçi görünen hem de tamamen sentetik içerikler üretilebilmesi, bu tür içeriklerin büyük ilgi görmesine neden olmuştur. 2018 yılında, kullanıcıların yüz değiştirme işlemi yapmasına olanak tanıyan ilk masaüstü uygulama olan "FakeApp" piyasaya sürülmüş ve derin sahte resmi olarak sektör haline gelmiştir. Derin sahte'ler, yapay zekâ algoritmaları kullanılarak oluşturulan ve gerçek kişilerin seslerini, yüzlerini ve hareketlerini taklit edebilen içeriklerdir. Bu içerikler, görsel ve işitsel örüntüleri öğrenip yeniden üretebilen veri temelli modeller tarafından oluşturulmaktadır.

Problem tanımı

Derin sahte üretimi esas olarak Çekişmeli Üretken Ağlar (GANs) teknolojisindeki gelişmelerden türemektedir. İlk dönemlerinde yalnızca yüz değiştirme işlemleri üzerine

odaklanan derin sahte teknolojileri zamanla daha da gelişerek gerçeğinden ayırt edilmesi oldukça güç olan hiper-gerçekçi videoların üretilmesini mümkün kılmıştır. Bunun gerçek ve sahte görüntüler arasındaki sınırın belirlenmesini son derece güçleştirecek bir olgunluk düzeyine ulaşmış olup, gözle bu ayrımı yapabilmek neredeyse imkânsız hale gelmiştir.

Derin Sahte Yönelimleri (Deepfake Trends 2024) raporuna göre, derin sahte video içeriklerinde ciddi bir artış yaşanmıştır. 2024 yılı itibarıyla dünya genelinde derin sahte içeriklerinde %900'lük bir artış söz konusudur ve bu içeriklerin %45'i rıza dışı müstehcen materyallerdir [3].

Gerçek dünyadaki bazı derin sahte örnekleri arasında, eski ABD başkanı Barak Obama'nın sahte bir açıklama yaptığı ve viral hale gelen bir derin sahte videosu yer almaktadır. Bu video, derin sahte içeriklerinin siyasette kötü amaçlarla kullanılabileceğini açıkça göstermiştir. Bir diğer örnekte ise, 2019 yılında Star Wars serisinden bir oyuncu, derin sahte teknolojisi kullanılarak dijital olarak yeniden yaratılmış ve böylece sahte ile gerçek eğlence içeriği arasındaki ayrım neredeyse tamamen ortadan kalkmıştır. Ayrıca TikTok ve Instagram gibi sosyal medya uygulamalarında, yalnızca 2023 yılı içerisinde 1 milyondan fazla yapay zekâ tabanlı videonun yüklenmesiyle, yüz değiştirme (face swap) içerikleri büyük ölçüde yayılmış ve viral olmuştur.

MIT Media Lab'in 2023 yılı raporuna göre, internet kullanıcılarının %68'inden fazlası en az bir derin sahte videosuyla karşılaşmıştır. Ayrıca YouTube, 2021–2024 arasında derin sahte içerik yüklemelerinde %450'lik bir artış olduğunu ve bu içeriklerin çoğunun siyasetle ilişkili olduğunu bildirmiştir.

Derin sahte'lerin yayılması, mahremiyetin ihlalinden siyasi manipülasyona kadar uzanan geniş etik sonuçlara yol açmaktadır. Sensity AI'nin 2023 raporuna göre, derin sahte içeriklerin %95'inden fazlası müstehcen içerik barındırmakta olup, bu içeriklerde en çok ünlüler ve politikacılar hedef alınmaktadır. Gelişmiş tespit yöntemlerinin geliştirilmesi sayesinde, araştırmacılar yapay zeka destekli araçların kötüye kullanımını engellemeyi, medya bütünlüğünü korumayı ve kişisel mahremiyeti savunmayı ummaktadır. Bu nedenlerle, derin sahte tespiti çalışılması gereken önemli bir araştırma alanıdır.

Derin sahte içeriklerin tespitinde karşılaşılan bazı zorluklar şunlardır: GAN'ler kullanılarak elde edilen gerçekçilik düzeyi, geleneksel tespit yöntemlerinin bu içerikleri tanımasını engellemektedir. Ayrıca, derin sahte içeriklerde görülebilen bazı saptanabilir kusurlar; gölgelerin tutarsızlığı, doğallıktan uzak göz hareketleri gibi bağlamsal uyumsuzluklar olabilir. Bir diğer önemli zorluk ise adversaryal saldırılardır; bu tür içerik üretkenleri, tespit algoritmalarını yanıltmak için veriler üzerinde kasıtlı değişiklikler yaparak tespiti daha da karmaşık hale getirir.

Çözüm yöntemleri

Derin sahte tespit yöntemleri genel olarak dört ana gruba ayrılabilir:
1. Derin öğrenme tabanlı yöntemler. Bu grupta genellikle CNN, RNN, LSTM gibi teknikler kullanılır. Bu yöntemler yüksek doğruluk oranı sağlar ve karmaşık desenleri öğrenebilir. Ancak tek dezavantajları, büyük veri kümelerine ihtiyaç duymaları ve hesaplama açısından maliyetli olmalarıdır.

2. Klasik makine öğrenmesi tabanlı yöntemler. Bu grupta el ile çıkarılmış öznitelikler ve istatistiksel yöntemler kullanılır. Bu grubun en dikkat çekici özelliği, hesaplama açısından maliyetli olmamalarıdır; ancak doğruluk oranları düşüktür ve yeni manipülasyon teknikleriyle başa çıkmakta zorlanırlar.

3. Üçüncü grup, hibrit/birleşim tabanlı yöntemlerdir. Bu yöntemlerin temel tekniği çok modlu (multimodal) füzyon kullanmaları ve adversaryal (karşıt) eğitim uygulamalarıdır. Bu yaklaşımlar, birden fazla tespit ipucunu aynı anda kullanabilme avantajına sahiptir; ancak model tasarımı ve eğitimi açısından karmaşıklık barındırırlar.

4. Dördüncü grup, blok zincir tabanlı yöntemlerdir. Bu yöntemlerde kullanılan temel teknikler, medya kimlik doğrulama ve köken (provenance) takibidir. Blokzincir tabanlı yöntemler, medya üzerindeki değişikliklerin izlenmesinde güvenilir çözümler sunar; ancak uygulama maliyetleri yüksektir. Ayrıca, bu yaklaşımlarda doğrudan bir tespit yöntemi bulunmamaktadır.

Çalışmanın amacı

Yapay zekâ ile oluşturulan içeriklerin kötüye kullanımının önlenmesi amacıyla sağlam tespit tekniklerinin geliştirilmesi hayati önem taşımaktadır. Bu, yanlış bilgilendirme yayılımının önüne geçmek, kimlik hırsızlığını engellemek ve bireyleri rızaları dışında oluşturulan içeriklerden korumak açısından kritik bir rol oynar. Derin sahte tespit sistemleri; kolluk kuvvetleri, sosyal medya platformları, haber doğrulama sistemleri ve siber güvenlik altyapıları gibi çeşitli alanlarda kullanılabilir.

Bu çalışmanın amacı, derin sahte tespiti konusundaki literatüre katkı sağlamak üzere CNN ve LSTM mimarilerinin başarımlarının, etkinliklerinin karşılaştırılması ve iyileştirilmesir. CNN'ler, görüntülerdeki ve karelerdeki mekansal tutarsızlıkları, örneğin piksel düzeyinde anormallikler, düzensiz dokular ve yüz artefaktları gibi yapıları tanıma konusunda başarılıdır. Öte yandan, LSTM ağları ise zamanla ilgili tutarsızlıkları yakalama açısından etkili olup, video içeriklerinde doğal olmayan geçişler ya da senkronizasyon hataları gibi ardışık anormallikleri analiz etmede etkilidir.

Motivasyon

Derin sahte içeriklerin ortaya çıkışı kişisel mahremiyet, medya bütünlüğü ve toplumsal güven açısından önemli etkiler yaratmaktadır. Derin sahtekârlıkların etkilerinin tespit edilmesi ve azaltılması, dijital içeriklerin özgünlüğünün korunması açısından oldukça önemlidir. Bu bağlamda, yapay zekâ destekli tespit yaklaşımlarının geliştirilmesi, hem bireylerin mahremiyetini korumak hem de medya güvenilirliğini sürdürmek adına kritik bir ihtiyaç haline gelmiştir.

CNN tabanlı ResNeXt50 ve RNN tabanı LSTM mimarilerinin uygulanmasıyla, bu araştırma, mekânsal ve zamansal verileri kullanan kapsamlı ve karşılaştırmalı bir tespit çerçevesi geliştirmeyi, aynı zamanda tespit doğruluğunu ve sağlamlığını artırmayı hedeflemektedir. Bu çerçevelerin gerçek dünya veri kümelerine, özellikle DFDC veri kümesine uygulanması, yapay zekâ tarafından oluşturulan içerikleri çeşitli medya formatlarında tanımlama etkinliğini doğrulayacaktır.

Geliştirilmiş tespit çerçeveleri, kötü amaçlı içeriklerin belirlenmesine ve ortadan kaldırılmasına yardımcı olabilir, medya bütünlüğünü koruyabilir ve kişisel gizliliği güvence altına alabilir.

CNN tabanlı ve LSTM mimarileriyle yapılan derin sahte tespitiyle ilgili olarak belirtilebilecek temel sınırlamalardan biri, hesaplama kaynaklarının sınırlı oluşudur. Derin sahte tespiti için gereken hesaplama araçları son derece pahalı olabilir ve herkes bu tür araçlara sahip olamayabilir. Bu bağlamda bir diğer sorun, derin sahte tespit modellerinin farklı veri kümelerinde genelleme yapmakta zorlanmasıdır; çünkü derin sahtelerin görsel ve bağlamsal özellikleri, verinin kaynağına, ışık koşullarına ve içerik türüne göre değişiklik gösterebilir. Ayrıca, derin sahte içerik üretkenleri, tespiti daha zor hale getirmek için adversaryal teknikler uygulayarak piksel düzeyinde veriyi değiştirebilir veya gürültü ekleyebilir. Veri kümeleriyle ilgili bir başka sınırlama, sağlanan veri kümesinin tüm manipülasyon tekniklerini veya tespit doğruluğunu etkileyebilecek demografik çeşitliliği kapsamıyor olabilmesidir. Bunların yanı sıra, özellikle GAN ve otokodlayıcı mimarilerindeki teknolojik ilerlemelerin hızı nedeniyle, her geçen gün daha gerçekçi derin sahte üretim teknikleri ortaya çıkmakta, dolayısıyla bu alanda sürekli güncel kalmak büyük önem arz etmektedir.

Tezin organizasyonu

Bu tez beş bölümden oluşmaktadır. Birinci bölümde, derin sahte (deepfake) ve derin sahte tespiti konularına giriş yapılmıştır. İkinci bölümde, derin sahte olgusu, derin sahte tespit yöntemleri, kullanılan veri kümeleri, bu alandaki zorluklar ve derin sahtenin sosyolojik etkilerine ilişkin literatür taraması sunulmuştur.

Üçüncü bölümde, araştırmada uygulanan yöntem ve araçlar Materyal ve Metot başlığı altında ele alnmıştır. Bu bölümde kullanılan mimariler ayrı ayrı tanıtılmış, yapıları ve matematiksel temelleri tartışılmıştır.

Dördüncü bölümde, modelin uygulanışı detaylandırılmıştır. Bu kapsamda araştırmanın akış şeması verilmiş, izlenen adımlar (veri kümesi seçimi, programlama dili ve kullanılan kütüphaneler gibi) ayrıntılı şekilde açıklanmıştır.

Beşinci bölümde, elde edilen sonuçlar sunulmuş, değerlendirme yöntemleri açıklanmış ve nihai olarak sonuçlar karşılaştırılmıştır.

2. DERİN SAHTE VE ETKİ DEĞERLENDİRMESİ

Derin sahte (deepfake) teknolojisi, ortaya çıktığı ilk dönemden günümüze kadar hızlı bir gelişim göstermiş ve toplumsal etkileri giderek artan bir konu haline gelmiştir. Başlangıçta yalnızca akademik çevrelerde deneysel çalışmalarla sınırlı olan bu teknoloji, zamanla daha erişilebilir hale gelmiş ve farklı kullanıcı kitleleri tarafından yaygın olarak kullanılmaya başlanmıştır. Bu süreçte, derin sahte içeriklerin yalnızca teknik bir yenilik olarak değil, aynı zamanda sosyal medya, siyaset, ekonomi ve hukuk alanlarında önemli sonuçlar doğuran çok boyutlu bir olgu olduğu görülmektedir. Literatürde yapılan çalışmalar, derin sahte teknolojisinin gelişim evrelerini, toplumsal etkilerini ve buna karşı geliştirilen tespit yöntemlerini ayrıntılı biçimde ele almaktadır.

2.1. Derin Sahte

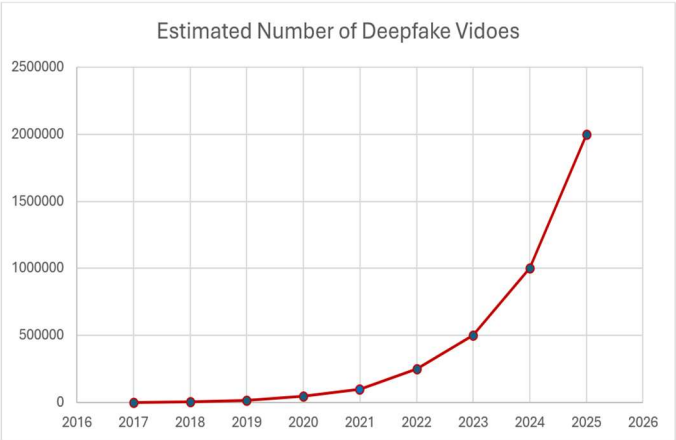
Derin sahte içerikleri, ortaya çıktığı zamandan bu yana yıllar içerisinde katlanarak artmıştır. Yapay zekâ, derin öğrenme ve Çekişmeli Üretken Aglar (GANs) teknolojilerindeki gelişmelerle birlikte farklı evrelerden geçmiştir.

i. 2017 öncesi aşamada, erken gelişim aşamasında, 2014'te GAN'ler tanıtılmış ve gerçekçi görüntü ve videoların üretimini kolaylaştırmıştır. Derin sahte teknolojisi o dönemde yüz değiştirme ve stil transferi ile ilgiliydi. Derin sahte genelde akademik çevrelerde kullanılıyor, araştırmacılar otokodlayıcı gibi yeni derin öğrenme modelleri ile denemeler yapılmıştır.
ii. Derin sahte içeriklerinin ikinci aşaması, 2017 ile 2019 arasındaki popülerlik dönemidir. Bu dönemde uzman olmayan kişiler kendi cihazlarında derin sahte uygulamaları kullanarak derin sahte içerikler üretmeye başlamışlardır. Aynı zamanda bu dönemde ilk derin sahte tespit modeli tanıtılmıştır.
iii. Üçüncü aşama, 2020 ile 2022 arasındaki derin sahte içerik patlaması dönemidir. Bu dönemde daha güçlü yapay zekâ modelleri tanıtılmış ve bu da derin sahte içeriklerinin sayısında büyük artışa neden olmuştur. Sonuç olarak, siyasi yanlış bilgilendirme ve finansal dolandırıcılıklar da artış göstermiştir. Hükümetler ve sosyal platformlar derin sahte tehdidini ciddiye almaya başlamış, bazı düzenlemeler ve yasaklamalar getirmiştir.

iv. Dördüncü aşama, 2023'ten 2025'e kadar olan gelişmiş yapay zekâ ve gerçekçilik dönemidir. Bu dönemde DALLE, Stable Diffusion ve Make-A-Video gibi modeller sayesinde daha karmaşık derin sahte içerikleri üretilmiştir. Derin sahte içeriklerinin kalitesi artmış, metinden videoya dönüşüm ve derin sahte uygulamalarının gelişimi ile siber suçlar ve dolandırıcılıklar da yükselmiştir. Bu yüzden, buna karşı önlem olarak daha fazla derin sahte tespit yöntemi geliştirilmiştir.

v. Son dönem ise derin sahtenin gelecekteki eğilimleridir. Görüldüğü üzere, 2025 yılı itibarıyla derin sahte videolarının sayısı yaklaşık 2 milyon seviyesine ulaşarak önceki yıllara kıyasla üstel bir artış eğilimi göstermektedir. Bu bulgu, derin sahte teknolojilerinin erişilebilirliğinin hızla arttığını ve toplumsal etkilerinin giderek daha kritik bir boyuta taşındığını ortaya koymaktadır.

Şekil 2.1'de, tarafımızca hazırlanmış olan grafik, 2017 yılından itibaren derin sahte ile üretilen videoların yıllara göre tahmini artışını göstermektedir. Grafikte yalnızca mevcut durum değil, aynı zamanda geleceğe yönelik eğilim de ortaya konmaktadır. Yapılan değerlendirmelere göre, ilerleyen yıllarda derin sahte videolarının sayısının üstel bir biçimde artacağı öngörülmektedir. Bu artış yalnızca miktar açısından değil, içerik kalitesinin giderek daha gerçekçi hale gelmesi bakımından da dikkat çekicidir. Literatürdeki eğilimler, tespit tekniklerinin davranışsal analizlere ve çok modlu yapay zekâ sistemlerine yöneldiğini göstermektedir. Buna ek olarak, olası olumsuz kullanımları sınırlamak amacıyla düzenlemeler ve yasaklamaların yanında etik kaygılar ve hukuki çerçevelerin de gelişmeye devam edeceği öngörülmektedir.



Şekil 2.1 Derin sahte videolarının yıllar içindeki artışı

Günümüzde derin sahte uygulamaları her geçen gün artmaktadır. Şu anda derin sahte'lerin şu alanlarda kullanıldığı söylenebilir [4]:

i. Eğlence ve Medya: Filmler ve video oyunlarında gerçekçi görsel efektler yaratmak ve Synthesia gibi uygulamalarda kurumsal eğitim videoları için kişiselleştirilmiş avatarlar oluşturmak.
ii. Müşteri Desteği: Yapay zekâ ile desteklenen ses klonlama, otomatik müşteri hizmeti sistemlerinde daha doğal sesli yanıtlar üretmek için kullanılmaktadır.
iii. Yanlış Bilgilendirme, Dezenformasyon ve Politik Propaganda: Günümüzde derin sahte, sahte bilgi yaymak, kamuoyunu manipüle etmek, seçimlere müdahale etmek, politik konuşmaları değiştirmek, sahte haber oluşturmak ve toplumsal tehdit yaratmak için de kullanılmaktadır.
iv. Dolandırıcılık ve Siber Güvenlik: Derin sahte ses ve video taklidi içeren dolandırıcılıklar dünya genelinde milyarlarca dolarlık kayıplara yol açmıştır. Örneğin, 2019 yılında bir derin sahte ses dolandırıcılığı, € 220 000 tutarında hileli para transferine neden olmuştur.
v. Eğitim ve Eğitim Süreçleri: Eğitim sistemleri ve öğretim süreçlerinde derin sahte'ler, gerçek dünya senaryolarını taklit etmek amacıyla kullanılmaktadır. Örneğin, tarih derslerinde tarihi kişilikler canlandırılabilir ya da polis ve ordu için sanal simülasyonlar hazırlanabilir.
vi. Pazarlama ve Reklam: Markalar, kullanıcılara yönelik kişiselleştirilmiş reklamlar oluşturmak için derin sahte teknolojisini kullanır. Bu reklamlarda sanal bir sözcü ya da yapay zekâ tarafından oluşturulmuş bir ünlü ürün tanıtır ve bu da hedef kitleyle daha güçlü bir duygusal bağ kurabilir.
vii. Oyun ve Sanal Gerçeklik: Derin sahte teknolojisi, gerçek karakterler gibi davranan oyuncu olmayan karakterler (NPC'ler) veya avatarlar yaratmak için kullanılabilir ve bu da sanal gerçeklik ile metaverse platformlarında kullanıcı deneyimini geliştirir.
viii. Sanat ve Yaratıcı Projeler: Sanatçılar, gerçek ve yapay öğeleri birleştirerek deneysel projeler oluşturmak için derin sahte araçlarını kullanır; bu projeler sanatın ve etkileşimli medyanın sınırlarını zorlar.

Günümüzde yapay zekadaki ilerlemeler nedeniyle derin sahte'ler hem daha erişilebilir hem de daha karmaşık hale gelmiştir. Teknik bilgisi olmayan kişiler bile DeepFaceLab gibi açık kaynak araçlara ve Zao gibi mobil uygulamalara erişerek son derece kaliteli derin sahte medya içerikleri üretebilmektedir. Bu tür teknolojik gelişmeler etkileyici olup, insanlığın

teknolojinin altın çağında yaşadığını gösterebilir. Ancak bu gelişmelerin sonuçları göz ardı edilmemelidir. Günümüzde derin sahte kullanılan dolandırıcılık yöntemleri artmaktadır ve bu kaybın 2028 yılına kadar 40 milyar dolara ulaşacağı tahmin edilmektedir. Ayrıca, derin sahte kullanılarak yapılan dezenformasyon kampanyaları medya güvenini sarsmakta ve neye inanılacağı her geçen gün daha da zorlaşmaktadır [5, 6].

Sentetik medya teknolojilerinin gelişim hızı o kadar yüksektir ki, görüntü adli bilimi ve yapay zekâ destekli tespit algoritmaları bu hıza yetişememektedir. Ancak buna rağmen, hükümetler, araştırmacılar ve özel sektör, zararlı derin sahte kullanımını durdurmak amacıyla her gün daha etkili araçlar geliştirmek için mücadele etmektedir [7]. Öte yandan, derin sahte, eğitim, eğlence ve erişilebilirlik gibi alanlarda oldukça faydalı ve meşru uygulamalara da sahiptir. Örneğin, görme engelli bireyler için içerik üretmek gibi [8].

2.1.1. Derin sahte tespiti

Derin sahte tespiti için farklı yaklaşımlar bulunmaktadır ancak mevcut tüm yöntemler arasında bazıları daha yaygın olarak kullanılmaktadır. Günümüzde uygulanmakta olan bazı derin sahte tespit yöntemleri çizelge 2.1'de açıklanmıştır. Bu yöntemler arasında derin öğrenme tabanlı yöntemler — CNN, RNN, LSTM, Transformers, Zaman Serisi Analizleri ve DNN — yer almakta olup, bu yöntemler yüksek doğruluk oranına sahiptir ve karmaşık örüntüleri öğrenme yeteneğine sahiptir. Derin öğrenme tabanlı yaklaşımlar, en etkili yöntemler arasında kabul edilmektedir. Öte yandan, el ile tasarlanmış özelliklere dayanan klasik makine öğrenmesi yöntemleri ve istatistiksel yöntemler, daha az hesaplama gücü gerektirir. Bunun yanında, birden fazla tespit işaretinin bir arada kullanılabildiği hibrit/birleşik (fusion) yöntemler de bulunmaktadır. Blockchain tabanlı yöntemler ise medyadaki değişiklikleri izleme açısından güvenilir derin sahte tespit teknikleridir; ancak, bu yöntemlerin uygulanması pahalıdır ve doğrudan bir tespit süreci içermezler.
Derin öğrenme yaklaşımlarında kullanılan mimari ve tekniklere göre, aşağıdaki derin sahte tespit modelleri örnek olarak gösterilebilir:

i. Evrişimli Sınır Ağlar (CNN'ler): CNN'ler, manipüle edilmiş video ve görüntülerdeki mekânsal tutarsızlıkları tespit etmek amacıyla kullanılabilir. Geleneksel CNN'lerin performansını artırmak için çoklu dikkat mekanizmaları (multi-attentional architectures)

veya topluluk yöntemleri (ensemble methods) gibi ek modüllere ihtiyaç duyulmaktadır. RNN modelleri, CNN'lerin bir alt dalı olarak değerlendirilmektedir [9, 10].

ii. Transformer Modelleri: Vision Transformers, derin sahte içeriklerdeki ince yapaylıkları tespit etmek amacıyla kendi kendine dikkat (self-attention) mekanizmalarını kullanır [11].

iii. Zamansal Analiz (Temporal Analysis): Video çerçevelerindeki zamanla ilgili tutarsızlıkları analiz eden modeller; örneğin doğal olmayan göz kırpma, düzensiz yüz hatları, doğal olmayan yüz hareketleri ve kareler arasındaki aydınlatma değişimleri gibi durumlar, derin sahte tespitinde oldukça etkili olmuştur [12].

iv. Biyolojik Sinyal Analizi: Derin sahte tespiti, videolarda yer alan fizyolojik ipuçlarının (örneğin nabız sinyalleri, göz kırpma, cilt tonu değişimleri gibi) izlenmesi yoluyla da yapılabilir. Bu biyolojik işaretlerdeki herhangi bir düzensizlik, sahte içeriklerin tespiti için önemli bir belirti olabilir [13].

Derin öğrenme tabanlı yaklaşımların yanı sıra, klasik makine öğrenmesi yaklaşımları da derin sahte tespitinde bir diğer çözümdür. Bu modeller, el ile tasarlanmış özelliklere (örneğin metin analizi veya istatistiksel örüntüler) dayanır. Bu yöntemler, derin öğrenmeye dayalı yaklaşımlara kıyasla daha az hesaplama gücü gerektirir [14].

Bu iki temel yönteme ek olarak, aşağıda yer alan hibrit ve birleşik (fusion) yöntemler de uygulanmaktadır:

i. Multimodal Birleşim (Multimodal Fusion): Görsel, metinsel ve işitsel verilerin birleştirilmesi, tamamlayıcı bilgilerin birlikte kullanılması yoluyla tespit doğruluğunu artırabilir [15].

ii. Adversaryal Eğitim (Adversarial Training): Modellerin, karşıt (adversaryal) senaryolarla eğitilmesi, gelişmiş derin sahte tespit tekniklerine karşı modelin dayanıklılığını artırabilir [15].

iii. Blockchain Tabanlı Teknikler: Medyanın özgünlüğünü doğrulamak amacıyla kaynağını ve üzerinde yapılan değişiklikleri izlemek için blockchain teknolojisinin kullanılması, etkin bir çözüm olarak kabul edilmektedir [14].

Çizelge 2.1. Derin sahte tespit yöntemlerinin karşılaştırılması

Yaklaşım	Anahtar Teknikler	Güçlü yönler	Zayıf yönler
Derin öğrenme temelli	CNN, RNN, DNN, LSTM, Transformers, Zamansal Analizler	Yüksek doğruluk, Karmaşık kalıpları öğrenebilir	Daha düşük doğruluk, yeni manipülasyon yöntemleriyle mücadele zorluğu
Klasik makine öğrenmesi	Öznitelik çıkarımları, istatistiksel yöntemler	Daha düşük hesaplama maliyeti	Büyük veri kümeleri gerektirir, yüksek hesaplama maliyeti
Hibrid/füzyon yöntemleri	Çok kipli Füzyon, Çekişmeli öğrenme	Birden fazla algılama ipucunu kullanabilir	Model tasarımı ve eğitimindeki karmaşıklık
Blokzinciri temelli	Medya kimlik doğrulama, Kaynak izleme	Medya değişikliklerinin güvenilir şekilde izlenmesi	Yüksek uygulama maliyeti, doğrudan bir tespit yöntemi değildir.

2.1.2. Derin sahte tespiti zorlukları

Derin sahte tespiti alanındaki dikkat çekici gelişmelere rağmen, hâlâ çözülmemiş bazı zorluklar mevcuttur. Bunlar, veri kümeleri arasında genelleme, gerçek zamanlı tespit ve gelişen teknolojiler olarak öne çıkmaktadır.

i. Veri Kümeleri Arasında Genelleme: Belirli veri kümeleri üzerinde eğitilen modeller, genellikle daha önce karşılaşılmamış manipülasyonları genelleme konusunda başarısız olurlar. Bunun başlıca nedeni, sahte içeriklerin oluşturulmasında kullanılan yöntemlerin farklılık göstermesidir [13, 16, 24].

ii. Gerçek Zamanlı Tespit: Gerçek zamanlı uygulamalarda ölçeklenebilirlik ve verimlilik sağlamak özellikle kaynaklara sınırlı erişimi olan mobil platformlar gibi cihazlarda oldukça zahmetli olabilir [15].

iii. Gelişen Teknolojiler: Çekişmeli Üretken Ağlar (GANs) mimarisinin hızlı gelişimi, derin sahte içeriklerin kalitesini artırmakta ve bu da tespit sürecini her geçen gün daha da zorlaştırmaktadır. Teknolojinin gelişmesiyle birlikte yeni eğilimler ortaya çıkmaktadır ve bunlardan bazıları aşağıdaki gibi sıralanabilir [13, 15, 17, 25].

a) Hafif Mimariler (Lightweight Architectures): Uç bilişim (edge computing) sistemleri ve mobil uygulamalardaki mevcut modellerin iyileştirilmesiyle erişilebilirlik artarken, aynı zamanda hesaplama kısıtları da çözülebilir.

b) Sıfır-Örnek ve Az-Örnekli Öğrenme (Zero-Shot ve Few-Shot Learning): Bu yaklaşımların amacı, modellerin daha önce görülmemiş veri kümeleri veya yeni manipülasyon tekniklerinde, yeniden eğitime gerek kalmadan dayanıklılığını artırmaktır.
c) Zamansal-Mekânsal Özellikler (Temporal-Spatial Features): Derin sahte tespit sistemlerinin yorumlanabilirliğini ve performansını artırmak amacıyla, zamansal mantığın mekânsal özelliklerle entegre edilmesi önerilmektedir.

2.2. Derin sahtelerin Sosyolojik Etkisi

Derin sahte teknolojileri, yalnızca teknik bir mesele olarak değil, aynı zamanda toplumsal yapılar üzerinde çok boyutlu etkiler yaratan önemli bir olgu olarak değerlendirilmektedir. Görsel-işitsel içeriklerin güvenilirliği, uzun yıllar boyunca toplumsal hafızanın, medya kurumlarına duyulan güvenin ve kamusal söylemin temel dayanaklarından biri olmuştur. Ancak derin öğrenme tabanlı üretim tekniklerindeki gelişmelerle birlikte, gerçek ve sahte içerikler arasındaki sınır giderek bulanıklaşmış ve bu durum bireylerin, kurumların ve toplumların bilgiye yaklaşım biçimlerini köklü bir şekilde dönüştürmüştür. Derin sahte içerikler; politik manipülasyon, toplumsal güvenin aşınması, cinsiyet ve güç dinamiklerinin yeniden şekillenmesi ve epistemik krizler gibi çeşitli sosyolojik sonuçlar doğurmaktadır. Bu bağlamda, derin sahte olgusunun sosyolojik açıdan incelenmesi, yalnızca teknolojinin potansiyel kötüye kullanım alanlarını anlamak açısından değil, aynı zamanda toplumsal güvenin yeniden inşası, medya okuryazarlığının geliştirilmesi ve dijital çağda bilgi ekosisteminin sürdürülebilirliği bakımından da kritik bir önem taşımaktadır.

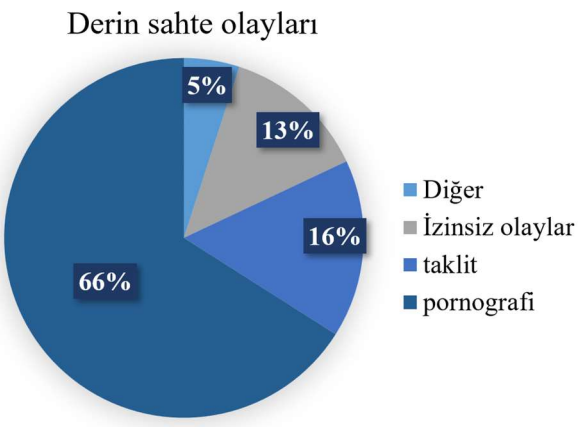
2.2.1. Derin sahte anlamada kuramsal yaklaşımlar

Derin sahtelerin sosyolojik önemi, dijital medya ile etkileşim biçimlerimizi kökten değiştirme ve görsel-işitsel kanıtların doğruluğuna dair yüzyıllardır süregelen varsayımları sarsma potansiyeline sahip olmalarıdır. Önceki medya manipülasyonları yalnızca profesyonellere özgü bir alanken, günümüzde herhangi biri modern teknolojileri kullanarak derin sahte üretebilir. Bu erişilebilirlik, kurumların ve bireylerin giderek karmaşıklaşan bilgi ortamında nasıl işledikleri üzerinde ciddi sonuçlar doğurur. Derin sahte'lere yönelik sosyolojik incelemeler, farklı kuramsal geleneklerden beslenmektedir. Gerçekliğin toplumsal inşası kuramı, derin sahte'lerin ortak gerçeklik ve otantiklik algımızı nasıl sarstığını açıklamada güçlü bir yaklaşımdır. Görsel-işitsel kanıtın güvenilirliği

bozulduğunda, gerçekliğe dair ortak bilgi üretme yolları zarar görür. Sembolik etkileşimcilik geleneği ise iletişimdeki güven temellerinin nasıl bozulduğunu anlamak açısından farklı bir bakış açısı sunar [18].

2.2.2. Medya ve politika üzerindeki etkileri

Derin sahte'lerin politik sonuçları özellikle önemli olmuştur. 2020 ABD başkanlık seçimlerinden önce internet, adaylar, onların politikaları ve seçim süreci hakkında yanlış ve yanıltıcı bilgiler yaymak amacıyla kullanılan derin sahte görüntü ve videolarla dolup taşmıştır. Bu olayların tamamı, halkın genel kanaatini manipüle etmek ve siyasi kurumlara olan güveni sarsmak yoluyla derin sahte'lerin demokratik süreçleri baltalama potansiyeline sahip olduğunu göstermektedir. Şekil 2.2'de de görülebileceği gibi, seçim bağlamının ötesinde, derin sahte'ler genel olarak medya sistemleri için de bir tehdit oluşturmaktadır. Pornografi, içerik sahibinin izni olmadan oluşturulan materyaller ve sahte kimlik taklidi, derin sahte içeriklerinin en sık üretildiği alanlardır. Derin sahte maruziyeti üzerine yapılan araştırmalar, belirsizlik yarattığını ve sosyal medya ile haber kaynaklarına olan genel güveni azalttığını ortaya koymuştur. Bu güven kaybı, hâlihazırda ana akım medya kurumlarına duyulan düşük güvenle tanınan bir çağda daha da endişe verici olmakta ve güvenilirliği istikrarsız olan diğer bilgi kaynaklarının da bu duruma eklenmesiyle daha da karmaşık hâle gelmektedir [19]. Hükümetler dijital platformlardan ekonomik katkılar sağladıkları için bu platformlara katı düzenlemeler getirememektedir. Politik açıdan ise, yeniliği teşvik etmek ve küresel pazarda rekabet edebilmek istemektedirler; bu nedenle düzenleyici yaklaşımlarda esneklik söz konusu olmaktadır [20].



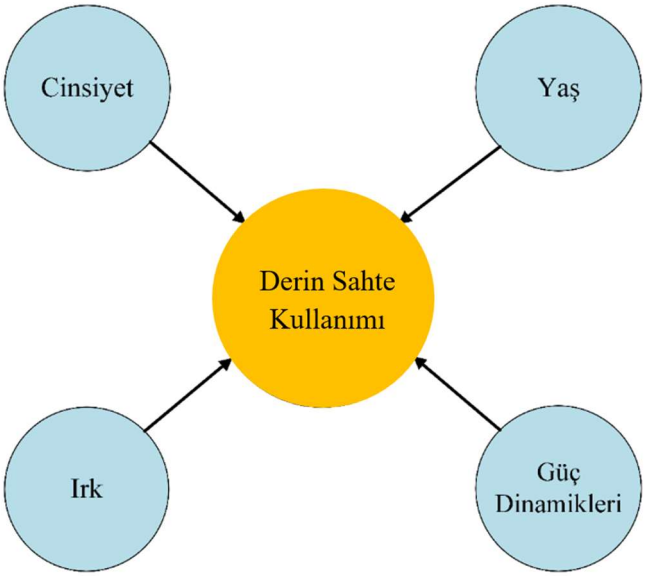
Şekil 2.2. Derin sahte olaylarının tiplerine göre dağılımı

2.2.3. Demografi ve güç dinamikleri

Derin sahte'lerin sosyolojik sonucu, özellikle ırk, cinsiyet ve yaş boyutlarında belirgindir. En güncel çalışmalar, kadınların bu teknolojinin kötüye kullanımından orantısız şekilde etkilendiğini, özellikle rıza dışı pornografi biçiminde olduğunu göstermektedir. Cinsiyet yönünden bu şiddet eğilimi, mevcut toplumsal eşitsizlikleri pekiştirmekte ve toplumsal cinsiyete dayalı şiddetin teknolojik bir uzantısı hâline gelmektedir. Benzer şekilde, derin sahte'ler ırka dayalı önyargıları temsil edebilir ve hatta bunları doğrulayabilir. Haut ve ark. (2021), ırka göre güvenilirlik algısı üzerindeki etkileri test etmiş ve videolarda ırksal görünüm değişikliğinin güvenilirlik algısını önemli ölçüde etkilediğini doğrulamıştır. Bu araştırmaya dayanarak, derin sahte'lerin mevcut toplumsal önyargılarla birleşerek zararlı stereotipleri ve ayrımcılığı güçlendirebileceği açıktır [21]. Çocuklar da derin sahte içeriklerin hedefi hâline gelebilmektedir. "Child sexual abuse images and youth produced images: The varieties of Image-based Sexual Exploitation and Abuse of Children," kaynağına göre, çocukların derin sahte içeriklerde istismara uğradığı beş farklı kategori bulunmaktadır. Bu kategoriler, ticari çocuk cinsel istismarı, ticari olmayan istismar, gençler tarafından üretilen cinsel içerik, zorla ya da istismarcı amacıyla gençler tarafından üretilmiş içeriklerve yapay olarak oluşturulmuş veya simüle edilmiş içerik olarak ortaya çıkmaktadır [22]. Bu tür eylemlerin durdurulması ve önlenmesi için yasal ve politik önlemlere dair zorlukların giderilmesi, teknolojik karşı önlemlerin geliştirilmesi, eğitsel girişimlerde bulunulması ve ebeveyne katılımın sağlanması zorunluluk olarak ortaya çıkmaktadır.

i. Yasal ve politik önlemlere dair zorluklar: Belirtilen sorunlarla ilgili önlemler alınmalı, bir çerçeve belirlenmeli ve uygunsuz eylemler durdurulmalıdır.
ii. Teknolojik karşı önlemler: Gerçek ve sahte sömürü içeriklerinin tespiti ve yayılmasını önlemek için gelişmiş araçlar geliştirilmelidir. Zaman geçtikçe kolluk kuvvetleri çocuklara yönelik cinsel istismar materyallerinin yayılmasına karşı daha etkili adımlar atmaya çalışmaktadır, bu da güncel teknolojik araçlara sahip olmayı gerektirir.
iii. Eğitsel girişim: Gençlerin açık içerik üretme ve paylaşma riskleri konusunda eğitilmesi, onları bu tür davranışlardan uzak tutmada önemli bir adımdır.
iv. Ebeveyn katılımı: Ebeveynler, çocuklarının çevrim içi etkinliklerine aktif şekilde katılmaya teşvik edilmeli ve onları internet güvenliği hakkında bilgilendirmek gerekmektedir [23-24]. Yaygın algı, çevrim içi tacizcilerin genellikle yabancılar olduğu yönündedir; ancak yapılan çalışma, bunların çoğunlukla akranlar ve mağdurun tanıdığı

kişiler, örneğin yakın arkadaşları olduğunu, yabancıların ya da mağdurun kimliğini bilmediği kişilerin ise bu gruptan sonra geldiğini ortaya koymuştur [25].



Şekil 2.3 Derin sahte kullanımını etkileyen faktörler

2.2.4. Sosyal bilim ve deneysel uygulamalar

Derin sahte'lerin taşıdığı gerçek tehlikelere ilişkin ciddi kaygılar göz önüne alınarak son dönem araştırmaların bu konu üzerinde kontrollü deneysel çalışmalar yaptığı görülmektedir. Haut ve arkadaşlarının (2021) gerçekleştirdiği çalışma, sosyal bilimlerde derin sahte teknolojisinin deneysel amaçla kullanımına dair ilk örnektir. Bu çalışma, "bilgisayar bilimi bilgisi sınırlı olan sosyal bilimcilerin kabul edilebilir sürede gerçekçi bir derin sahte oluşturabileceğini" göstermektedir [26]. Öğrencilerin bir öğretmenin fiziksel çekiciliğine dair algılarını inceleyen çalışmaları, derin sahte'ler kullanılarak deneydeki değişkenlerin hassas şekilde nasıl kontrol altına alınabileceğini ve aynı zamanda ekolojik geçerliliğin nasıl korunabileceğini ortaya koymaktadır.

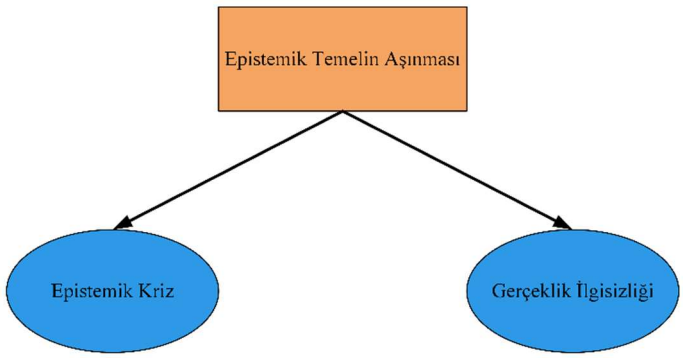
Bu tür bir çalışma, sosyal bilim metodolojisi için güçlü bir araç olabilir; belirli faktörleri sistematik olarak değiştirirken diğerlerini sabit tutarak kontrollü görsel ve işitsel uyarıcılarla deney yapma olanağı sunabilir. Araştırmacılar, bu yaklaşımla nedensel etkileri geçmişe kıyasla daha etkili şekilde ortaya koyabilirler. Ancak, bu tür bir çalışmanın uygulanması esnasında, deneylerde bilgilendirilmiş onam ve aldatma gibi önemli etik sorunlar da gündeme gelebilir. Diğer yandan dijital platformların devreye girmesiyle sosyal etkileşim ve

bilgi yayılımı dönüşmüş, iletişim, eğitim ve eğlencenin ayrılmaz bir parçası hâline gelmiştir. Toplumun bu platformlara olan bağımlılığı, özellikle erişimi veya ifade özgürlüğünü kısıtlıyor gibi göründüğünde, düzenleyici önlemler sansür olarak algılanabilir. Kullanıcı kitlesinin çok uluslu yapısı ve farklı bölgelerdeki kültürel standartların çeşitliliği hem etkili hem de kültürel olarak uygun düzenlemelerin uygulanmasını daha da karmaşık hâle getirebilmektedir [27].

2.2.5. Güven aşınması ve epistemik krizler

Derin sahte'lerin en büyük sosyolojik etkisi belki de pek çok kuramcının "epistemik kriz" olarak adlandırdığı durumu yaratma gücünde yatmaktadır. Duyusal verilerin güvenilirliğine dair en temel varsayımlar, bu teknoloji tarafından sarsılmakta ve bu durum, toplu gerçeklik bilgisinin giderek ulaşılması zor hâle geldiği bir "gerçek sonrası" toplumun itici gücü olabilmektedir [28].

Bu epistemik temelin aşınması, tekil yanlış temsil vakalarının çok ötesine uzanır. Derin sahte'ler ne kadar yaygın ve gelişmiş olursa, bu durum genel olarak medyaya, hatta doğru şekilde sunulmuş içeriklere duyulan güvenin çökmesine yol açabilir. Bu durum, zaman zaman "gerçeklik kayıtsızlığı" olarak adlandırılır ve kolektif eylem ile toplumsal dayanışma için gerekli olan ortak epistemolojik temeli zayıflattığı anlamında temel bir sosyolojik mesele teşkil eder.



Şekil 2.4 Derin Sahtelerin sosyolojik etkileri

2.2.6. Toplumsal tepkiler ve uyum

Derin sahte tehlikesine karşı çeşitli toplumsal tepkiler ortaya çıkmıştır. Bu tepkiler; yapay medya içeriklerini tespit etmeye yönelik teknolojik çözümlerden, kötü niyetli kullanımı caydırmayı amaçlayan yasal ve düzenleyici önlemlere, medya okuryazarlığını teşvik etmeyi hedefleyen pedagojik müdahalelere kadar uzanmaktadır. Bu tepkilerin her biri, sorunun doğasına ve en uygun çözüm yoluna dair farklı varsayımlar içermektedir [29].

Derin sahte tehlikelerine verilen toplumsal tepkiler teknolojik, hukuki ve eğitsel olmak üzere üç ana eksende şekillenmektedir. Teknolojik yaklaşım, gerçek ve yapay medya içeriklerini birbirinden ayırt edebilecek daha gelişmiş ve hassas algoritmalar geliştirmeye odaklanmaktadır. Bu algoritmalar, yalnızca yüzeysel farklılıkları değil, aynı zamanda içerikteki derin yapaylık izlerini de tespit edebilecek şekilde tasarlanmaktadır. Hukuki yanıt ise özellikle rıza dışı oluşturulan derin sahte içerikleri ve teknolojinin diğer kötüye kullanım alanlarını doğrudan yasaklayan düzenlemeler şeklinde ortaya çıkmaktadır. Bu tür düzenlemeler bireylerin mahremiyetini korumayı ve toplum genelinde zarara yol açabilecek içeriklerin yayılmasını engellemeyi hedeflemektedir. Eğitsel yaklaşım ise yapay içerik çağında bireylerin eleştirel düşünme ve medya okuryazarlığı becerilerini geliştirmesinin zorunluluğunu vurgulamaktadır; çünkü günümüzde bilgiye erişimin kolaylaşması aynı zamanda sahte içeriklere maruz kalma riskini de artırmaktadır. Son olarak, tüm bu ortamın sosyolojik bir değerlendirmesi, etkili bir yanıtın yalnızca tek bir alandaki müdahaleyle değil; teknoloji, hukuk, eğitim ve medya gibi alanlarda uyumlu ve koordineli bir iş birliğiyle mümkün olacağını açıkça ortaya koymaktadır [30].

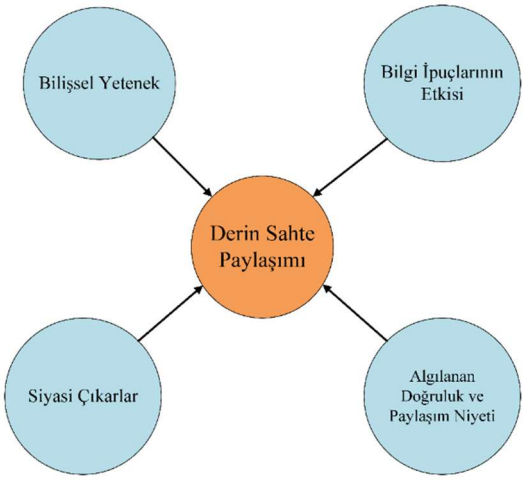
Derin sahte içerikleri, çok katmanlı ve çok yönlü bir gerçeklik taşıyan sosyolojik bir etkiye sahiptir; bu etkinin sonuçları, toplumsal yaşamın pek çok alanına yayılmaktadır. Siyasetten toplumsal cinsiyet ilişkilerine, bilimsel çalışmalardan bilgi kuramına (epistemoloji) kadar uzanan geniş bir yelpazede, derin sahte'lerin ortaya koyduğu gerçeklik, mevcut sosyal yapıları ve kabulleri sarsmakta, alışılmış doğrulara meydan okumaktadır. Bu durum, yalnızca yüzeysel bir medya sorunu değil, toplumun değer sistemlerini, bilgiye ulaşma yöntemlerini ve hatta bireyler arası güven ilişkilerini temelden etkileyen bir dönüşüm anlamına gelmektedir. Bu nedenle, derin sahte olgusunun yarattığı sosyolojik kırılmalarla baş edebilmek için yansıtıcı, eleştirel ve çok taraflı karşı önlemler alınması gerekmektedir.

Bu önlemler; teknolojik, hukuki, kültürel ve etik düzeylerde kapsamlı analizleri ve iş birliklerini zorunlu kılmaktadır.

Sonuç olarak, bu teknolojiyi anlamak, doğru değerlendirmek ve topluma uyumlu bir şekilde entegre edebilmek için gelecekte sosyolojik bir kavrayışa her zamankinden daha fazla ihtiyaç duyulacaktır. Gelecekte yapılacak araştırmaların, derin sahte'lerin farklı toplumsal kurumlar üzerindeki etkilerini deneysel ve ampirik olarak incelemesi, çeşitli kültürel ve siyasi ortamlarda benimsenen farklı düzenleyici yaklaşımları karşılaştırmalı olarak değerlendirmesi ve potansiyel faydalı kullanım alanlarını, zararları en aza indirme çabalarıyla birlikte keşfetmeye odaklanması büyük önem taşımaktadır. Bu tür çalışmalar, hem derin sahte teknolojisinin risklerini yönetmek hem de topluma sağlayabileceği yararları ortaya koymak adına belirleyici olacaktır.

2.2.7. Derin sahte paylaşımını etkileyen faktörler

Derin sahte içeriklerin paylaşılmasında etkili olabilecek bazı önemli faktörler bulunmaktadır. Şekil 2.5'te derin sahte paylaşımını etkileyen genel faktörler görülmektedir.



Şekil 2.5 Derin sahte paylaşımını etkileyen faktörler

Bu faktörler, bireylerin sosyal, bilişsel ve politik eğilimlerine göre değişiklik gösterebilmekte ve bu değişkenler, bir kişinin sahte içerikleri fark etmeden ya da kasıtlı olarak paylaşma olasılığını doğrudan etkileyebilmektedir. Bu faktörler genelde karşımıza politik ilgi, bilişsel yetenek, sosyal ağın büyüklüğü olarak çıkmaktadır.

i. Politik İlgi: Siyasi konulara yüksek ilgi duyan bireyler, diğer bireylere kıyasla derin sahte içeriklerini paylaşmaya daha yatkındır. Bu durumun temelinde, bu kişilerin siyasete aktif olarak katılım göstermeleri ve politik gündemi yakından takip etmeleri yatmaktadır. Siyasal bilgiye ulaşma arzusu, onları çeşitli bilgi kaynaklarına yöneltmekte; ancak bu kaynaklar arasında sahte ya da manipüle edilmiş içeriklere maruz kalma ihtimali de artmaktadır. Politik olarak angaje bireylerin, özellikle kendi görüşlerini destekleyen veya karşıt görüşleri zayıflatmayı hedefleyen derin sahte içerikleri bilinçli ya da bilinçsiz şekilde yayma olasılığı daha yüksektir. Bu durum, yanlış bilgi döngüsünün hızla yayılmasına neden olabilmektedir.
ii. Bilişsel Yetenek: Daha düşük düzeyde bilişsel yeteneğe sahip bireyler, karşılaştıkları içeriklerin doğruluğunu ve güvenilirliğini eleştirel bir şekilde değerlendirme konusunda zorluk yaşayabilirler. İçeriğin ne kadar mantıklı veya gerçek dışı olduğunu analiz etmek için gerekli olan eleştirel düşünme becerileri eksik olduğunda, derin sahte gibi görsel ve işitsel olarak ikna edici içeriklerin sahte olduğunu anlamak zorlaşır. Bu bireyler, içeriğin doğruluğunu sorgulamadan onu gerçekmiş gibi kabul edebilir ve bu nedenle farkında olmadan sahte içerikleri yayabilir. Bu durum, özellikle sosyal medya ortamlarında yaygın olarak gözlemlenmektedir.
iii. Sosyal Ağın Büyüklüğü: Siyasi olarak ilgili bireylerden oluşan geniş sosyal çevrelere sahip kişiler, derin sahte içerikleri paylaşma konusunda daha büyük bir etkiye sahiptir. Çünkü bu bireyler, sahip oldukları sosyal etkileşim ağı sayesinde içerikleri daha fazla kişiye ulaştırma potansiyeline sahiptirler. Ayrıca, çevresel baskı ve grup normları da paylaşım kararlarını etkileyebilir. Eğer kişinin çevresi belirli bir ideolojik yapıya sahipse ve bu yapı içerisinde derin sahte içerikler sıkça paylaşılıyorsa, kişi de bu davranışı normalleştirerek aynı şekilde paylaşımda bulunabilir. Sosyal medya gruplarının ve dijital toplulukların etkisi, bu tür davranışların pekişmesinde önemli rol oynamaktadır.

Bu faktörlerin her biri, tek başına etkili olabileceği gibi birbirleriyle etkileşim hâlinde de çalışabilir. Dolayısıyla, derin sahte içeriklerin neden bu kadar hızlı yayıldığını anlamak için bu etkenlerin bütüncül bir yaklaşımla değerlendirilmesi gerekmektedir [31].

Bilgi İpuçlarının Etkisi: Politik olmayan içerikler söz konusu olduğunda, bağlamsal bilginin bulunmaması, derin sahte içeriğin gerçek olduğuna inanılma olasılığını artırmaktadır. Kişi, içerikle ilgili arka plan bilgisine sahip değilse, içeriği sorgulamadan kabul etme eğilimi gösterebilir.

Algılanan Doğruluk ve Paylaşım Niyeti: Derin sahte içeriğin gerçekliğine duyulan inanç ne kadar yüksekse, bireyin bu içeriği paylaşma isteği de o ölçüde artmaktadır. Başka bir deyişle, içeriğin doğru olduğuna inanmak, paylaşım kararını doğrudan etkileyen önemli bir faktördür [32].

3. MATERYAL VE METOT

Toplu Dönüşümlü Artık Ağlar (Residual Networks with Aggregated Transformations-ResNeXt50), derin sahte tespit uygulamalarında kullanılabilen bir çeşit Evrişimli Sinir Ağı (Convolutional Neural Network- CNN) tabanlı ağ yapısıdır. Uzun Kısa Süreli Bellek (Long Short Time Memory – LSTM) ise yine derin sahte tespiti görevlerinde kullanılabilen bir Tekrarlayan Sinir Ağı (Recurrent Neural Network-RNN) tabanlı ağ yapısıdır. Bu tezde, veri kümesi olarak Derin Sahte Algılama Veri Kümesi (DeepFake Detection Dataset- DFDC) veri kümesi kullanılmıştır. Hem ResNeXt50 hem de LSTM yöntemleri her birisi tek tek ve en sonunda hibrid model olarak uygulanmış ve derin sahte videoları tespit edebilmek amacıyla her biri 20 epoch boyunca eğitilmiştir. Derin sahte videoların'ın tespitine yönelik önerilen çözümün performansı, sonuçların görselleştirme teknikleriyle analiz edilmesi ve karışıklık matrisi kullanılarak değerlendirilmiştir.

Tezin bu bölümünde, çalışmada kullanılan veri kümeleri ve önerilen yöntemin altyapısını oluşturan bütün mimariler detaylı olarak açıklanmıştır. Bu kapsamda öncelikle derin sahte veri kümeleri incelenmiş, derin sahte geliştirmede etkili bir şekilde kullanılan üretken çekişmeli ağlar hakkında bilgi verilmiştir. Bölümde ayrıca CNN ve ResNet mimarileri ve türevleri ile LSTM yapısı açıklanmıştır.

3.1. Derin Sahte Veri Kümeleri

Derin sahte tespitinde kullanılan algoritmaların yanı sıra, veri kümeleri de geliştirilen tespit yöntemlerinin eğitilmesi ve değerlendirilmesi sürecinde önemli bir rol oynamaktadır. Veri kümeleri, derin sahte tespiti görevlerinde temel bir öneme sahiptir. Veri kümesi ne kadar gerçekçi, çeşitli ve büyük olursa, elde edilecek sonuçlar da o kadar kesin ve anlamlı olacaktır. Bu bağlamda kullanılabilecek öne çıkan veri kümeleri çizelge 3.1'de gösterilmekte ve aşağıda açıklanmaktadır.

FaceForensics++ (FF++): En erken dönemlerden bu yana en yaygın kullanılan veri kümelerinden biri olan FaceForensics++ (FF++), dört otomatik yüz manipülasyon yöntemi DeepFakes, Face2Face, FaceSwap ve NeuralTextures kullanılarak manipüle edilmiş 1000 orijinal videodan oluşur. Veri kümesinin ana kaynağı, çeşitli yüz görünümleri ve arka

planlara sahip 977 YouTube videosudur. Bu veri kümesinde, segmentasyon görevleri için ikili maske (binary mask) etiketleri sağlanmakta ve görüntü ile video sınıflandırma görevlerini desteklemektedir. FF++, üzerinde birçok tespit yönteminin değerlendirildiği bir kıyaslama standardı hâline gelmiştir. Sistematik derlemeler ve anket çalışmalarında yaygın olarak kullanılmıştır ve alandaki en fazla deneysel çalışmaya sahip veri kümesi olarak öne çıkmaktadır [16-18, 33-35].

DFDC: Derin Sahte Tespit Yarışma (Deep Fake Detection Challenge- DFDC) veri kümesi güncel derin sahte araştırmalarında kullanılmak üzere tasarlanmış geniş çaplı bir veri kümesidir. Facebook tarafından sunulmuş olup, 3400'den fazla ücretli oyuncunun yer aldığı 1000 000'dan fazla klip içermektedir. Bu videolar, üretken yapay zekâ tabanlı ve öğrenmeye dayanmayan farklı manipülasyon teknikleri kullanılarak üretilmiştir. Amacı, sahne ve oyuncu çeşitliliği açısından gerçek dünya derin sahtekârlıklarını yansıtmaktır. DFDC'nin ölçeği ve çeşitliliği nedeniyle, "gerçek dünyadaki" derin sahte'lere iyi genelleme yapabilen tespit modellerinin geliştirilmesinde değerli bir kaynak olarak kabul edilir [19, 20, 36].

Celeb-DF: Celeb-DF, 590 gerçek video ve 5639 sahte video içeren büyük ve çalışması zor olan bir veri kümesidir. Her bir video toplamda iki milyondan fazla film karesine sahiptir. Bu veri kümesinin geliştirilmesindeki amaç, erken dönem veri kümelerinin sınırlılıklarını — düşük kalite ve gerçek dışı yapaylıklar gibi — ortaya koyarak, daha yüksek kaliteli ve daha zorlayıcı sahte videolar sağlamaktı. Celeb-DF, geliştirilmiş gerçekçilik ve çeşitliliğe sahip ikinci nesil bir derin sahte veri kümesidir ve tespit algoritmalarının dayanıklılığını değerlendirmek için bir kıyaslama noktası hâline gelmiştir [13, 14, 21, 38].

Wild Deepfake: Bu veri kümesi, internetten elde edilmiş 707 derin sahte videosundan çıkarılan 7314 yüz sekansı içermektedir. Bu veri kümesinin en büyük avantajı, gerçek dünya derin sahte'lerinin çeşitliliğini ve öngörülemezliğini temsil etmesidir; bu da çoğu zaman tespit modellerinin zorlandığı bir alandır [39].

DeeperForensics-1.0: Bu veri kümesi, farklı düzeylerde çeşitli bozulmalar (perturbation) uygulayarak, gerçek dünya bozulmalarına odaklanır ve sonuç olarak tespit dayanıklılığını artırmayı amaçlar.

Google/Jigsaw Deep Fake Detection Dataset: FF++ veri kümesinin bir parçasıdır. Mevcut verinin çeşitliliğini artırmaya yardımcı olmak amacıyla farklı sahnelerde yer alan 28 oyuncunun oynadığı 3000'den fazla manipüle edilmiş videodan oluşmaktadır [40].

UADFV, DF-TIMIT, FF-DF: Bu veri kümeleri birinci nesil veri kümeleridir. Boyutları küçüktür, çeşitlilikleri düşüktür ve genellikle tespit edilmesi kolay belirgin yapaylıklar barındırırlar. Bu nedenle, bunlar üzerinde tespit yapmak genellikle zor bir görev olarak kabul edilmez.

Çizelge 3.1. Derin sahte çalışmalarında kullanılan veri kümeleri

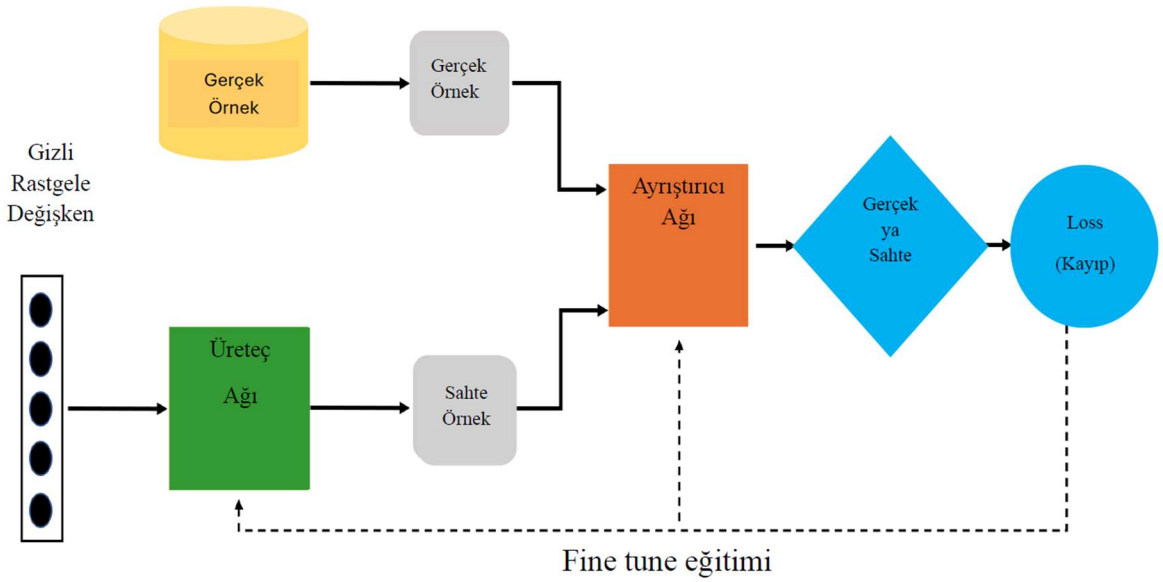
Veriseti	Yıl	Video Boyutu	Kaynak/ Çeşitlilik	Manipülasyon Biçimleri	Gerçekçilik /Zorluk	Kullanım Örnekleri
FaceForensics++	2018	1 000+	YouTube	Derin sahteler	Orta	Algılama vb. için karşılaştırmalı değerlendirme
DFDC	2019	100 000+	Ücretli oyuncular	Çoklu	Yüksek	Büyük ölçekli, gerçek dünyaya genelleme
Celeb-DF	2020	6229	Ünlüler	Derin sahteler	Yüksek	Sağlam, tespiti zorlu çalışmalar
Wild Deepfake	2019	707	Internet	Gerçek dünya sahteleri	Çok yüksek	Sahteliklere karşı genelleme
Deeper forensics-1.0	2020	60 000+	Çoklu	Yüz değiştirme	Yüksek	Bozulmalara karşı dayanıklılık

Derin sahte tespit çalışmalarında kullanılabilecek mevcut veri kümeleri için henüz standartlaşmış kıyaslama ölçütleri bulunmamaktadır. DeepFakeBench gibi projeler, test protokollerini, veri işleme süreçlerini ve değerlendirme metriklerini daha standart hâle getirerek karşılaştırmaların daha adil olmasını ve araştırmaların daha tekrarlanabilir olmasını amaçlamaktadır. Ayrıca, DeeperForensics-1.0 ve WildDerin sahte gibi veri kümeleri, modellerin gerçek dünya koşullarına ne ölçüde genelleme yapabildiğini doğrulamak açısından kritiktir; çünkü kontrollü ortamlarda eğitilmiş modeller, gerçek hayata genellenemez. Yöntemsel etkilere bakıldığında, derin öğrenme — özellikle CNN tabanlı yaklaşımlar — baskın yöntemlerdir ve performans üzerinde kullanılan veri kümesinin önemli bir etkisi olduğu gözlemlenmiştir.

3.2. Üretken Çekişmeli Ağlar

Üretken çekişmeli Ağlar (Generative AdversarialNetwork- GAN), genellikle denetimsiz öğrenme yöntemine dayanan makine öğrenmesi modelleridir ve iki ayrı sinir ağı bileşeninden oluşurlar. Şekil 3.1'de GAN blok diyagramı görülemektedir. GAN lar da iki sinir ağı, biri Üreteç (Generator – G) diğeri ise Ayrıştırıcı (Discriminator – D) olarak adlandırılır. Üreteç ağı, ters evrişimli (deconvolutional) sinir ağı yapısına sahiptir ve gerçeğe çok yakın, yapay olarak üretilmiş içerikler oluşturarak bu içeriklerin gerçekmiş gibi algılanmasına neden olabilir. Öte yandan, Ayrıştırıcı ağı, CNN temellidir ve bu yapay içeriklerle gerçek içerikler arasındaki farkı ayırt etmeye çalışır; aldığı girdileri analiz ederek hangisinin gerçek hangisinin sahte olduğunu belirlemeye çalışır.

Bu iki sinir ağı, sistem içinde sürekli olarak birbirleriyle rekabet hâlindedir. Üreteç, ayrıştırıcıyı kandıracak kadar gerçekçi içerikler üretmeye çalışırken; ayrıştırıcı da gelen verinin gerçek mi yoksa yapay mı olduğunu doğru bir şekilde tespit etmeye odaklanır. Bu süreçte, her iki ağ da performanslarını artırmak için kendilerini geliştirir ve doğruluk açısından bir yarış içerisindedir. Sonuç olarak, bu doğruluk yarışının galibi olan ağ, diğerine göre daha yüksek bir başarı oranı elde eder. Bu etkileşim sayesinde, GAN yapıları zamanla daha sofistike ve ikna edici sahte içerikler üretme kapasitesine ulaşır.



Şekil 3.1 GAN blok diyagramı

Veri x üzerinde bir üretken dağılımı p_g'yi öğrenebilmek için, üretken (generator), veri uzayı üzerinde tanımlı olan ön gürültü dağılımı p_z(z) 'den G(z;θ_g) şeklinde bir eşleme (mapping) fonksiyonu oluşturur. Ayrıştırıcı (discriminator) ise D(x;θ_d) şeklinde tanımlanır ve x'in gerçek eğitim verisinden mi yoksa üretken tarafından üretilen p_g 'den mi geldiğini belirten tek bir skaler olasılık çıktısı üretir.

$$\min_G \max_D V(D,G) = \mathbb{E}_x \sim p_{data}(x)[\log D(x)] + \mathbb{E}_z \sim p_z(z)[\log(1 - D(G(z)))] \quad (1)$$

Üretken ağın temel amacı, üretim miktarını artırmak ve bu sayede ayrıştırıcının (discriminator) hata oranını yükseltmektir; böylece oluşturulan bazı sahte çıktılar, ayrıştırıcı tarafından gerçek olarak algılanabilir. Üretken ile ayrıştırıcı arasındaki etkileşim biçimleri ve matematiksel formüller doğrultusunda, farklı türlerde GAN yapıları ortaya çıkmıştır. Bu GAN türlerinden bazıları şunlardır: Vanilya GAN (Vanilla GAN), Koşullu GAN (Conditional GAN- cGAN), Derin evrişimli GAN (Deep Convolutional GAN- DCGAN), CycleGAN, StyleGAN, DiscoGAN ve Süper Çözünürlüklü GAN (Super Resolution GAN SRGAN). Vanilla GAN, GAN ailesinin en temel ve en sade türüdür. Bir üretken ve bir ayrıştırıcı ağdan oluşur. Bu yapı, ayrıştırıcıdan doğrudan geri bildirim almadan çıktılar üretir. Vanilya GAN, orijinal görüntülerden öğrenerek yeni ve benzer görüntüler oluşturur. Bu süreçte, üretken ağ rastgele bir gürültü dağılımı üzerinden çıktılar üretirken, ayrıştırıcı bu çıktının gerçek olup olmadığını anlamaya çalışır. Her iki ağın bu şekilde yarışması, sistemin öğrenme kapasitesini geliştirir. cGAN yapılarında ise, üretilecek veriye ilişkin belirli bir kriter veya koşul ağlara önceden tanıtılır. Bu durumda ağlar, rastgele gürültü kullanarak hata oranını artırmak yerine, kendilerine verilen bu koşul doğrultusunda çalışırlar. Koşul uygulanmış veri kümeleri, klasik GAN'lara kıyasla çok daha doğru ve amaca yönelik sonuçlar elde edilmesini sağlar. Özellikle, veri kümelerine ait sınıf etiketlerinin hem üretkenye hem de ayrıştırıcıya gönderilmesiyle, bu koşullara uygun daha net ve yönlendirilmiş çıktılar elde edilir. Yani, elde edilen sonuçlar yalnızca rastlantısal değil; verideki belirli özellikler gözetilerek üretilir. Bu yöntem hem görsel kaliteyi hem de hedef doğruluğunu artırır [41].

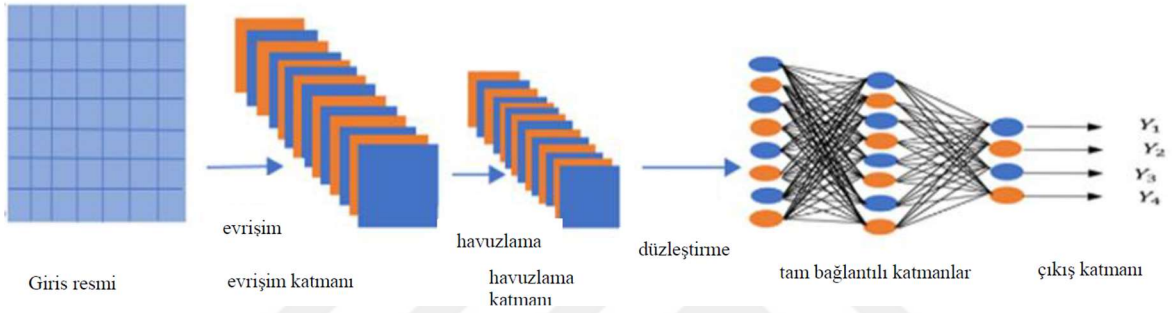
$$\min_G \max_D V(D,G) = \mathbb{E}_x \sim p_{data}(x)[\log D(x \mid y)] + \mathbb{E}_z \sim p_z(z)[\log(1 - D(G(z \mid y)))] \quad (2)$$

DCGAN, diğer GAN türlerinde olduğu gibi üretken ağ ve ayrıştırıcı ağ olmak üzere iki ana ağdan oluşur. Bu tür GAN'ların temel farkı, özellikle üretken ağ kısmında yer alan ekstra evrişim katmanlarıdır. Yani, giriş ve çıkış arasına eklenen bu evrişimli katmanlar sayesinde daha karmaşık ve detaylı görüntü üretimi mümkün hâle gelir.

Öte yandan, ayrıştırıcı ağda sahte ve gerçek görüntüleri sınıflandırabilmek için evrişimsel katmanlar kullanır. Bu katmanlar, görsel verinin daha iyi analiz edilmesini sağlar ve modelin gerçek ile yapay görüntüler arasındaki farkı ayırt etme yeteneğini geliştirir. DCGAN yapıları, özellikle görüntü üretimi alanında daha yüksek başarı oranlarıyla öne çıkar [42]. CycleGAN'da amaç, görüntüden görüntüye çeviri yapabilen bir derin evrişimli sinir ağı eğitmektir. Örneğin, bir manzaranın yaz mevsiminden kış mevsimine dönüştürülmesi gibi. Bu tür ağlarda, giriş ve çıkış görüntüleri arasındaki eşleştirme, birbirleriyle eşleştirilmiş veri kümeleri kullanılarak gerçekleştirilir. Bu modelde hem üretken hem de ayrıştırıcı ağlar eğitilmelidir. Üretken ağ, bir görüntüyü başka bir stile çevirirken, ayrıştırıcı ağ bu çevirinin gerçek mi yoksa yapay mı olduğunu değerlendirmeye çalışır. Böylece sistem, iki yönlü dönüşümleri öğrenerek giriş görüntüsünden hedef görüntüye ve tekrar geri dönüş yapmayı öğrenir [43]. SRGAN, görüntülerin çözünürlüğünü artırarak onları çok daha yüksek kaliteye ulaştırmak amacıyla kullanılan üretken çekişmeli sinir ağı türleridir. Bu tür ağlarda, özellikle düşük çözünürlüklü görüntülerin detaylarını ve keskinliğini artırarak, insan gözüne hitap eden ve daha doğal görünen sonuçlar elde edilmeye çalışılır. SRGAN'larda, sadece çekişmeli kayıp fonksiyonu (adversarial loss function) değil, aynı zamanda algısal ya da içerik kaybı (perceptual/content loss) adı verilen ikinci bir kayıp fonksiyonu da hesaplanır. Bu ek kayıp fonksiyonu, görüntülerin sadece sayısal olarak değil, algısal düzeyde de insan görüşüne yakınlığını ölçebilmek için kullanılır. Yani görüntünün kalitesini değerlendirirken, çıplak gözle algılanabilecek detaylara da dikkat edilir. Ağın üretici kısmı, eğitim veri kümesindeki örneklerden öğrenerek yüksek çözünürlüklü sahte görüntüler üretmeye çalışırken ayrıştırıcı kısmı ise, bu üretilen görüntülerin gerçek mi yoksa yapay mı olduğunu belirlemeye çalışır. SRGAN'ın temel amacı, o kadar yüksek kalitede görüntüler üretmektir ki, ayrıştırıcı ağ bile bu görüntülerin gerçek olup olmadığını ayırt edemez hâle gelsin. Bu da modelin gerçekçiliğini ve üretim gücünü kanıtlar [2, 44-47].

3.3. Evrişimli Sinir Ağı

Evrişimli Sinir Ağı (Convolutional Neural Network - CNN), özellikle görüntü tanıma ve görüntü işleme için tasarlanmış özelleştirilmiş bir derin öğrenme mimarisidir. CNN'in temel odak noktası, hiyerarşik özellik çıkarımı kullanarak görsel verilerdeki desenleri tanımlamaktır. CNN'in bu özelliği, onu nesne tespiti, yüz tanıma ve medikal görüntü analizi gibi bilgisayarlı görü (computer vision) uygulamaları için uygun kılar. Şekil 3.2 de Evrişimli sinir ağı katmanları görülmektedir.



Şekil 3.2 Evrişimli sinir ağı

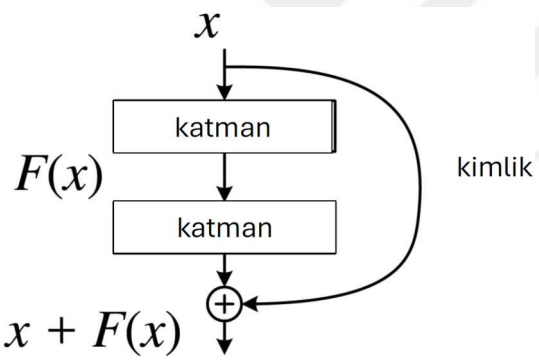
CNN, giriş görüntülerine filtreler uygular ve kenarlar ile dokular gibi düşük seviyeli özellikleri, şekiller ve nesneler gibi yüksek seviyeli desenleri tanımlar. Havuzlama katmanları (pooling layers), maksimum havuzlama (max pooling) veya ortalama havuzlama (average pooling) işlemleriyle özellik haritalarının uzamsal boyutlarını azaltarak hesaplama verimliliğini artırır. Tam bağlantılı katmanlar (fully connected layers), yüksek boyutlu veriyi vektörlere indirger ve son sınıflandırmayı gerçekleştirir. Eğer giriş verisi bir 3D görüntü ise, CNN'in ilk katmanları çizgiler ve kenarlar gibi temel özellikleri belirler, ardından üst katmanlar bu temelleri kullanarak daha karmaşık nesneleri oluşturur. Daha sonra aşağı örnekleme (downsampling) aşaması gelir. Bu aşamada havuzlama katmanları, uzamsal boyutları azaltarak önemli özelliklere odaklanılmasını sağlar. Sonraki adım sınıflandırmadır. Bu aşamada tam bağlantılı katmanlar çıktıyı üretir ve nesneleri sınıflandırır. CNN'lerin en büyük avantajlarından biri, özellikleri (features) otomatik olarak öğrenmeleridir. Diğer bir ifadeyle özellik mühendisliği işlemleri elle yapılmaz, desenler doğrudan veriden öğrenilir. Ayrıca CNN'ler, katmanlı işleme yoluyla yerel ve küresel detayları ve bağlamı yakalayan uzamsal bir hiyerarşiden yararlanır. Bunlara ek olarak, verimlidirler: Ağırlıklar filtreler arasında paylaşılır, bu da hesaplama maliyetini azaltır. Ancak CNN'lerin dezavantajı, eğitim için büyük ve etiketlenmiş veri kümelerine ihtiyaç duymalarıdır. Ayrıca eğitim süreci için

güçlü Grafik İşlem Birimi (Graphical Process Unit – GPU) veya Tensör İşlem Birimi (Tensor Process Unit – TPU) gerekir [48-49].

3.4. RESNET

Artık ağları (Residual Network - ResNet), kaybolan gradyan (vanishing gradient) problemini çözmek amacıyla 2015 yılında tanıtılmıştır. Bu ağlar, şekil 3.3 de gösterüldüğü gibi atlamalı bağlantılar (residual blocks) kullanarak derin öğrenme dünyasında devrim yaratmıştır [50]. Bu bloklar sayesinde ağlar yüzlerce katmanla eğitilebilmekte, gradyanlar doğrusal olmayan katmanları atlayarak geçiş yapabilmekte ve böylece bilgi akışının durması önlenmektedir [51]. Residual mapping,

$$F(x) = H(x) - x \tag{3}$$



Şekil 3.3 ResNet

ResNet, tıbbi görüntüleme, tarım ve nesne tespiti gibi alanlarda kullanılabilir. ResNet uygulamasındaki en büyük zorluk aşırı öğrenme (overfitting) problemidir. Derin ResNet modellerinde bu sorunu önlemek için agresif veri artırımı (augmentation) ve etiket yumuşatma (label smoothing) yöntemlerinin uygulanması gereklidir [52].

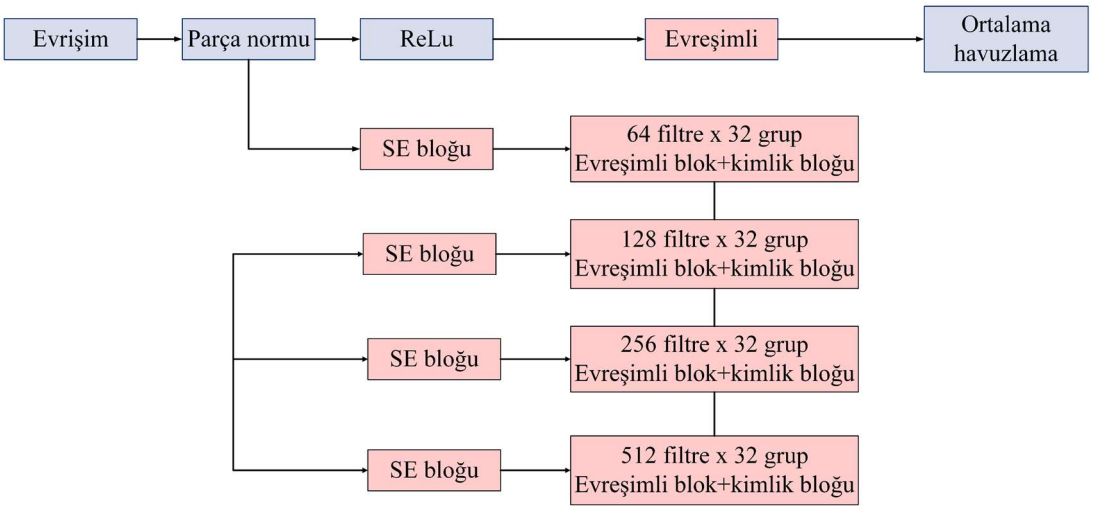
3.5. ResNeXt50

Toplu Dönüşümlü Artık Ağlar (Residual Networks with Aggregated Transformations-ResNeXt50), 2016 yılında ResNet50'nin evrimi olarak tanıtılmış ve özellikle özellik çeşitliliğini artırmayı hedefleyen bir mimari olarak geliştirilmiştir. Bu mimarinin en temel katkısı, "cardinality" (kardinalite) kavramıdır. Kardinalite, bir evrişim katmanında giriş

verisinin işlendiği paralel yolların (grupların) sayısını ifade eder. Böylece ağın yapısını değiştirmeden, daha çeşitli ve zengin özellik dönüşümleri elde edilebilir. ResNeXt50, VGG/ResNet'in katman yığma stratejisini, Inception mimarisindeki "böl-parça-birleştir" yaklaşımıyla birleştirerek, katman sayısını artırmadan ve hesaplama maliyetini yükseltmeden modelin doğruluğunu ve genel performansını iyileştirmeyi amaçlar.

ResNeXt50'nin bir diğer önemli özelliği gruplandırılmış evrişimler (grouped convolutions) kullanmasıdır. Bu yaklaşım, filtreleri 32 gruba ayırarak daha az parametreyle daha verimli özellik çıkarımı sağlar. Ayrıca mimari, Sıkıştırma ve Uyarma (Squeeze-and-Excitation, SE) bloklarıyla güçlendirilebilir. SE blokları, kanallara özgü yanıtları ortalama havuzlama ve uyarlanabilir yeniden ağırlıklandırma teknikleriyle kalibre ederek doğruluk oranını özellikle tıbbi görüntüleme gibi alanlarda %2–3 seviyesinde artırabilmektedir. Mimari aynı zamanda modüler tasarımı ile öne çıkar. Kimlik (identity) ve evrişim blokları esnek bir biçimde birleştirilebilir; bu da 50 ila 101 katman arasında derinlik ölçeklenebilirliği sağlar. Böylelikle ResNeXt50, farklı görevler ve veri kümelerine kolayca uyarlanabilen esnek bir yapı sunar [53].

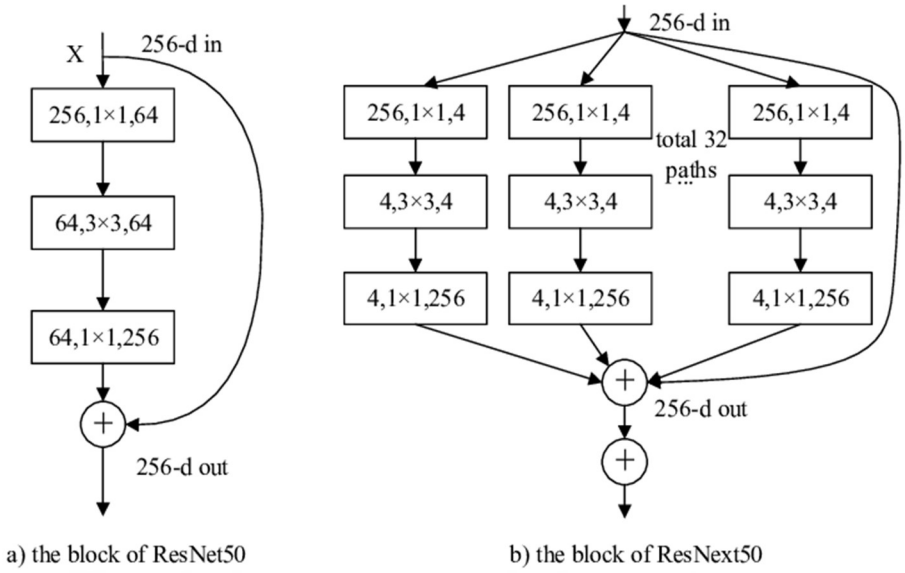
Genel olarak ResNeXt yapısı şekil 3.4 de görülmektedir. Yapı, evrişim blokları, kimlik blokları ve isteğe bağlı SE blokları üzerine kuruludur. Bu birleşim, ResNeXt50'nin ResNet50'ye kıyasla daha yüksek doğruluk ve esneklik elde etmesini sağlar. Adındaki "son ek" (örneğin 50 veya 101), toplam katman sayısını belirtir. Modüler bileşenleri ve kardinalite boyutları sayesinde ResNeXt mimarisi şekil 3.5 de görüleceği gibi farklı bloklarla oluşturularak, karmaşık görevleri etkin biçimde yönetebilen güçlü bir derin öğrenme modeli olarak öne çıkmaktadır [54-56].



Şekil 3.4. ResNeXt50 yapısı

ResNet50 ile ResNeXt50 karşılaştırıldığında, ResNet50'nin her katmanda sabit sayıda filtre kullandığı, buna karşılık ResNeXt50'nin işlem maliyeti ile özellik çeşitliliği arasında denge sağlamak için gruplanmış evrişimler (grouped convolutions) kullandığı söylenebilir. Ayrıca, SE bloğu, bilgilendirici kanalları vurgulayarak ResNeXt50'nin performansını daha da artırır ve karmaşık görevlerde daha yüksek doğruluk elde edilmesini sağlar.

Genel olarak, ResNeXt50, kardinalite, SE blokları ve modüler tasarımı kullanarak hesaplama maliyeti ile performans arasında etkili bir denge kurar. ResNet, Inception ve SE mimarilerinin prensiplerini birleştirerek daha az parametreyle üstün özellik çıkarımı elde eder; bu da onu tıbbi görüntüleme ve nesne tespiti gibi derin öğrenme uygulamaları için güçlü bir seçenek haline getirir.

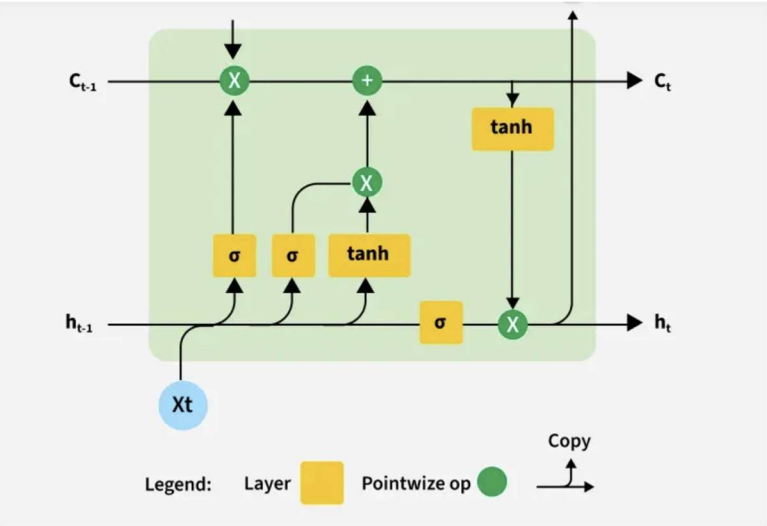


Şekil 3.5. ResNet50 ve ResNeXt50 blokları [57]

3.6. Uzun Kısa Süreli Bellek

Uzun Kısa Süreli Bellek (Long Short Term Memory- LSTM) ağları, sıralı verilerde uzun vadeli bağımlılıkları yönetmek için tasarlanmış özel bir Yinelemeli Sinir Ağı (Recurrent Neural Network- RNN) türü olarak tanımlanabilir. Bu ağlar, kapılı bellek yapıları kullanarak kaybolan gradyan problemlerini azaltmak suretiyle uzun vadeli bilgiyi öğrenmeyi kolaylaştırır. LSTM'in yapısı şekil 3.6 da gösterilmiştir ve temelde üç kapı ve bir hücre durumundan oluşan bellek bloklarına dayanır. İlk kapı olan Unutma Kapısı (Forget Gate), sigmoid aktivasyon fonksiyonu aracılığıyla önceki hücre durumundan hangi bilginin silinmesi gerektiğine karar verir. İkinci kapı olan Girdi Kapısı (Input Gate), hücre durumuna yeni eklenecek bilgiyi sigmoid ve tanh aktivasyonlarının kombinasyonu ile kontrol eder. Üçüncü kapı olan Çıktı Kapısı (Output Gate), mevcut hücre durumu ve önceki gizli duruma göre son gizli durum çıktısını ayarlar. Hücre durumu (cell state), kalıcı bellek gibi davranarak, birkaç ardışık adımdan sonra bile bilgilerin geri çağrılmasını mümkün kılar. LSTM'ler, RNN'lerle karşılaştırıldığında belleği daha iyi yönetebilir çünkü hücre durumu ve kapıları kullanarak uzun süreli verileri tutabilirken, RNN'ler sadece gizli durumlara (hidden states) dayanır. Ayrıca LSTM'ler dizi işlemede (sequence handling) daha üstündür; 1 000'den fazla adımdan oluşan bağımlılıkları işleyebilirken, RNN'ler kısa dizilerde daha etkilidir. LSTM ağları konuşma tanıma, zaman serisi analizi, doğal dil işleme (NLP), video altyazı üretimi, sağlık verisi analizi ve verimlilik geliştirme gibi alanlarda uygulanabilir.

Ancak LSTM'lerin dezavantajı yüksek hesaplama yükü taşımaları ve yorumlana bilirliklerinin düşük olmasıdır [58-62].



Şekil 3.6. LSTM diyagramı [63]

LSTM'in çalışma şekli şu şekilde açıklanabilir:
Bir kelime dizisi $(w_1, w_2, w_3, \dots, w_n)$, içinde, LSTM'in t zaman adımındaki durumu (h_t, c_t) olarak ifade edilir; burada h_t gizli katman (hidden layer) ve c_t hücre durumudur (cell state). İlk olarak, giriş vektörü x_t ile önceki durum (h_{t-1}, c_{t-1}) LSTM tarafından alınır. Ardından unutma kapısı (f_t) tarafından hücre durumundan hangi bilginin silinmesi gerektiğine karar verilir. Bu karar, giriş vektörü ile önceki gizli durum temel alınarak, hücre durumundaki (c_{t-1}) her sayı için 0 ile 1 arasında bir değer üretilerek verilir. Burada 1 sayısı "tamamen koru", 0 sayısı ise "tamamen sil" anlamına gelir.

Forget Gate: $f_t = \sigma(W_f.[h_{(t-1)}, x_t] + b_f)$ (4)

Sonra, hücre durumuna hangi bilginin kaydedileceğine karar vermek için giriş kapısı (i_t) devreye girer ve bu kapı iki bölümden oluşur. İlk bölüm, hangi değerlerin güncelleneceğine karar veren sigmoid katmanı olan "giriş kapısı katmanı"dır. İkinci bölüm ise yeni aday değerleri (ct~) üreten tanh katmanıdır ve bu değerler hücre durumuna eklenebilir.

Input Gate: $i_t = \sigma(W_i.[h_{(t-1)}, x_t] + b_i)$ (5)

Cell State Update: $\tilde{c}_t = \tanh(W_c.[h_{(t-1)}, x_t] + b_0)$ (6)

Daha sonra eski hücre durumu (c_{t-1}) yeni hücre durumuna (c_t) güncellenecektir. Eski hücre durumu f_t and then the new candidate values will be added. ile çarpılacak ve ardından yeni aday değerler buna eklenecektir.

Final Cell State: $c_t = f_t \times c_{(t-1)} + i_t \times \tilde{c}_t$ (7)

Daha sonra çıktı belirlenecektir. Öncelikle bir sigmoid katmanı çalıştırılır, ardından hücre durumu tanh işlevinden geçirilir ve elde edilen değer sigmoid kapısının çıktısı ile çarpılır.

Output Gate: $o_t = \sigma(w_o.[h_{(t-1)}, x_t] + b_0)$ (8)

Hidden State: $h_t = o_t \times \tanh(c_t)$ (9)

3.7. Hibrit Model

CNN ile RNN ağlarını birleştiren hibrit modeller, yapay zekâ tarafından üretilen medya manipülasyonlarının tespitinde güncel literatürde öne çıkan en etkili yaklaşımlardan biri olarak değerlendirilmektedir. Bu modeller, özellikle uzamsal ve zamansal özelliklerin birlikte işlenmesini sağlayarak, yalnızca tekil karelere dayalı yöntemlerin ötesine geçmektedir. CNN tabanlı yapılar, video karelerinden mikro düzeyde mekânsal bilgileri çıkarmada oldukça etkilidir. Yüz dokuları, kenar hatları, göz kırpma frekansı ve cilt desenlerindeki bozulmalar gibi yüzeysel anomaliler, ResNet, Inception ResNetV2 veya ResNeXt gibi güçlü derin mimariler aracılığıyla yakalanabilmektedir [64] . Ayrıca, transfer öğrenme ve ince ayar (fine-tuning) yöntemleriyle önceden eğitilmiş CNN modellerinin yeniden uyarlanması, daha geniş veri kümelerinin etkin kullanımını sağlamakta ve derin sahte tespit performansını artırmaktadır [64].

Buna karşın RNN tabanlı LSTM mimarileri, videoların ardışık karelerindeki zamansal bağımlılıkları değerlendirme kapasitesiyle öne çıkmaktadır. Doğal olmayan yüz hareketleri, aydınlatma farklılıkları, dudak senkronizasyonu hataları veya ardışık kareler arasındaki

tutarsızlıklar gibi zamana bağlı anomaliler, LSTM tabanlı yaklaşımlar sayesinde daha doğru bir şekilde yakalanabilmektedir [65] . Böylelikle, CNN'in uzamsal duyarlılığı ile LSTM'in zamansal tutarlılık analizi birbirini tamamlamakta ve hibrit mimari altında bütünleşmektedir.

CNN ve LSTM'nin entegrasyonu, literatürde genellikle "mekânsal-zamansal özellik birleşimi" (spatio-temporal feature fusion) olarak tanımlanmaktadır. CNN katmanları mikro düzeyde doku analizini üstlenirken, LSTM katmanları 30–50 karelik ardışık sekanslar boyunca makro düzeyde hareket örüntülerini ortaya çıkarmaktadır [66] . Bu bütünleşik yapı, yalnızca kare bazlı analizlerde gözden kaçabilecek düzensiz göz kırpma sıklığı, doğal olmayan baş hareketleri, ritmik olmayan yüz kas oynamaları ya da dudak senkronizasyonundaki milisaniyelik zaman kaymaları gibi ince anomalilerin tespit edilmesine imkân tanımaktadır [67]. Ayrıca, yüz kırpma (face cropping), normalizasyon ve veri artırma (data augmentation) gibi ön işleme teknikleri, arka plan gürültüsünü azaltarak sistemin daha dayanıklı ve genelleştirilebilir hale gelmesine katkı sağlamaktadır [32][66].

Her ne kadar CNN–LSTM hibrit mimarisi doğruluk ve güvenilirlik açısından önemli avantajlar sunsa da, hesaplama maliyetlerinin yüksekliği ve kaynak yoğunluğu nedeniyle gerçek zamanlı uygulamalardaki uygulanabilirliği sınırlı kalmaktadır. Bu nedenle, güncel araştırmalar hafifletilmiş mimarilerin (lightweight architectures) geliştirilmesine ve görsel-işitsel ipuçlarını birlikte işleyen çok modlu çerçevelere odaklanmaktadır[68]. Literatürde raporlanan ampirik bulgular, hibrit CNN–LSTM sistemlerinin tek modlu yaklaşımlara kıyasla kayda değer performans artışları sağladığını göstermektedir. Örneğin, Al Dulaimi ve arkadaşları tarafından gerçekleştirilen bir çalışmada [64], yalnızca kare bazlı yöntemlere göre %23'e varan bir doğruluk artışı elde edilmiştir. Ancak, derin sahte üretim tekniklerinin sürekli evrim geçirmesi, mevcut hibrit yapıların zamanla yetersiz hale gelmesine yol açmakta; bu da uyarlanabilir, daha verimli ve gerçek zamanlı kullanılabilir hibrit mimarilerin geliştirilmesini gelecekteki araştırmalar için kritik bir gereklilik haline getirmektedir [69].

Tezde önerilen hibrit (ResNeXt50/LSTM) modelde, manipüle edilmiş video içeriklerinde mekânsal (spatial) ve zamansal bozulmaları yakalayabilmek için evrişimli bir omurga yapısı, yinelenen (recurrent) zamansal modelleme ile bütünleştirilmiştir. Bir video dizisi, X= $\{x_t\}_{t=1}^{T}$ biçiminde tanımlanmakta olup, her bir kare $x_t \epsilon \mathbb{R}^{3\times H\times W}$ uzayında yer almaktadır. Ayırt edici mekânsal temsilleri çıkartabilmek amacıyla, her kare ResNeXt-50 evrişimli ağı üzerinden

işlenmekte ve tam bağlı (fully connected) katmana ulaşmadan önce kısaltılarak aşağıdaki özellik haritasının elde edilmesini sağlamaktadır:

$$F_t = g_{\theta cnn}(x_t) \in \mathbb{R}^{(C \times h \times w)} \tag{10}$$

Burada c=2048 olarak tanımlanmaktadır. Küresel ortalama havuzlama (global average pooling) işlemi, bu temsili sabit uzunlukta bir vektöre sıkıştırmaktadır.

$$f_t = GAP(F_t) = \frac{1}{hw} \sum_{(i=1)}^{h} \sum_{(j=1)}^{w} F_t[:, i, j] \in \mathbb{R}^C \tag{11}$$

Kareler arasındaki zamansal dinamikleri bütünleştirmek amacıyla, $\{f_t\}_{t=1}^{T}$ dizisi, gizli boyutu d=2048 olan tek katmanlı bir Long Short Term Memory (LSTM) ağına aktarılmaktadır. Her bir adımda gerçekleşen yinelenen güncellemeler, önceki bölümde ayrıntılı olarak sunulan standart LSTM denklemleri ile denetlenmektedir.

$$i_t = \sigma(W_i f_t + U_i h_{t-1} + b_i) \tag{12}$$

$$f_t = \sigma(W_f f_t + U_f h_{t-1} + b_f) \tag{13}$$

$$o_t = \sigma(W_o f_t + U_o h_{t-1} + b_o) \tag{14}$$

$$\tilde{c}_t = \tanh(W_c f_t + U_c h_{t-1} + b_c) \tag{15}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{16}$$

$$h_t = o_t \odot \tan h(c_t) \tag{17}$$

Burada başlangıç durumları $h_0 = 0$ ve $C_0 = 0$ olarak tanımlanmaktadır. Ayrıca, σ fonksiyonunu, ⊙ ise Hadamard çarpımını temsil etmektedir. Zamansal birleştirme (temporal

aggregation), tüm dizi boyunca elde edilen gizil durumların (hidden states) ortalaması alınarak gerçekleştirilmektedir.

$$h' = \frac{1}{T}\sum_{t=1}^{T} h_t \tag{18}$$

Yukarıda elde edilen temsil, olasılığı 0.4 olan dropout düzenlileştirmesine tabi tutulmaktadır. Bu, yalnızca eğitim aşamasında geçerli olmak üzere, zamansal özellik vektörünün bileşenlerinin %40'ının rastgele elenerek düzenlileştirildiği anlamına gelmektedir. Ardından, elde edilen vektör tam bağlı (fully connected) bir sınıflandırma katmanına aktarılmaktadır.

$$Z = Wh' + b,\ \hat{y} = \text{softmax}(z) \tag{19}$$

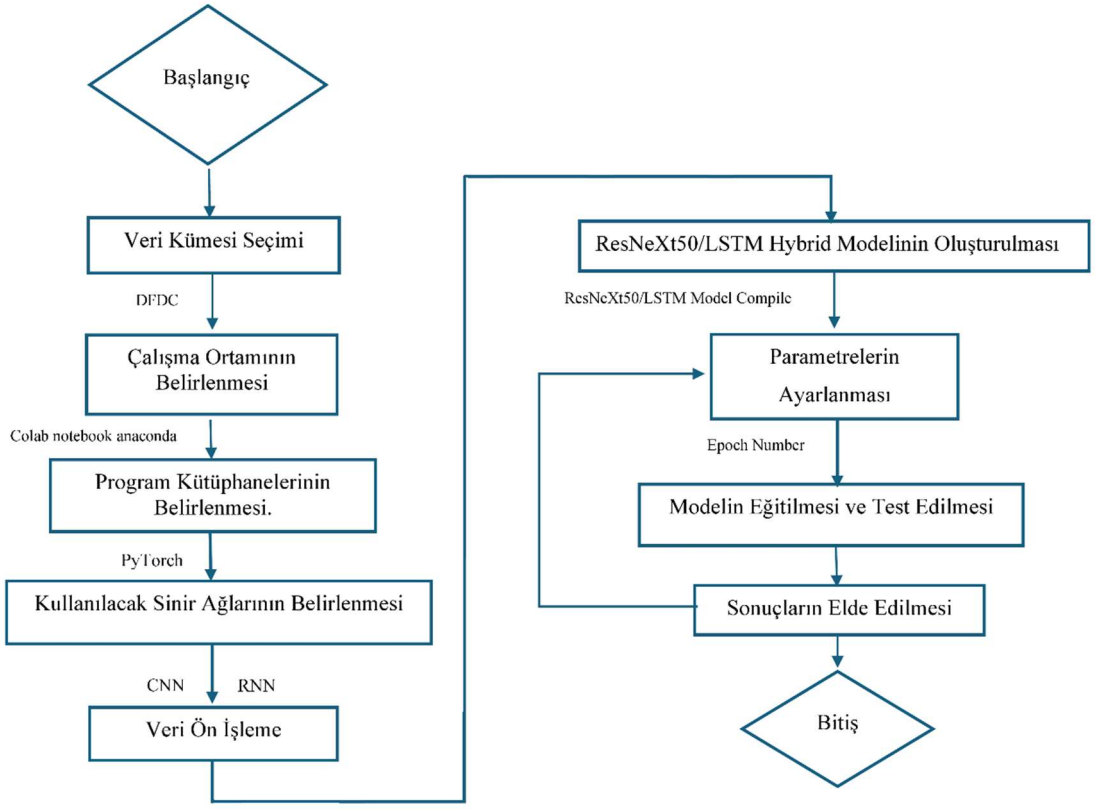
Son adımda, $0 = \{\theta_{cnn}, \theta_{lstm}$,W,b} parametrelerini içeren model, kategorik çapraz entropi (categorical cross-entropy) kaybı kullanılarak optimize edilmektedir.

$$\mathcal{L}(o) = -\frac{1}{N}\sum_{(n=1)}^{N}\sum_{k=1}^{k} y_k^{(n)} \log \hat{y}_k^{(n)} \tag{20}$$

Burada *k=2*, gerçek ve sahte sınıflandırması için kullanılmaktadır ve $(X^{(n)}, y^{(n)})$ ifadesi, n-inci video örneğini ve ona karşılık gelen kodlanmış etiketi göstermektedir.

4. UYGULAMA

Bu bölümde, önerilen yöntemin gerçekleştirildiği aşamalar sırasıyla ele alınmuıştır. Şekil 4.1'de tez sürecinin akış diyagramı görülmektedir. Bu çalışmada verimli ve yüksek görsel gerçekçiliğe sahip olması nedeniyle detayları Bölüm 3.1'de verilen DFDC veri kümesi kullanılmıştır. Kontrollü deneyler içeren bu veri kümesi yalnızca yüz ifadelerindeki değişikliklere odaklanmakta ve arka plan neredeyse hiç değiştirilmemektedir. Ayrıca, bu veri kümesi hedef aktörün yüzünü kontrol etmek için kaynak aktörün yüz hareketlerini kullanır [40, 70].



Şekil 4.1. Tez süreci akış diyagramı

Modeli geliştirebilmek için yüksek işlem gücüne sahip ve güçlü bir GPU içeren bir platforma ihtiyaç duyulmaktadır. Bu bağlamda, Google şirketinin bulut tabanlı hizmetlerinden biri olan Google Colaboratory Pro+ kullanılmıştır. Kullanılan GPU, NVIDIA A100'dür ve CPU, 8vCPU kapasitesindedir (ücretli).

Programlama işlemi Python dili kullanılarak Jupyter Notebook ortamında gerçekleştirilmiştir. Program, kullanılacak kütüphaneleri belirlemek üzere uygulamaya

yüklenmiştir. Bu programlamada derin öğrenme için Torch (PyTorch), video ve görüntü işleme için OpenCV (cv2), görselleştirme için seaborn, sayısal işlemler için NumPy ve işletim sistemiyle etkileşim kurmak için os kütüphaneleri kullanılmıştır. PyTorch, açık kaynaklı bir derin öğrenme çatısıdır ve esnek yapısı, dinamik hesaplama grafiği oluşturabilmesi ve kullanıcı dostu olmasıyla popülerdir. PyTorch ağırlıklı olarak Sinir Ağları (CNN, RNN gibi) tasarlamak, eğitmek ve değerlendirmek amacıyla kullanılır. İçerisinde Linear, Conv2d ve LSTM gibi ön tanımlı katmanlar bulunmaktadır; bu da karmaşık modellerin oluşturulmasını kolaylaştırır. Ayrıca, PyTorch içerisinde ResNet, VGG, Inception gibi önceden eğitilmiş modellere de erişim sağlar. NumPy'ye benzer şekilde güçlü tensör işlemleri sunabilir ancak bunu GPU hızlandırmasıyla yapar. Bununla birlikte, veri yükleme ve toplu veri işleme için kullanılan DataLoader ve Dataset sınıflarını da içerir.

OpenCV, görüntülerin yeniden boyutlandırılması veya gri tonlamaya çevrilmesi gibi görüntü işlemlerini gerçekleştirme imkânı sunar. Ayrıca video işlemleri de sağlar; bu işlemler arasında video okuma, yazma ve kare yakalama gibi özellikler yer alır. OpenCV'nin bir diğer özelliği ise yüz tespiti yapabilmesidir ve bu çalışma kapsamında bu özellikten yararlanılmıştır. Os modülü ise dizinlerde gerçekleşen dosya işlemleri ve ortam değişkenlerini yönetme gibi görevleri yerine getirebilir. Ayrıca sistem düzeyinde çeşitli işlemleri gerçekleştirme kapasitesine sahiptir. Bu modül, işletim sistemiyle platformlar arası etkileşimde önemli bir rol oynar ve bu da dosya yollarının işlenmesi, dosya yönetimi ve sistem komutlarının çalıştırılması gibi işlemler için gereklidir.

Modelin uygulanabilmesi için öncelikle verilerin Google Drive ortamına yüklenmesi gerekmektedir. Veriler Google Drive'a başarıyla yüklendikten sonra uygulamanın ilk aşaması olan ön işleme (preprocessing) süreci başlatılabilir. Bu aşama, modelin nihai sonuçları üzerinde büyük etkiye sahip olduğu için oldukça kritik bir rol oynamaktadır.

Bu tez kapsamında kullanılan DFDC (DeepFake Detection Challenge) veri kümesi, oldukça büyük boyutlu ve kapsamlı bir veri kümesidir. Bu nedenle, ilgili kodların tamamının veri kümesi üzerinde tek seferde çalıştırılması hem çok uzun zaman alabilmekte hem de işlem süresinin aşırı uzaması nedeniyle hata ile sonuçlanabilmektedir.

DFDC veri kümesi, 00'dan 49'a kadar numaralandırılmış toplam 50 alt bölüme (split) ayrılmıştır. Bu tezde kullanılan yöntemsel yaklaşımda, aşağıda ayrıntıları verilen tüm ön

işleme ve modelleme süreçleri her bir alt bölüm için ayrı ayrı yürütülmüştür. Başka bir ifadeyle, ilgili kodlar veri kümesinin her bir bölümünde 50 kez tekrarlanmış ve böylece her split için ayrı sonuçlar elde edilmiştir.

Tüm bölümler üzerinde yapılan tekrarların ardından, elde edilen çıktılar istatistiksel olarak birleştirilmiş ve ortalama değerler hesaplanmıştır. Nihai olarak, tezde sunulan Çizelge lar ve şekillerde gösterilen sonuçlar bu ortalama değerler üzerinden raporlanmıştır. Böylece, hem veri kümesinin büyüklüğünden kaynaklanan işlem zorlukları aşılmış hem de daha güvenilir ve genellenebilir sonuçlar elde edilmiştir.

Ön işleme aşamasının temel bir gerekliliği, yüz tanıma işlemlerini kolaylaştırmak amacıyla bir yüz tanıma modelinin kurulmasıdır. Bu modelin entegrasyonunun ardından, veri kümesindeki videoların bozuk olup olmadığı kontrol edilmiş, her bir videodan 20 kare (frame extraction) çıkarılarak mini-batch'ler oluşturulmuştur. Sonraki aşamada, tüm kareler 112×112 piksel boyutuna yeniden ölçeklendirilmiş (resizing) ve geçersiz olduğu tespit edilen video sayısı raporlanmıştır.

Bunu takiben, veri kümesinin sırası rastgeleleştirilmiş ve 100 kareden daha az içeren, kısa olarak sınıflandırılan videolar veri kümesinden çıkarılmıştır. Ardından, her bir video için kare sayıları hesaplanmış ve eğitime hazır hale getilimiştir. Görüntülerin doğru şekilde işlenebilmesi ve görselleştirilebilmesi için kanal sıralaması RGB olarak ayarlanmış, normalizasyonu geri almak amacıyla ortalama ve standart sapma değerleri uygulanarak orijinal piksel değerleri yeniden elde edilmiştir. Daha sonra, tüm görüntüler 0–255 aralığına ölçeklendirilmiş ve yardımcı işlevler aracılığıyla uygun biçimde düzenlenmiştir. Bu ön işleme adımları, yüz odaklı ve standartlaştırılmış bir veri kümesinin oluşturulmasını sağlayarak derin sahte tespit sürecinde hem doğruluk hem de güvenilirlik açısından önemli katkılar sunmuştur.

Ön işleme sürecinin tamamlanmasının ardından, veri kümesi eğitim ve doğrulama ikiye ayrılmıştır. İlk olarak sahte ve gerçek video sayıları belirlenmiş, ardından train_test_split fonksiyonu ile veri kümesi %80 eğitim ve %20 doğrulama olacak şekilde ayrıştırılmıştır. Her iki alt kümede de gerçek ve sahte videoların dağılımı yeniden hesaplanmıştır. Son aşamada, tüm görüntü boyutları 112×112 pikselde sabitlenmiş ve bu görüntüler için ortalama ve standart sapma değerleri hesaplanmıştır.

Veriler video_dataset sınıfı kullanılarak yüklemiş ve ardından DataLoader aracılığıyla veriler işlem verimliliğini artırmak amacıyla gruplara ayrılmıştır. Son olarak, örnek bir kare görselleştirilerek sistemin test edilmesi sağlanmıştır.

Bir sonraki süreçte, bu tezde çalışma konusu yapay sinir ağı modelleri seçilmişir. Bu modeller, "Materyal ve Metot" bölümünde açıklanan özellikleri ve yetenekleri göz önünde bulundurularak tercih edilmiştir. CNN tabanlı bir model olan ResNeXt50'nin yanı sıra, RNN tabanlı bir model olan LSTM de seçilmiştir. LSTM'ler, hücre durumu (cell state) ve geçitler (gates) aracılığıyla uzun vadeli bağıntıları daha etkili şekilde yönetebildikleri için hafıza yönetimi konusunda avantaj sağlarlar. Ayrıca, LSTM'ler sıralı verileri işleme konusunda daha başarılıdır; 1 000 adımdan daha uzun olan bağıntıları dahi takip edebilirler.

Bu süreçte ResNeXt50 modeli, videolardan gelen kareler üzerinden mekansal özellikleri (spatial features) çıkarır ve tam bağlantılı (fully connected) katman kullanarak sınıflandırma işlemini gerçekleştirir. Sonrasında LSTM modeli, bu karelerin zaman sırasına göre düzenlenmiş dizilerini işler ve zaman bağımlılıklarını yakalar. Ardından LeakyReLU aktivasyon fonksiyonu ve Dropout katmanı uygulanır; nihai sınıflandırma sonuçları ise Linear katman ile üretilir.

Eğitim ve test sürecinde seçilen modeller ön işlemden geçirilmiş veriler üzerinde 20 epoch boyunca eğitime ve teste tabi tutulmuştur. Şekil 4.2'de, LSTM modeli için 20 epoch süresince gerçekleşen eğitim ve test süreci sonunda elde edilen sonuçlardan bir kesit görülmektedir.

```
Epoch 18 - Test Loss: 1.5304, Acc: 55.35%
Predictions: Class 0=315 (59.3%), Class 1=216 (40.7%)
--------------------------------------------------
Epoch [19/20], Train Loss: 0.8706, Acc: 68.58%
Epoch 19 - Test Loss: 1.5490, Acc: 56.66%
Predictions: Class 0=312 (58.8%), Class 1=219 (41.2%)
--------------------------------------------------
Epoch [20/20], Train Loss: 0.8994, Acc: 68.05%
Epoch 20 - Test Loss: 1.5630, Acc: 55.97%
```

Şekil 4.2. LSTM modeli eğitim sürecinden bir kesit

Şekil 4.3, DFDC veri kümesi üzerinde ResNeXt50 modelinin 20 epoch süresince gerçekleşen eğitim sürecini göstermektedir.

```
Epoch [19/20], Train Loss: 1.6320, Acc: 0.4486
Epoch [19] Test Loss: 1.6271, Acc: 0.4480
Epoch [20/20], Train Loss: 1.6223, Acc: 0.4523
```

Şekil 4.3 ResNeXt50 modeli eğitim sürecinden bir kesit

Bu süreç aynen elde edilen hibrid (LSTM/ResNeXt50) modelin üzerinde tekrar yapılmıştır. Şekil 4.4, DFDC veri kümesi üzerinde Hybrid (LSTM/ResNeXt50) modelinin 20 epoch süresince gerçekleşen eğitim sürecini göstermektedir.

```
Test Loss: 0.4765, Acc: 0.8572
Epoch [19/20], Train Loss: 0.0541, Acc: 98.58%
Test Loss: 0.4980, Acc: 0.8515
Epoch [20/20], Train Loss: 0.0383, Acc: 99.41%
Test Loss: 0.5177, Acc: 0.8426
```

Şekil 4.4 Hybrid (LSTM/ResNeXt50) modeli eğitim sürecinden bir kesit

Ayrıca, doğru sınıflandırılan örneklerin yüzdesini görmek ve LSTM ile ResNeXt50 modellerinin performansını karşılaştırmak amacıyla eğitim kaybı (training loss), test kaybı (test loss), eğitim doğruluğu (training accuracy) ve test doğruluğu (test accuracy) hesaplanmıştır.

5. BULGULAR VE DEĞERLENDİRME

Bu bölümde, gerçekleştirilen deneysel çalışmalar sonucunda elde edilen bulgular sunulmakta ve bu bulguların değerlendirilmesi yapılmaktadır. Öncelikle, kullanılan modellerin eğitim ve test süreçlerinden elde edilen performans metrikleri olarak verilmiş, ardından bu sonuçlar karşılaştırmalı olarak tartışılmıştır. Elde edilen çıktılar, derin sahte tespitinde kullanılan farklı yaklaşımların güçlü ve zayıf yönlerini ortaya koymakta ve ilerleyen bölümlerde yapılacak tartışmalar için temel oluşturmaktadır.

5.1. Değerlendirme Ölçütleri

Model uygulandıktan sonra, performanslarının değerlendirilmesi açısından elde edilen sonuçlar dikkate alınmalıdır. Hibrit LSTM/ResNeXt50 modelin performansı, deneysel analizler aracılığıyla değerlendirilmiştir.

Eğitim ve test aşamalarından sonra, modelin performansını görmek amacıyla sonuçların görselleştirilmesi gerçekleştirilmiş ve bu aşamada elde edilen karışıklık matrisinden (confusion matrix) yaralanılmıştır. Karışıklık matrisine ek hem eğitim hem de test setleri için modellerin performansları kayıp (loss) ve doğruluk (accuracy) olmak üzere iki grafik olarak sunulmuştur.

Çalışmanın sonuçlarını karşılaştırmak amacıyla, değerlendirme ölçütleri olarak doğruluk (accuracy), duyarlılık (recall), kesinlik (precision) ve F_1- skoru değerleri performans kriterleri bölümünde verilmiştir. Bu elde edilen değerlere göre, bu alanda yapılmış diğer çalışmaların sonuçları da tartışılmıştır. Değerlendirmeler, standart sınıflandırma metrikleri kullanılarak gerçekleştirilmiştir. TP (gerçek pozitifler), TN (gerçek negatifler), FP (yanlış pozitifler) ve FN (yanlış negatifler) değerleri, karışıklık matrisinden elde edilen model tahminlerine dayanarak tanımlanmıştır.

Doğruluk oranı (accuracy rate), modelin eğitim verileri ile eğitildikten sonra test verileri üzerinde doğru şekilde yaptığı tahminlerin oranını gösterir. Ancak, yalnızca doğruluk oranını dikkate almak bir modelin başarısını değerlendirmek için yeterli olmayabilir; bu nedenle, diğer kriterler de göz önünde bulundurulmalıdır. Bu değerlendirme metrikleri,

modellerimizin doğruluk tahminindeki güçlü ve zayıf yönleri ile yanlış pozitif ve yanlış negatif oranlarını en aza indirme kabiliyetleri hakkında fikir verir. Karışıklık matrisi, modellerin yaptığı gerçek tahmin hatalarının ayrıntılı analizini de sağlar ve bu modellerin pratik derin sahte tespiti görevlerinde ne kadar kullanılabilir olduğunu daha iyi anlamamıza yardımcı olur. Precision, modelin pozitif olarak sınıflandırdığı örneklerden gerçekten pozitif olanların oranıdır. Yüksek kesinlik, daha fazla ilgili (doğru) sonuç üretildiğini ifade eder.

Doğruluk bir modelin doğru tahminlerde bulunma sıklığını ifade eden bir ölçüttür.

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

Kesinlik model tarafından yapılan pozitif tahminlerin ne kadarının gerçekten doğru olduğunu gösteren bir ölçüttür.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \tag{22}$$

Duyarlılık bir modelin mevcut tüm gerçek pozitif örnekler arasından ne kadarını doğru şekilde tespit edebildiğini ölçen bir değerlendirme ölçütüdür.

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \tag{23}$$

F_1 Skoru, modelin doğruluğunu ölçmek için kesinlik ve duyarlılığın harmonik ortalamasıdır. 0 ile 1 arasında bir değer alır. Modelin en iyi performansı F_1 Skoru'nun 1 olmasıyla, en kötü performansı ise 0 olmasıyla gösterilir.

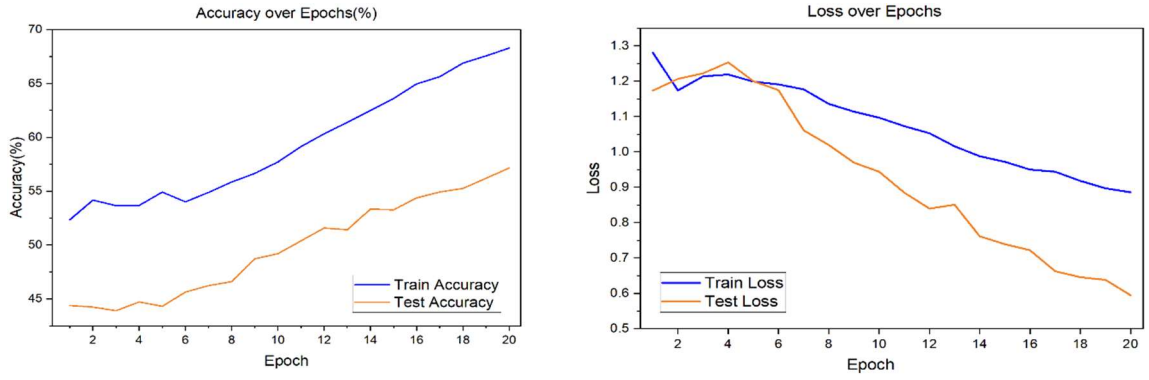
$$F_1\ Skor = 2.\frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{24}$$

5.2. Deneysel Bulgular

CNN'ler mekânsal bilgiyi öğrenirken, uzun-kısa vadeli bellek (LSTM) ağı, derin sahte (deepfake) içeriklerin çok boyutlu yapısını değerlendirmek için zamansal dizileri analiz etmek amacıyla kullanılmıştır. CNN bileşeni, bireysel video karelerinden mekânsal özellikleri çıkarır ve özellikle dokular, kenarlar ve yüz anomalileri (örneğin doğallıktan uzak göz kırpma hareketleri veya cilt dokuları) üzerinde yoğunlaşmaktadır. Mekânsal-zamansal özelliklerin birleştirilmesi bağlamında, CNN mimarisi kullanılarak kare düzeyinde mekânsal desenler çıkarılır; bu desenler arasında anormal yüz dokuları, ışık-gölge tutarsızlıkları ve yüz bölgelerinde şekil bozuklukları yer almaktadır. Ayrıca CNN, gözler ve ağız gibi yüzün belirli alanlarına özgü özelleştirilmiş konvolüsyon filtreleri kullanarak ince yüz anomalilerini de algılayabilmektedir.

Diğer yandan LSTM ağı, 30 ila 50 karelik dizi boyunca zamansal tutarlılığı inceleyerek anormal göz kırpma düzenleri, anormal baş hareketleri, kareler arasında aydınlatmadaki anormallikler gibi dinamik yüz ifadelerini tespit ederek CNN tarafından çıkarılan özellikleri ardışık olarak işleymekte ve zamanla değişen mimikleri modellendirmektedir.

Her iki ağın derin sahte videoları tespit etme performanslarını karşılaştırmak amacıyla Accuracy, Kayıp ve Karışıklık Matrisi gibi değerlendirme metrikleri kullanılmıştır. Şekil 5.1.a'da LSTM modeli için DFDC veri kümesi üzerinde 20 epoch süresince yapılan eğitim ve doğrulama doğruluk grafiği gösterilmektedir.



(a) (b)

Şekil 5.1. LSTM modeli (a) Doğruluk, (b) Kayıp grafiği

Eğitim doğruluğu, modelin eğitim verisi üzerinde ne kadar başarılı olduğunu gösterir. Modelde eğitim doğruluğu yaklaşık %52'den başlayarak %68'in üzerine kadar istikrarlı bir şekilde artmıştır, bu da LSTM modelinin eğitim verisini öğrenme konusunda başarılı olduğunu göstermektedir. Öte yandan, turuncu çizgi modelin doğrulama verisindeki performansını temsil etmektedir. Başlangıçta %44 civarında, yaklaşık 3 ve 5 epoch'larda düşüşler görülse de sonrasında istikrarlı bir şekilde artarak yaklaşık %57'e ulaşmıştır. Genel olarak bakıldığında modelin sürekli artan bir performans sergilediği söylenebilir. Test doğruluğu, eğitim doğruluğu ile beraber artması modelin sadece ezber yapmadığını, aynı zamanda genelleme kabiliyeti kazandığını göstermektedir.

Eğitim ve test eğrilerinin arasındaki fark yaklaşık %8-12 arasındadır. İki eğrinin arasında olan fark model de aşırı uyum kuşkusunu ortaya çıkarmaktadır. Derin sahte modelleri, çoğu zaman eğitim verisinde yer alan belirli yapaylıklar veya üretim tekniklerine aşırı uyum (overfitting) sağlama eğilimindedir. Bu durum, modelin genelleştirilebilir derin sahte tespit özellikleri öğrenmek yerine, veri kümesine özgü örüntülere odaklandığını gösterebilir.

Genel olarak, modelin, artan epok sayısı ile birlikte öğrenme performansını geliştirdiği gözlemlenmektedir. Eğitim ve test doğruluk değerlerinin paralel bir şekilde yükselme eğiliminde olması, eğitim sürecinin etkin biçimde işlediğini göstermektedir. Bununla birlikte, test doğruluğunun yaklaşık %57 seviyesinde kalması, modelin henüz veri kümesinin karmaşıklığını tam anlamıyla yakalayamadığını ve kısmi bir yetersiz öğrenme durumu sergilediğini ortaya koymaktadır. Modelin doğruluk eğrisinde gözlemlenen kademeli ve zaman zaman dalgalı iyileşme, DFDC veri kümesinin doğasından kaynaklanan zorluklarla ilişkili olabilir. DFDC, FaceSwap, Face2Face gibi birden fazla derin sahte üretim yöntemini bünyesinde barındırdığından, modelin tutarlı bir öğrenme gerçekleştirmesi güçleşmektedir. Ayrıca veri kümesinde yer alan sıkıştırma yapaylıkları ve kalite farklılıkları, eğitim sürecinde modele aktarılan sinyallerin tutarlılığını bozarak öğrenme performansını olumsuz yönde etkileyebilmektedir.

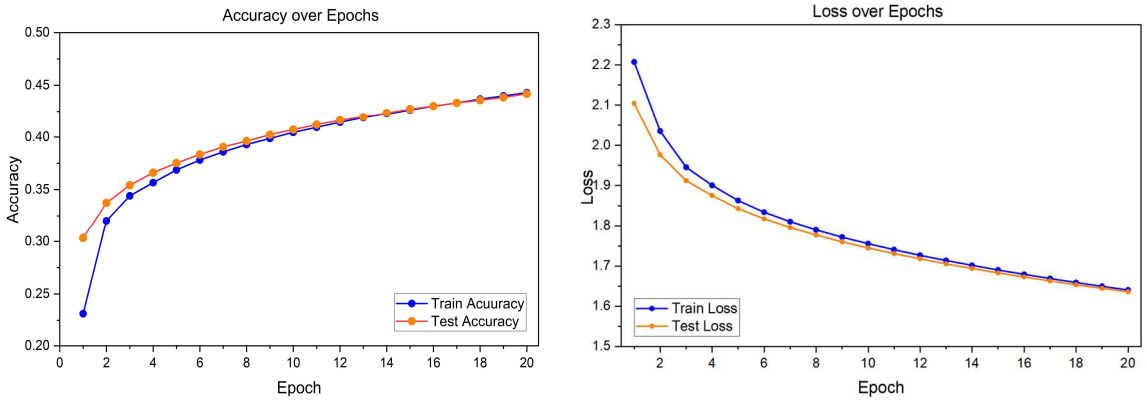
Şekil 5.1.b'de, DFDC veri kümesi üzerinde 20 epoch boyunca LSTM modelinin kayıp grafiğini göstermektedir. Elde edilen kayıp eğrileri incelendiğinde, eğitim kaybının yaklaşık 1,28 seviyesinden başlayarak epoch ilerledikçe düzenli bir biçimde azaldığı ve 20 epokta yaklaşık 0,88 değerine azaldığı görülmektedir. Benzer şekilde test kaybı da başlangıçta 1.18 seviyelerinde olup daha keskin bir düşüş göstermiş ve 20 numaralı epokta 0,60 seviyesine

kadar gerilemiştir. Her iki kayıp değeri de 1–5 epoklar arasında dalgalanma göstermekte, hatta başlangıçta test kaybı eğitim kaybının üzerine çıkmaktadır. Bu durum, modelin DFDC'nin karmaşık veri dağılımı içerisinde istikrarlı öğrenme örüntülerini bulabilmesi için zamana ihtiyaç duyduğunu göstermektedir. Özellikle 6'inci epoktan itibaren test kaybının eğitim kaybından daha düşük seyretmesi dikkat çekici bir durumdur. Ayrıca test veri kümesinin dağılımının görece daha kolay örnekler içermesi de bu duruma katkıda bulunabilir.

Her iki eğrinin de sürekli olarak düşüş eğiliminde olması, modelin öğrenme sürecinin etkin bir şekilde işlediğini ve herhangi bir aşırı uyum (overfitting) belirtisinin ortaya çıkmadığını göstermektedir. Eğitim ve test doğruluk eğrilerinde gözlemlenen artışla birlikte düşünüldüğünde, bu sonuçlar modelin ezberleme eğiliminden ziyade genelleştirilebilir özellikler öğrendiğine işaret etmektedir.

LSTM modeli DFDC veri kümesinin üzerinde, elde edilen Karışıklık Matrisinden hesaplanan sonuçları ve performans değerlendirme metrikleri:

Kesinlik = %62,55
Duyarlılık = %81,44
Doğruluk = %57,27
F_1 Skor = %67



(a) (b)

Şekil 5.2. ResNeXt50 Modeli (a) Doğruluk, (b) Kayıp grafiği

Şekil 5.2.a, DFDC veri kümesi üzerinde 20 epoch boyunca ResNeXt50 modelinin doğruluk oranını göstermektedir. Eğitim ve test doğruluk eğrileri incelendiğinde, modelin ilk epoklarda hızlı bir öğrenme süreci sergilediği, ardından doğruluk değerlerindeki artışın daha istikrarlı ve kademeli bir şekilde devam ettiği gözlemlenmektedir. İlk 1–4 epok arasında test doğruluk değerlerinin eğitim doğruluklarından daha yüksek gerçekleşmesi, eğitim verisine uygulanan düzenlileştirme (ör. veri artırma veya dropout) yöntemlerinin modeli başlangıçta daha zor bir öğrenme sürecine yönlendirmesi ya da test veri kümesinin görece daha basit örnekler içermesi ile açıklanabilir. 5'inci epoktan itibaren her iki eğri birbirine yakın seyretmekte ve 20 numara epok sonunda yaklaşık %44 doğruluk seviyesine ulaşmaktadır.

Eğitim ve test doğruluklarının birbirine oldukça yakın olması, modelin aşırı uyum (overfitting) sergilemediğini ve genelleme kabiliyetinin güçlü olduğunu göstermektedir. Bununla birlikte, elde edilen nihai doğruluk değerlerinin düşük seviyede kalması, modelin veri setinin karmaşıklığını tam anlamıyla yakalayamadığını ve kısmi bir yetersiz öğrenme (underfitting) durumunun söz konusu olduğunu ortaya koymaktadır.

Şekil 5.2.b, DFDC veri kümesi üzerinde ResNeXt50 modelinin 20 epoch boyunca eğitim ve test kaybı grafiğini göstermektedir.

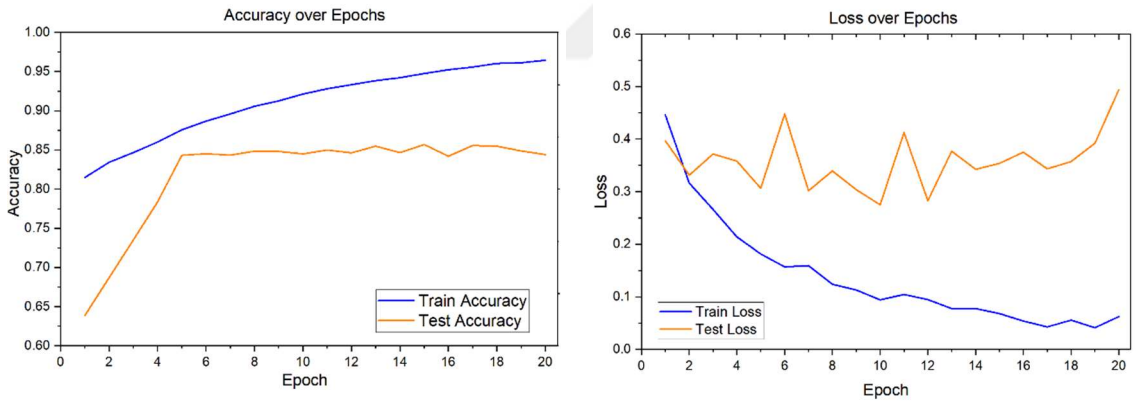
Kayıp eğrileri incelendiğinde hem eğitim hem de test kayıplarının ilk epoklardan itibaren düzenli bir biçimde azalma eğilimi gösterdiği görülmektedir. Eğitim kaybı yaklaşık 2,2 seviyesinden başlayarak 20 numara epok sonunda 1,63 değerine gerilemiştir. Benzer şekilde test kaybı da başlangıçta yaklaşık 2,1 düzeyinde olup benzer bir seyir izlemiş ve 20 numara epokta eğitim kaybına oldukça yakın bir değer olan 1,63 seviyesine ulaşmıştır.

Eğitim ve test kayıpları arasındaki farkın tüm süreç boyunca oldukça düşük seviyede kalması, modelin aşırı uyum (overfitting) sergilemediğini ve genelleme kabiliyetinin güçlü olduğunu ortaya koymaktadır. İlk 5 epokta kayıplardaki hızlı düşüş, modelin temel örüntüleri kısa sürede öğrenebildiğini; sonraki epoklarda ise azalmanın daha kademeli seyretmesi, modelin performans açısından bir plato seviyesine yaklaşmaya başladığını göstermektedir. Bununla birlikte, nihai kayıp değerlerinin görece yüksek kalması (~1,63), modelin veri kümesindeki (DFDC) karmaşıklığı tam olarak yakalayamadığını ve kısmi bir yetersiz öğrenme (underfitting) durumu sergilediğini düşündürmektedir.

Ikili sınıflandırma (binary classification) problemi için yaklaşık 1,65 seviyesinde bir kayıp değeri, model tahminlerinin %50–%50 olasılıklara yakın gerçekleştiğini göstermektedir. 20 epok boyunca modelin kaybı yalnızca yaklaşık 0,5 oranında azaltabilmiş olması, ayırt edici özellikleri öğrenmede güçlük yaşadığını ortaya koymaktadır. 1,65 seviyesindeki kayıp değeri, genellikle %50 güven düzeyinde tahminlere karşılık gelmekte olup, bu durum modelin neredeyse rastgele sınıflandırmaya yakın bir performans sergilemesini açıklamaktadır.

ResNeXt50 modelini DFDC veri kümesinin üzerinde uyguladığımızda Karışıklık Matrisi sonuçlarından alınan sonuçlar:

Kesinlik =%44,18
Duyarlılık = %42,2
Doğruluk = %44,53
F_1Skor =%44,2



(a) (b)

Şekil 5.3. Hybrit (ResNeXt50/LSTM) modeli (a) Doğruluk, (b) Kayıp grafiği

Şekil 5.3.a da eğitim doğruluğu, birinci epokta yaklaşık %82 seviyesinde başlamış ve yirminci epoka kadar %97 seviyesine ulaşmıştır. Bu istikrarlı artış eğilimi, modelin eğitim verilerinden etkili bir şekilde öğrenmekte olduğunu ve durağanlaşma göstermediğini ortaya koymaktadır. Test doğruluğu ise yaklaşık %64 seviyesinde başlamış ve beşinci epoka kadar %85 düzeyine yükselmiştir; bu durum, modelin başlangıçta iyi bir genelleme kabiliyeti sergilediğini göstermektedir. Beşinci epoktan sonra test doğruluğu yaklaşık %85 seviyesinde plato yapmış ve yirminci epoka kadar küçük dalgalanmalarla bu seviyede

kalmıştır. Her iki eğri dikkate alındığında, modelin derin sahte ve gerçek video karelerini ayırt etmede faydalı örüntüleri yakaladığı, özellikle erken dönemlerdeki hızlı artışla gözlemlenmektedir. Bununla birlikte, test doğruluğundaki plato, modelin kapasitesi veya düzenlileştirme yöntemlerinin genelleme yeteneğini sınırladığını göstermektedir. DFDC veri kümesi; çeşitliliği, sıkıştırma artefaktları ve gerçek dünya varyasyonları nedeniyle zorlu bir yapıya sahiptir ve yaklaşık %84–85 seviyelerinde plato görülmesi, temel yöntemlerde yaygın bir durumdur.

Şekil 5.3.b de eğitim kaybı (Training Loss), birinci epokta yaklaşık 0,44 seviyesinde başlamış ve epoklar boyunca tutarlı ve düzenli bir şekilde azalmıştır. Bu durum, modelin eğitim verilerini giderek daha iyi uyumladığını ve durağanlaşma göstermeden optimizasyonunu sürdürdüğünü ortaya koymaktadır. Eğitim kaybının ~0,45 seviyesinden ~0,05 seviyesine kadar sağlıklı ve düzenli bir düşüş göstermesi, modelin eğitim verileri üzerinde başarılı bir şekilde optimize olduğunu kanıtlamaktadır.

Test kaybı ise birinci epokta eğitim kaybına benzer şekilde yaklaşık 0,39 seviyesinde başlamıştır. İlk olarak 3–5. epoklar arasında hafif bir düşüş göstermiş, sonrasında ise 0,28 ile 0,45 aralığında dalgalanmıştır. Test kaybı ile eğitim kaybı arasındaki ayrışma (divergence) 5. epoktan sonra başlamaktadır. Bu durum, aşırı uyum (overfitting) belirtisi olup, model eğitim örnekleri üzerinde optimizasyonunu sürdürse de DFDC veri kümesindeki görülmemiş veriler için genelleme kabiliyetinin sınırlı kaldığını göstermektedir.

Model, güçlü bir kapasite sergileyerek (düşük eğitim kaybı ~0,05) eğitim verilerini iyi öğrenmiş olsa da test performansının istikrarını düzenleyici (regularization) mekanizmalar açısından etkili olamamaktadır. Test kaybındaki dalgalanmalar, veri kümesindeki çeşitliliklere (farklı sıkıştırma seviyeleri, kimlikler ve manipülasyon türleri) karşı duyarlılığı işaret etmektedir. Bununla birlikte, modelin erken evrelerde hızlı bir şekilde faydalı özellikleri öğrendiği; ancak 7. epoktan sonraki ek eğitim sürecinin düzenlileştirmeden ziyade ezberlemeyi (memorization) geliştirdiği söylenebilir.

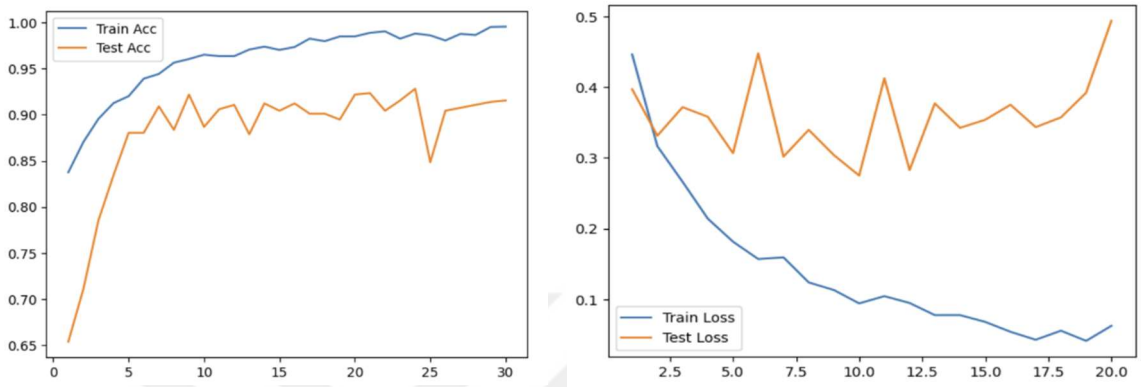
Hybrid (LSTM/ResNeXt50) modelini DFDC veri kümesinin üzerinde uyguladığımızda Karışıklık Matrisi (Confusion Matrix) sonuçlarından alınan sonuçlar:

Kesinlik = %91,85

Duyarlılık =%90,68

Doğruluk =%85,26

F_1 Skor =%91,1



(a) (b)

Şekil 5.4. Hybrid (ResNeXt50/LSTM) Modelinin 30 ekoka kadar (a) Doğruluk, (b) Kayıp grafiği

Şekil 5.4.a da 30 epoch boyunca eğitilen model, başlangıç aşamalarında hem eğitim hem de test doğruluğunda hızlı bir artış göstermiştir. Yaklaşık 10 numaralı epoch'tan itibaren eğitim doğruluğu %100'e yaklaşırken test doğruluğu %90 civarında sabitlenmiştir. 20 epoch'luk eğitime kıyasla daha yüksek doğruluk değerleri elde edilmiş olsa da, test eğrisinde gözlemlenen dalgalanmalar ve ani düşüşler, modelin eğitim verisine aşırı uyum (overfitting) eğiliminde olduğunu göstermektedir. Eğitim ve test performansı arasındaki giderek artan genelleme farkı da bu durumu desteklemektedir; zira model, eğitim verisini neredeyse mükemmel şekilde öğrenirken, görülmemiş veriler üzerinde istikrarsız sonuçlar üretmektedir. Buna karşılık, 20 epoch süren eğitimde test performansı daha dengeli ve istikrarlı bir seyir izlemiş, ancak ulaşılan en yüksek doğruluk %85 seviyesinde kalmıştır. Dolayısıyla, 30 epoch'luk model daha yüksek performans sağlamasına karşın, bunu istikrar ve genelleme pahasına gerçekleştirmekte, 20 epoch'luk model ise daha dengeli ve tutarlı sonuçlar sunmaktadır.

20 ve 30 epoch için elde edilen kayıp (loss) eğrileri, modelin genelleme performansına ilişkin önemli ipuçları sunmaktadır. Şekil 5.4.b de 30 epoch boyunca yapılan eğitimde, eğitim kaybı düzenli bir şekilde azalarak çok düşük bir seviyeye (~0,05) ulaşmış ve modelin

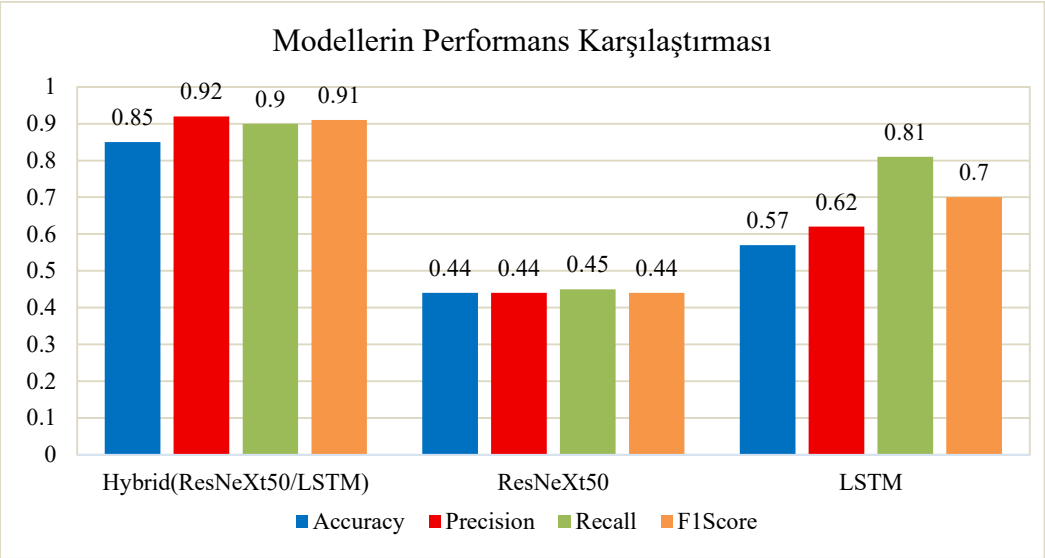
eğitim verisini neredeyse mükemmel biçimde öğrendiğini göstermiştir. Ancak test kaybı 0,3–0,45 aralığında dalgalanmalar göstermiş ve eğitimin sonunda yaklaşık 0,5 seviyesine kadar yükselmiştir. Eğitim ve test kaybı arasındaki bu ayrışma, modelin eğitim verisine aşırı uyum (overfitting) sergilediğini ve görülmemiş verilerdeki başarımının zayıfladığını açıkça ortaya koymaktadır. Buna karşılık, 20 epoch süren eğitimde eğitim kaybı ~0,1 seviyesine kadar azalmış, test kaybı ise daha istikrarlı bir seyir izlemiş ve 0,3–0,4 aralığında dalgalanmakla birlikte 30 epoch'ta gözlenen keskin artışı göstermemiştir. 20 epoch'luk model, 30 epoch'luk modele kıyasla daha yüksek bir eğitim kaybı ile sonlansa da, daha dar bir genelleme farkı koruyarak test performansında daha güvenilir sonuçlar sunmuştur. Dolayısıyla karşılaştırma, 30 epoch eğitiminin eğitim verisine uyumu artırmasına karşın istikrar ve genelleme açısından olumsuzluk yarattığını; 20 epoch'un ise daha dengeli bir sonuç sağladığını göstermektedir.

ResNeXt50 modelinde eğitim ve test doğruluk değerleri birbirine oldukça yakın seyretmiş, 20'uncu epoch sonunda yaklaşık %44 seviyesine ulaşmıştır. Eğitim ve test eğrileri arasında ciddi bir ayrışma olmaması, modelin genelleme kapasitesinin yüksek olduğuna işaret etse de düşük doğruluk değerleri modelin karmaşık örüntüleri yakalayamadığını ve aşırı basit/kapasitesi yetersiz bir yapıya sahip olduğunu göstermektedir. Bu nedenle, pratik uygulamalar için yeterli başarı düzeyine ulaşamamıştır.

LSTM modelinde, kayıp eğrileri incelendiğinde, eğitim kaybının düzenli biçimde azalmasına karşın test kaybının artış göstermesi dikkat çekmektedir. Bu sonuç, ikinci modelde gözlenen aşırı uyum eğilimini doğrulamakla birlikte, test doğruluk değerleri birinci modele kıyasla yine daha yüksek seviyededir.

Hybrid (ResNeXt50/LSTM) modeli ise, eğitim süreci boyunca doğruluk değerlerinde istikrarlı bir artış sergilemiş, 20 numaralı epoch sonunda yaklaşık %97 eğitim doğruluğuna ulaşmıştır. Test doğruluk değeri daha sınırlı kalmış olmakla birlikte, %85 seviyesinde sabitlenerek birinci modele kıyasla oldukça üstün bir performans göstermiştir. Bu durum, modelin eğitim verisini etkin bir şekilde öğrenebildiğini ve derin sahtecilik tespitinde yüksek ayırt edici kapasiteye sahip olduğunu kanıtlamaktadır. Her ne kadar eğitim ve test doğrulukları arasında yaklaşık %12–13'lük bir fark bulunsa da, hibrit modelin eğitim verilerini etkin bir biçimde öğrenebildiği görülmektedir. Bununla birlikte, söz konusu fark modelin DFDC veri kümesindeki görülmemiş örnekler üzerinde genelleme kapasitesinin

sınırlı kaldığını işaret etmektedir. Ancak, literatürde mevcut olan yöntemler göz önünde bulundurulduğunda, tezde önerilen hibrit modelin elde ettiği genel performansın tahmin edici ve rekabetçi bir düzeyde olduğu söylenebilir.



Şekil 5.5. LSTM ve ResNeXt50 ve Hybrid modellerin performans karşılaştırması

Şekil 5.5, üç farklı modelin (Hybrid LSTM/ResNeXt50, ResNet ve LSTM) Kesinlik (Precision), Duyarlılık (Recall), Doğruluk (Accuracy) ve F_1skoru metrikleri bakımından karşılaştırmalı performansını göstermektedir. Görüldüğü üzere, Hybrid (LSTM/ResNeXt50) modeli tüm metriklerde en yüksek değerleri elde etmiş ve özellikle Kesinlik (%91,85), Duyarlılık (%90,68), Doğruluk (85,26%) ve F_1 skoru (%91,10) oranlarıyla diğer modellere kıyasla üstün bir performans sergilediği gözlemlenmiştir. Bu sonuç, hibrit mimarinin hem uzamsal hem de zamansal özellikleri etkili biçimde yakalayarak derin sahtecilik tespitinde daha başarılı olduğunu ortaya koymaktadır. Hibrit modelin eğitim ve test verileri için elde edilen bazı rastgele doğruluk sonuçları ile eğitim ve test verilerine ait kayıp sonuçları, ayrıca modelin çeşitli rastgele çalıştırmalarına ilişkin çıktılar EK1, EK2 ve EK3'te sunulmuştur.

ResNeXt50 modeli tüm metriklerde yaklaşık %44 seviyelerinde kalarak en düşük performansı göstermiştir. Bu durum, ResNet'in tek başına DFDC veri kümesindeki sahtecilik artefaktlarını ayırt etmede yetersiz kaldığını düşündürmektedir. LSTM modeli, ResNeXt50'e kıyasla daha iyi sonuçlar üretmiş ve özellikle Duyarlilik (%81,44**)** değerinde güçlü bir performans göstermiştir. Ancak, doğruluk (%57,27) ve F_1 skoru (%69,95) bakımından Hybrid modele kıyasla geride kalmıştır. Genel olarak değerlendirildiğinde,

Hybrid (ResNeXt50/ LSTM) diğer iki modele kıyasla hem doğruluk hem de F_1 skor bakımından DFDC veri kümesi üzerinde en güçlü performansı sunmaktadır.

Çizelge 5.1. Farklı modellerin performans değerleri

Model Adı	*F1*-Score	Kesinlik	Duyarlılık
Optical Flow+RNN	%29,14	%24,96	-
OF+RNN+CNN	%66,51	%67,11	-
Optical Flow+CNN	%68,99	%69,36	-
Resnet	%71,00	%69,80	
ResNet-50	%75,10	%74,10	
Mobile Net V2	%79,20	%79,80	-
Efficient Net-B5	%86,82	%70<x<%80	-
VGG	%86,90	%84,60	
VGG16	%87.00	%86,30	%90
Efficient Net-Vision Transformer	%88	%70<x<%80	-
Xception	%88,60	%87,70	
LSTM/ResNeXt-50	%91,10	%92	90%

Çizelge 5.1. de, derin sahte tespitine yönelik çeşitli derin öğrenme modellerinin performansları F_1skor ve Precision metrikleri açısından sunulmaktadır. Elde edilen sonuçlara göre, ResNeXt50/ LSTM hibrit modeli %91,10 F_1 skoru ve %92 kesinlik değeri ile tüm diğer modellerden daha yüksek performans göstermiştir. EfficientNet-B5, Xception ve VGG16 gibi güçlü modeller de yüksek başarı oranları sergilemiş olsa da, önerilen hibrit modelin ulaştığı değerler, mekânsal ve zamansal özellik çıkarımının birleştirilmesinin performansa önemli katkılar sağladığını ortaya koymaktadır. Bu bulgu, önerilen hibrit yaklaşımın derin sahte tespitinde daha güvenilir ve dengeli sonuçlar sunduğunu göstermektedir.

Erken dönem DFDC temel modelleri (baselines) ve bunları izleyen pek çok çalışma, eşik bağımsız AUC değerlerini rapor etmiş; FF++'tan DFDC'nin daha zorlu, "gerçek yaşam" dağılımına geçildiğinde ise eşiklendirilmiş F_1Skoru ve doğruluk değerlerinin görece sınırlı kaldığını göstermiştir. Özgün DFDC makalesi ve devamındaki analizler, DFDC üzerinde sağlam bir genellemenin güç olduğunu vurgular; FF++'ta güçlü sonuçlar veren birçok model DFDC'de belirgin biçimde performans kaybı yaşamaktadır [71].

Son yıllarda DFDC (veya Kaggle-DFDC) üzerinde eğitilen tek akışlı (single-stream) CNN/ViT boru hatlarında F_1 Skoru çoğunlukla %80'lerin ortasında; kesinlik ve duyarlılık ise genellikle orta-yüksek %80'ler düzeyinde raporlanmaktadır. Örneğin Hasan (2025), Kaggle-DFDC üzerinde EfficientNet-B5'i ince ayarlayarak %86,82 F_1 skoru elde ettiğini bildirmektedir. Çalışmada kesinlik ve duyarlılık değerlerinin her ikisi de %90 değeri altındadır [72]. Buna karşılık, tezde önerilen hibrit modelimizin F_1 Skoru %91,10 ve kesinlik değeri %91,85 olup söz konusu eşiği aşmakta; dağıtıma alınmaya elverişli bir işletim noktasında yalancı-pozitif ve yalancı-negatifler arasında daha iyi bir dengeye sahip olduğuna işaret etmektedir.

DFDC'yi içeren daha geniş karşılaştırmalı değerlendirmeler, klasik CNN tabanlı modellerin (Xception/ResNet/VGG) saygın doğruluk düzeylerine ulaşsa da DFDC ve FF++ üzerinde ortak metriklerle ayarlandıklarında modelimizin F_1 Skoru/kesinlik değerlerine kıyasla tipik olarak daha düşük kaldığını göstermektedir; bu durum, FF++'a kıyasla DFDC'de süregelen genelleme boşluklarını da görünür kılmaktadır. Tezde önerilen modelin kesinlik ve duyarlılık değerleri (%91,85 ve %90,68), benzer tek akışlı görsel kurulumlarda bu standart omurgaların DFDC üzerinde çoğunlukla eriştiği seviyelerin [73] üzerindedir.

Yöntemsel çalışmalar ve meta-karşılaştırmalar şu saptamaları güçlendirmektedir: (i) DFDC, FF++ gibi özenle derlenmiş veri kümelerinden daha zorludur; (ii) birçok çalışma F_1 Skoru/kesinlik/duyarlılık yerine AUC bildirmektedir; (iii) bire bir, yeniden üretilebilir karşılaştırmalar seyrektir. F_1 Skoru raporlandığında da, topluluk/ensemble veya işitsel-görsel füzyon kullanılmadan tek akışlı görsel boru hatlarında DFDC üzerinde ≥%90 F_1 Skoru değerlerinin nispeten ender görüldüğü anlaşılmaktadır. Bu bağlam, tezde önerilen modelin F_1 Skoru ve kesinlilik değerlerinin (%91,10 ve %91,85) DFDC'de tipik olarak bildirilen aralıkların üzerinde olduğu yönündeki iddiayı güçlendirmektedir [74].

Veri kümeleri arası sağlamlığa (cross-dataset robustness) odaklanan son çalışmaların bir bölümü, DFDC test alanı olarak kullanıldığında, ileri mimariler için dahi DFDC AUC değerlerinin %70'lerin ortası ile %80'lerin başı aralığında kaldığını bildirmektedir [75]. Bu bulgu, DFDC'nin zorluğunu bir kez daha ortaya koymakta ve tezde önerilen modelin modelin %90'ın üzerindeki F_1 Skoru/kesinlik değerleriyle, sabit bir işletim noktasında güçlü tek akışlı algılayıcılar arasında yer aldığını göstermektedir.

DFDC temelli, tek akışlı görsel algılayıcılarda F_1 Skoru/kesinlik/duyarlılık raporlayan çalışmalarla karşılaştırıldığında, modelimizin %91,10 F_1 Skoru ve %91,85 kesinlik, değerleri literatürde sıkça görülen (çoğu zaman %80'lerin ortalarında kalan) sonuçlardan yüksektir. Önerilen modelin %90,68 olan duyarlılığı da güçlü ve dengelidir. Yanlış pozitiflerin maliyetli olduğu bağlamlarda F_1 Skoru ve kesinlik tercih edilen metriklerdir.

6. SONUÇ

Bu çalışmada, DFDC veri kümesi üzerinde derin sahte tespiti amacıyla ResNeXt50, LSTM ve hibrit ResNeXt50/LSTM mimarileri kapsamlı bir biçimde karşılaştırılmıştır. Bulgular, yalnızca mekânsal özelliklere odaklanan ResNeXt50 mimarisinin, DFDC veri kümesine özgü karmaşık artefaktları yakalamada sınırlı kaldığını ortaya koymuştur. Öte yandan, zamansal bağımlılıkları modelleme kapasitesine sahip olan LSTM mimarisi daha başarılı sonuçlar elde etmiş, ancak doğruluk ve kesinlik gibi metriklerde denge sağlamada yetersiz kalmıştır.

Bu bağlamda geliştirilen hibrit ResNeXt50/LSTM modeli, mekânsal ve zamansal özellik çıkarımlarını bir araya getirerek tüm performans metriklerinde en yüksek değerlere ulaşmıştır. Model, %91,10 F_1skoru ve %92 kesinlik değeri ile hem ResNeXt50 hem de LSTM tabanlı yaklaşımları geride bırakmış, aynı zamanda literatürde sıkça kullanılan EfficientNet, Xception ve VGG16 gibi güçlü mimarilerden de daha iyi sonuçlar üretmiştir. Bu durum, derin sahte tespitinde tek boyutlu yaklaşımların sınırlılıklarını aşmak için hibrit mimarilerin önemli avantajlar sunduğunu göstermektedir.

Elde edilen bulgular, hibrit mimarilerin yalnızca daha yüksek performans sağlamadığını, aynı zamanda sahte içeriklerin giderek daha gerçekçi hale geldiği günümüzde daha güvenilir bir çözüm oluşturduğunu ortaya koymaktadır. Gelecekte yapılacak araştırmalarda, bu yaklaşımın daha verimli hale getirilmesi, farklı veri kümeleri üzerinde genelleme kabiliyetinin test edilmesi ve çok modlu (multimodal) entegrasyon yöntemleriyle desteklenmesi, sahtecilik yöntemlerinin evrimleşen yapısına karşı daha dayanıklı tespit sistemlerinin geliştirilmesine katkı sağlayacaktır.

KAYNAKLAR

1. İnternet: Somers, M., (July 2020). Deepfakes, explained. *MIT Sloan, 9*. Web: https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained / Erişim Tarihi: 26.09.2025.

2. Mahdizadehaghdam, S., Panahi, A., Krim, H. (2019). *Sparse Generative Adversarial Network*. International Conference on Computer Vision Workshop, Seoul, 3063-3071.

3. İnternet: Regula and Sapio Research. (2024). *Deepfake Trends 2024*. Web: https://regulaforensics.com/resources/deepfake-trends-2024-report Erişim Tarihi: 26.09.2025.

4. Kalpokas, I., Kalpokiene, J. (2022). *Deepfakes: A Realistic Assessment of Potentials, Risks, and Policy Regulation* (1). Zurich: Springer International Publishing, 87.

5. İnternet: Kishore, A. (May 2019). What is a Deepfake and How Are They Made?. *Online Tech Tips*, 1. Web: https://www.online-tech-tips.com/what-is-a-deepfake-and-how-are-they-made / Erişim Tarihi: 26.09.2025

6. Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., Liu, Y. (2022). Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision*, 130(7), 1678–1734.

7. Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C. (2024). Deepfake Generation and Detection: A Benchmark and Survey. *Computer Vision and Pattern Recognition,* 1(4), 1-28.

8. Sharma, V. K., Garg, R., Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. *Multimedia Tools and Applications*, 84(6), 22187–22229.

9. Stroebel, L., Llewellyn, M., Hartley, T., Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, 7(2), 83–113.

10. Rana, M. S., Nobi, M. N., Murali, B., Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*, 10(1), 25494–25513.

11. Gupta, G., Raja, K., Gupta, M., Jan, T., Whiteside, S. T., Prasad, M. (2024). A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *MDPI Electronics*, 13(1), 1–27.

12. Naitali, A., Ridouani, M., Salahdine, F., Kaabouch, N. (2023). Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. *Computers*, 12(10), 1–26.

13. Passos, L. A., Jodas, D., Costa, K. A. P., Souza Júnior, L. A., Rodrigues, D., del Ser, J., Camacho, D., Papa, J. P. (2024). A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, 41(8), 13570.

14. Rana, M. S., Nobi, M. N., Murali, B., Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*, 10(2), 25494–25513.

15. Abbas, F., Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 238, 124260.

16. İnternet: Rossler, A. (2018). FaceForensics++: Learning to Detect Manipulated Facial Images. *Github*. Web: https://justusthies.github.io/posts/faceforensics++ / Erişim Tarihi: 26.09.2025.

17. Lakshmi, B. H., Rıshıka, V., Chandana, S. S., Sree, N. B. (2024). Deepfake Detectıon A Systematıc Lıterature Revıew. *International Journal of Engineering, Science and Advanced Technology*, 24(10), 214–222.

18. Gong, L. Y., Li, X. J. (2024). A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. *Multidisciplinary Digital Publishing Institute,* 13(3), 565-585.

19. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C. C. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. *Arxiv,* 2(1), 1-4.

20. Shifa Mohd Imran, S. and Devendra Tawde, P. (2024). Deepfake Detection: A Literature Review. *International Research Journal of Engineering and Technology*, 11(3), 1-6.

21. İnternet: Li, Y. Yang, X. Sun, P. Qi, H. and Lyu, S. (2024). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. Web: https://deepfakedetectionchallenge.ai / Erişim Tarihi: 26.09.2025.

22. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, U. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, 5(3) 2382–2390.

23. internet: Yan, Z. Zhang, Y. Yuan, X. Lyu, S. and Wu, B. (2014). DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. *Fixed alignment on faceshifter, 1.* Web: https://github.com/ondyari/FaceForensics / Erişim Tarihi: 26.09.2025.

24. Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, 7(2), 83–113.

25. Raza, A., Munir, K., Almutairi, M. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences*, 12(19), 1-7.

26. Patarlapati, N. (2023). Unmasking Reality: Exploring the Sociological Impacts of Deepfake Technology. *International Journal for Research in Applied Science and Engineering Technology*, 11(11), 882–889.

27. Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy and Technology*, 34(4), 623–643.

28. Garcia-Murillo, M., MacInnes, I. (2025). Economic, political, and social factors impeding the regulation of digital platforms. *Telecommunications Policy*, 49(3), 1-102915.

29. Eberl, A., Kühn, J., Wolbring, T. (2022). Using deepfakes for experiments in the social sciences- A pilot study. *Frontiers in Sociology*. 7(1), 1-907199.

30. Finkelhor, D., Turner, H., Colburn, D., Mitchell, K., Mathews, B. (2023). Child sexual abuse images and youth produced images: The varieties of Image-based Sexual Exploitation and Abuse of Children. *Child Abuse and Neglect*, 143(June), 106269.

31. Sunkari, V., Nagesh, A. S. (2024). *Artificial intelligence for deepfake detection: systematic review and impact analysis*. IAES International Journal of Artificial Intelligence, 13(4), 3786–3792.

32. Walsh, K., Mathews, B., Parvin, K., Smith, R., Burton, M., Nicholas, M., Napier, S., Cubitt, T., Erskine, H., Thomas, H. J., Finkelhor, D., Higgins, D. J., Scott, J. G., Flynn, A., Noll, J., Malacova, E., Le, H., Tran, N. (2025). Prevalence and characteristics of online child sexual victimization: Findings from the Australian Child Maltreatment Study. *Child Abuse and Neglect*, 160(August 2024), 107186.

33. Finkelhor, D., Turner, H., Colburn, D. (2023). Which dynamics make online child sexual abuse and cyberstalking more emotionally impactful: Perpetrator identity and images? *Child Abuse and Neglect*, 137(August 2022), 106020.

34. Fallis, D. (2021b). The Epistemic Threat of Deepfakes. *Philosophy and Technology*, 34(4), 623–643.

35. Garcia-Murillo, M., MacInnes, I. (2025). Economic, political, and social factors impeding the regulation of digital platforms. *Telecommunications Policy*, 49(3), 102915.

36. Patarlapati, N. (2023). Unmasking Reality: Exploring the Sociological Impacts of Deepfake Technology. *International Journal for Research in Applied Science and Engineering Technology*, 11(11), 882–889.

37. Vaccari, C., Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media and Society*, 6(1), 1-12.

38. Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39–52.

39. Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics*, 57(May 2020).

40. Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, 182(February), 111074.

41. Mirza, M., Osindero, S. (2014). Conditional Generative Adversarial Nets. *Arxiv,* 10(1), 1–7.

42. Radford, A., Metz, L., Chintala, S. (2026). *Unsupervised representation learning with deep convolutional generative adversarial networks*. 4th International Conference on Learning Representations, Rio de Janeiro, 1–16.

43. Zhu, J. Y., Park, T., Isola, P., Efros, A. A. (2017). *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.* Proceedings of the IEEE International Conference on Computer Vision, Venice, 2242–2251.

44. Yang, Z., Zhan, F., Liu, K., Xu, M., Lu, S. (2023). AI-Generated Images as Data Source: The Dawn of Synthetic Era. *Computer Vision and Pattern Recognition*, 12 (2), 1–20.

45. Tian, C., Zhang, X., Lin, J. C.-W., Zuo, W., Zhang, Y., Lin, C.-W. (2022). Generative Adversarial Networks for Image Super-Resolution: A Survey. *ACM Computing Surveys*, 3(1).

46. İnternet: From Wikipedia contributors. (n.d.). (2019) Generative adversarial network. *Wikipedia,* Web: https://en.wikipedia.org/wiki/Generative_adversarial_network / Erişim Tarihi: 26.09.2025.

47. İnternet: From amazon information. (2024). What is a GAN?. *Amazon AWS*. Web: https://aws.amazon.com/what-is/gan / Erişim Tarihi: 26.09.2025.

48. Simonyan, K., Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. 3rd International Conference on Learning Representations, ICLR 2015- Conference Track Proceedings, San Diego,1–14.

49. Maerten, A.-S., Soydaner, D. (2023). From paintbrush to pixel: A review of deep neural networks in AI-generated art. *Arxiv*, 5(1), 1–17.

50. He, K., Zhang, X., Ren, S., Sun, J. (2016). *Deep residual learning for image recognition*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, 770–778.

51. Ebrahimi, M. S., Abadi, H. K. (2021). *Study of Residual Networks for Image Recognition*. Intelligent Computing- Proceedings of the 2021 Computing Conference, Stanford, 754–763.

52. Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T. Y., Shlens, J., Zoph, B. (2021). Revisiting ResNets: Improved Training and Scaling Strategies. *Advances in Neural Information Processing Systems*, 27(NeurIPS), 22614–22627.

53. Moreno-Lozano, M. I., Ticlavilca-Inche, E. J., Castañeda, P., Wong-Durand, S., Mauricio, D., Oñate-Andino, A. (2024). A Performance Evaluation of Convolutional Neural Network Architectures for Pterygium Detection in Anterior Segment Eye Images. *Diagnostics*, 14(18), 2-14.

54. TiclavilcaInche, E. J., Isabel, M. M.-L., Castañeda, P., Wong-Durand, S., Alejandra, Oñate-Andino. (2023). Online and Biomedical Engineering. *International Journal of Online and Biomedical Engineering*, 19(9), 122–130.

55. Choi, M. K., Jung, H. (2021). Development of Fast Refinement Detectors on AI Edge Platforms. *Notes in Computer Science*, 12664(1), 592–606.

56. Kolekar, S., Gite, S., Pradhan, B., Alamri, A. (2024). Predicting Vehicle Pose in Six Degrees of Freedom from Single Image in Real-World Traffic Environments Using Deep Pretrained Convolutional Networks and Modified Centernet. *International Journal on Smart Sensing and Intelligent Systems*, 17(1), 1-6.

57. İnternet: From Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. (2016). Aggregated Residual Transformations for Deep Neural Networks. *Arxiv*, Web: http://arxiv.org/abs/1611.05431 / Erişim Tarihi: 26.09.2025.

58. Ayadi, A., Truong, T. M. D., Frydman, C., Wemmert, C. (2024). *Explaining LSTM neural networks: A DEVS-based approach*. European Modeling and Simulation Symposium*, EMSS*, Fes, 1–10.

59. Lee, D., Lim, M., Park, H., Kang, Y., Park, J. S., Jang, G. J., Kim, J. H. (2017). Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus. *China Communications*, 14(9), 23–31.

60. Hochreiter, S., Schmidhuber, J. (1997). LSTM paper. *Neural Computation*, 9(8), 1735–1780.

61. Elsayed, N., Elsayed, Z., Maida, A. S. (2022). *LiteLSTM Architecture for Deep Recurrent Neural Networks*. Proceedings- IEEE International Symposium on Circuits and Systems, Austin, 1304–1308.

62. Hochreiter, S., Schmidhuber, J. (1997). LSTM Fundometals. *Neural Computation*, 9(8), 1735–1780.

63. O'Shea, T. J., Hitefield, S., Corgan, J. (2017). *End-to-end radio traffic sequence recognition with recurrent neural networks*. IEEE Global Conference on Signal and Information Processing, Montreal, 277–281.

64. Al-Dulaimi, O. A. H. H., Kurnaz, S. (2024). A Hybrid CNN-LSTM Approach for Precision Deepfake Image Detection Based on Transfer Learning. *Electronics*, 13(9), 1–22.

65. İnternet: Tariq, S., Lee, S., Woo, S. S. (2020). A Convolutional LSTM based Residual Network for Deepfake Video Detection. *Arxiv*, Web: http://arxiv.org/abs/2009.07480 / Erişim Tarihi: 26.09.2025.

66. Saikia, P., Dholaria, D., Yadav, P., Patel, V., Roy, M. (2022). *A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features*. 2022 International Joint Conference on Neural Networks (IJCNN), Padua, 1-7.

67. Lee, G., Kim, M. (2021). Deepfake detection using the rate of change between frames based on computer vision. *Sensors*, 21(21).7367.

68. Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., Al Ghamdi, M. A. (2024). Deepfake forensics: A survey of digital forensic methods for multimodal deepfake identification on social media. *Frontiers in Psychology*, 15, 11157519.

69. Chen, Y., Kumar, A., Singh, R. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications, 238*, 1242600.

70. İnterne: Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M. (2016). Face2Face: Real-time Face Capture and Reenactment of RGB Videos. *Arxiv*, Web: http://arxiv.org/abs/2007.14808 / Erişim Tarihi: 26.09.2025.

71. İnternet: Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C. C. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. *Arxiv*, Web: http://arxiv.org/abs/1910.08854 / Erişim Tarihi: 26.09.2025.

72. İnternet: Hasan, M., Ruhama, S., Sithi, S. T., Samit, C. M. M., Saha, O. (2025). Unmasking Deep Fakes: Leveraging Deep Learning for Video Authenticity Detection. *Arxiv*, Web: http://arxiv.org/abs/2505.06528 / Erişim Tarihi: 26.09.2025.

73. Abbasi, M., Váz, P., Silva, J., Martins, P. (2025). Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. *Applied Sciences*, 15(3).1225.

74. İnternet: Lin, C., Deng, J., Hu, P., Shen, C., Wang, Q., Li, Q. (2024). Towards Benchmarking and Evaluating Deepfake Detection. *Arxiv*, Web: http://arxiv.org/abs/2203.02115 / Erişim Tarihi: 26.09.2025.

75. İnternet: Fu, X., Yan, Z., Yao, T., Chen, S., Li, X. (2025). Exploring Unbiased Deepfake Detection via Token-Level Shuffling and Mixing. *Arxiv*, Web http://arxiv.org/abs/2501.04376 / Erişim Tarihi: 26.09.2025.

EKLER

EK-1. DFDC doğruluk veri kümeleri

Random bölümlerden (00,10,19,37,49) doğruluk sonuçları.



Şekil Ek-1.1. (a,b,c,d,e) DFDC üzerinde Accuracy sonuçları

EK-2. DFDC kayıp veri kümeleri

Random bölümlerden (09,21,22,26,47) kayıp sonuçları.



Şekil Ek-2.1 (a,b,c,d,e) DFDC üzerinde Kayıp sonuçları

EK-3. DFDC eğitim veri kümeleri

Random bölümlerden (00,10,23,38,47) eğitim sonuçları.

```
Epoch [11/20], Train Loss: 0.1135, Acc: 96.78%
Test Loss: 0.2319, Acc: 92.88%
  Predictions: REAL=264, FAKE=3
  REAL accuracy: 98.80%
  FAKE accuracy: 0.00%
Epoch [12/20], Train Loss: 0.0750, Acc: 97.71%
Test Loss: 0.3041, Acc: 93.26%
  Predictions: REAL=265, FAKE=2
  REAL accuracy: 99.20%
  FAKE accuracy: 0.00%
Epoch [13/20], Train Loss: 0.0692, Acc: 98.08%
Test Loss: 0.2739, Acc: 93.63%
  Predictions: REAL=264, FAKE=3
  REAL accuracy: 99.20%
  FAKE accuracy: 6.25%
Epoch [14/20], Train Loss: 0.0552, Acc: 98.62%
Test Loss: 0.5362, Acc: 85.02%
  Predictions: REAL=237, FAKE=30
  REAL accuracy: 89.24%
  FAKE accuracy: 18.75%
Epoch [15/20], Train Loss: 0.0438, Acc: 98.44%
Test Loss: 0.3149, Acc: 93.63%
  Predictions: REAL=260, FAKE=7
  REAL accuracy: 98.41%
  FAKE accuracy: 18.75%
Epoch [16/20], Train Loss: 0.0784, Acc: 97.43%
Test Loss: 0.3909, Acc: 94.01%
  Predictions: REAL=265, FAKE=2
  REAL accuracy: 99.60%
  FAKE accuracy: 6.25%
Epoch [17/20], Train Loss: 0.0469, Acc: 98.71%
Test Loss: 0.3813, Acc: 93.63%
  Predictions: REAL=266, FAKE=1
  REAL accuracy: 99.60%
  FAKE accuracy: 0.00%
Epoch [18/20], Train Loss: 0.0463, Acc: 98.90%
Test Loss: 0.4173, Acc: 89.51%
  Predictions: REAL=247, FAKE=20
  REAL accuracy: 93.63%
  FAKE accuracy: 25.00%
Epoch [19/20], Train Loss: 0.0550, Acc: 98.45%
Test Loss: 0.4214, Acc: 93.63%
  Predictions: REAL=260, FAKE=7
  REAL accuracy: 98.41%
  FAKE accuracy: 18.75%
Epoch [20/20], Train Loss: 0.0256, Acc: 99.08%
Test Loss: 0.4580, Acc: 93.63%
  Predictions: REAL=266, FAKE=1
  REAL accuracy: 99.60%
  FAKE accuracy: 0.00%
```

```
  Predictions: REAL=611, FAKE=28
  REAL accuracy: 98.74%
  FAKE accuracy: 24.71%
Epoch [12/20], Train Loss: 0.2210, Acc: 91.23%
Test Loss: 0.4023, Acc: 88.42%
  Predictions: REAL=622, FAKE=17
  REAL accuracy: 99.46%
  FAKE accuracy: 16.47%
Epoch [13/20], Train Loss: 0.1793, Acc: 92.54%
Test Loss: 0.3712, Acc: 87.48%
  Predictions: REAL=622, FAKE=17
  REAL accuracy: 98.92%
  FAKE accuracy: 12.94%
Epoch [14/20], Train Loss: 0.1567, Acc: 93.09%
Test Loss: 0.3709, Acc: 86.23%
  Predictions: REAL=600, FAKE=39
  REAL accuracy: 96.21%
  FAKE accuracy: 21.18%
Epoch [15/20], Train Loss: 0.1651, Acc: 93.15%
Test Loss: 0.3919, Acc: 88.73%
  Predictions: REAL=608, FAKE=31
  REAL accuracy: 98.38%
  FAKE accuracy: 25.88%
Epoch [16/20], Train Loss: 0.1226, Acc: 94.68%
Test Loss: 0.4366, Acc: 81.38%
  Predictions: REAL=515, FAKE=124
  REAL accuracy: 85.74%
  FAKE accuracy: 52.94%
Epoch [17/20], Train Loss: 0.1148, Acc: 94.86%
Test Loss: 0.5611, Acc: 87.95%
  Predictions: REAL=615, FAKE=24
  REAL accuracy: 98.56%
  FAKE accuracy: 18.82%
Epoch [18/20], Train Loss: 0.1052, Acc: 95.41%
Test Loss: 0.5213, Acc: 85.13%
  Predictions: REAL=583, FAKE=56
  REAL accuracy: 94.04%
  FAKE accuracy: 27.06%
Epoch [19/20], Train Loss: 0.1043, Acc: 95.45%
Test Loss: 0.5885, Acc: 86.85%
  Predictions: REAL=604, FAKE=35
  REAL accuracy: 96.93%
  FAKE accuracy: 21.18%
Epoch [20/20], Train Loss: 0.0828, Acc: 96.56%
Test Loss: 0.5964, Acc: 87.48%
  Predictions: REAL=610, FAKE=29
  REAL accuracy: 97.83%
  FAKE accuracy: 20.00%
```

```
Epoch [11/20], Train Loss: 0.1691, Acc: 93.70%
Test Loss: 0.4355, Acc: 84.65%
  Predictions: REAL=440, FAKE=42
  REAL accuracy: 95.75%
  FAKE accuracy: 30.49%
Epoch [12/20], Train Loss: 0.1636, Acc: 93.29%
Test Loss: 0.6513, Acc: 84.85%
  Predictions: REAL=469, FAKE=13
  REAL accuracy: 99.50%
  FAKE accuracy: 13.41%
Epoch [13/20], Train Loss: 0.1294, Acc: 95.13%
Test Loss: 0.6015, Acc: 85.89%
  Predictions: REAL=450, FAKE=32
  REAL accuracy: 97.75%
  FAKE accuracy: 28.05%
Epoch [14/20], Train Loss: 0.1232, Acc: 95.18%
Test Loss: 0.5194, Acc: 85.68%
  Predictions: REAL=435, FAKE=47
  REAL accuracy: 95.75%
  FAKE accuracy: 36.59%
Epoch [15/20], Train Loss: 0.0891, Acc: 96.77%
Test Loss: 0.6996, Acc: 85.48%
  Predictions: REAL=454, FAKE=28
  REAL accuracy: 98.00%
  FAKE accuracy: 24.39%
Epoch [16/20], Train Loss: 0.0783, Acc: 97.08%
Test Loss: 0.6352, Acc: 76.14%
  Predictions: REAL=365, FAKE=117
  REAL accuracy: 81.25%
  FAKE accuracy: 51.22%
Epoch [17/20], Train Loss: 0.0757, Acc: 97.54%
Test Loss: 0.8183, Acc: 83.61%
  Predictions: REAL=445, FAKE=37
  REAL accuracy: 95.75%
  FAKE accuracy: 24.39%
Epoch [18/20], Train Loss: 0.0967, Acc: 96.72%
Test Loss: 0.5062, Acc: 81.33%
  Predictions: REAL=394, FAKE=88
  REAL accuracy: 88.00%
  FAKE accuracy: 48.78%
Epoch [19/20], Train Loss: 0.0701, Acc: 97.49%
Test Loss: 0.6083, Acc: 85.48%
  Predictions: REAL=436, FAKE=46
  REAL accuracy: 95.75%
  FAKE accuracy: 35.37%
Epoch [20/20], Train Loss: 0.0597, Acc: 97.75%
Test Loss: 0.8221, Acc: 80.50%
  Predictions: REAL=398, FAKE=84
  REAL accuracy: 88.00%
```

```
Epoch [11/20], Train Loss: 0.1039, Acc: 95.56%
Test Loss: 0.3107, Acc: 87.30%
  Predictions: REAL=386, FAKE=110
  REAL accuracy: 88.54%
  FAKE accuracy: 80.52%
Epoch [12/20], Train Loss: 0.0827, Acc: 96.82%
Test Loss: 0.2599, Acc: 91.53%
  Predictions: REAL=429, FAKE=67
  REAL accuracy: 96.18%
  FAKE accuracy: 66.23%
Epoch [13/20], Train Loss: 0.0561, Acc: 97.78%
Test Loss: 0.3260, Acc: 90.93%
  Predictions: REAL=432, FAKE=64
  REAL accuracy: 96.18%
  FAKE accuracy: 62.34%
Epoch [14/20], Train Loss: 0.0555, Acc: 97.63%
Test Loss: 0.2908, Acc: 88.71%
  Predictions: REAL=409, FAKE=87
  REAL accuracy: 92.12%
  FAKE accuracy: 70.13%
Epoch [15/20], Train Loss: 0.0828, Acc: 97.10%
Test Loss: 0.2569, Acc: 90.32%
  Predictions: REAL=435, FAKE=61
  REAL accuracy: 96.18%
  FAKE accuracy: 58.44%
Epoch [16/20], Train Loss: 0.0785, Acc: 97.48%
Test Loss: 0.4187, Acc: 88.71%
  Predictions: REAL=445, FAKE=51
  REAL accuracy: 96.42%
  FAKE accuracy: 46.75%
Epoch [17/20], Train Loss: 0.1095, Acc: 96.32%
Test Loss: 0.2927, Acc: 89.92%
  Predictions: REAL=421, FAKE=75
  REAL accuracy: 94.27%
  FAKE accuracy: 66.23%
Epoch [18/20], Train Loss: 0.0580, Acc: 97.93%
Test Loss: 0.5275, Acc: 85.28%
  Predictions: REAL=382, FAKE=114
  REAL accuracy: 86.87%
  FAKE accuracy: 76.62%
Epoch [19/20], Train Loss: 0.0569, Acc: 98.03%
Test Loss: 0.2991, Acc: 91.33%
  Predictions: REAL=432, FAKE=64
  REAL accuracy: 96.42%
  FAKE accuracy: 63.64%
Epoch [20/20], Train Loss: 0.0487, Acc: 98.13%
Test Loss: 0.3576, Acc: 90.93%
  Predictions: REAL=418, FAKE=78
  REAL accuracy: 94.51%
```

```
Epoch [11/20], Train Loss: 0.1632, Acc: 93.71%
Test Loss: 0.5496, Acc: 86.00%
  Predictions: REAL=451, FAKE=42
  REAL accuracy: 97.04%
  FAKE accuracy: 34.48%
Epoch [12/20], Train Loss: 0.1218, Acc: 94.97%
Test Loss: 0.5763, Acc: 85.40%
  Predictions: REAL=438, FAKE=55
  REAL accuracy: 95.07%
  FAKE accuracy: 40.23%
Epoch [13/20], Train Loss: 0.1421, Acc: 94.68%
Test Loss: 0.4586, Acc: 85.19%
  Predictions: REAL=431, FAKE=62
  REAL accuracy: 94.09%
  FAKE accuracy: 43.68%
Epoch [14/20], Train Loss: 0.1019, Acc: 95.93%
Test Loss: 0.5891, Acc: 84.79%
  Predictions: REAL=415, FAKE=78
  REAL accuracy: 91.87%
  FAKE accuracy: 51.72%
Epoch [15/20], Train Loss: 0.0829, Acc: 96.23%
Test Loss: 0.7176, Acc: 82.76%
  Predictions: REAL=391, FAKE=102
  REAL accuracy: 87.68%
  FAKE accuracy: 59.77%
Epoch [16/20], Train Loss: 0.1051, Acc: 96.19%
Test Loss: 0.9500, Acc: 79.31%
  Predictions: REAL=378, FAKE=115
  REAL accuracy: 83.99%
  FAKE accuracy: 57.47%
Epoch [17/20], Train Loss: 0.1235, Acc: 95.61%
Test Loss: 0.5881, Acc: 81.14%
  Predictions: REAL=383, FAKE=110
  REAL accuracy: 85.71%
  FAKE accuracy: 59.77%
Epoch [18/20], Train Loss: 0.1031, Acc: 95.94%
Test Loss: 0.5634, Acc: 86.00%
  Predictions: REAL=435, FAKE=58
  REAL accuracy: 95.07%
  FAKE accuracy: 43.68%
Epoch [19/20], Train Loss: 0.0611, Acc: 97.89%
Test Loss: 0.6501, Acc: 84.99%
  Predictions: REAL=428, FAKE=65
  REAL accuracy: 93.60%
  FAKE accuracy: 44.83%
Epoch [20/20], Train Loss: 0.0695, Acc: 97.68%
Test Loss: 0.6661, Acc: 86.00%
  Predictions: REAL=421, FAKE=72
  REAL accuracy: 93.35%
```

Şekil Ek-3.1 (a,b,c,d,e) DFDC üzerinde run sonuçları



Gazili olmak ayrıcalıktır