



**ÇEKİŞMELİ ÜRETİCİ AĞ KALINTILARININ ALGILANMASI İLE  
DEEPPAKE TESPİTİ**

**Tamer SAY**

**DOKTORA TEZİ  
BİLGİ GÜVENLİĞİ MÜHENDİSLİĞİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ  
FEN BİLİMLERİ ENSTİTÜSÜ**

**MAYIS 2025**

## ETİK BEYAN

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

- 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 çalışmada 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.

Tamer SAY

05/05/2025

# ÇEKİŞMELİ ÜRETİCİ AĞ KALINTILARININ ALGILANMASI İLE DEEPFAKE TESPİTİ

(Doktora Tezi)

Tamer SAY

GAZİ ÜNİVERSİTESİ  
FEN BİLİMLERİ ENSTİTÜSÜ

Mayıs 2025

## ÖZET

Bu tez, deepfake/yüz manipülasyonunun tarihsel gelişimini inceleyerek makine öğrenmesi tekniklerinin bu alandaki uygulamalarına odaklanmaktadır. Bu bağlamda, yıllar içerisinde başarısını kanıtlamış ve öne çıkan GAN (Generative Adversarial Networks) tabanlı deepfake çalışmaları ele alınmıştır. GAN ile üretilen deepfake veri setleri kapsamlı bir şekilde analiz edilmiş; literatürde yer alan çalışmalarda GAN tabanlı veri setlerinin tespit yöntemleri ve bu yöntemlerin başarımları araştırılmıştır. Deepfake tespiti açısından GAN kalıntılarının incelenmesi, literatürde yeterince ele alınmamış bir konu olup bu alanda sınırlı sayıda araştırma bulunmaktadır. Mevcut araştırmalara alternatif olarak, bu tezde GAN kalıntıları üzerinde deepfake tespiti amacıyla özellik çıkarımı yöntemi kullanılmış ve çeşitli sınıflandırma algoritmaları ile tespit simülasyonları geliştirilmiştir. Fourier, Wavelet ve HOG (Histogram of Oriented Gradients) dönüşüm metodundan yararlanılan çalışmada, GAN kalıntılarını tespit etmeye yönelik çoklu GAN modeli içeren bir veri seti oluşturulmuştur. Bu özgün veri seti ile gerçek veri seti üzerinde uygulanan özellik çıkarım yönteminin başarımları test edilmiştir. Elde edilen sonuçlar doğrultusunda, farklı GAN modellerinin performansları karşılaştırılmış ve geliştirilen simülasyonun tespit başarımları incelenmiştir. GAN kalıntısı tespit ve analiz simülasyonu kapsamında, belirli GAN modellerinin ürettiği görseller üzerine RGB renk uzayında dönüşüm metodları uygulanmış ve elde edilen veriler frekans uzayında incelenerek sınıflandırma algoritmalarına sunulularak sonuç elde edilmiştir. Deepfake tespiti amacıyla geliştirilen bu yöntem ile GAN modeli ile üretilen görsellerde %100'e varan tespit başarımlarına ulaşılmıştır. Ayrıca, farklı sınıflandırma algoritmalarının entegrasyonu ile gerçekleştirilen bu çalışmada, uygun sınıflandırma algoritmasının seçimine yönelik bulgular da paylaşılmıştır.

Bilim Kodu : 92432

Anahtar Kelimeler : Deepfake tespiti, derin öğrenme, makine öğrenmesi, yapay sinir ağları, GAN tespiti

Sayfa Adedi : 81

Danışman : Prof. Dr. Mustafa ALKAN

DEEPFAKE DETECTION BY DETECTING GENERATIVE ADVERSARIAL  
NETWORK ARTIFACTS

(Ph. D. Thesis)

Tamer SAY

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

May 2025

ABSTRACT

This thesis examines the historical development of deepfake/face manipulation technologies, with a focus on machine learning techniques and their applications in this field. It specifically investigates GAN (Generative Adversarial Networks)-generated deepfake studies that have demonstrated success over the years. GAN-generated deepfake datasets are comprehensively analyzed, along with detection methods and their performance metrics documented in existing literature. The examination of GAN artifacts for deepfake detection represents an understudied area, with limited research currently available. As an alternative to existing approaches, this study employs feature extraction methods using Fourier, Wavelet, and HOG (Histogram of Oriented Gradients) transformations to detect GAN artifacts, developing detection simulations through various classification algorithms. A novel dataset containing multiple GAN models is created to test the effectiveness of the proposed feature extraction method on both synthetic and real-world data. Experimental results demonstrate 100% detection accuracy for images generated by specific GAN models when analyzing their artifacts in the frequency domain after applying RGB color space transformations. The study compares performance across different GAN architectures and evaluates the detection efficacy of the developed simulation framework. Additionally, it provides insights into optimal classifier selection through the integration of multiple classification algorithms. This methodology advances deepfake detection by focusing on GAN-generated visual artifacts, offering a robust approach that combines frequency domain analysis with machine learning classification. The findings contribute to understanding model-specific patterns in synthetic media and establish benchmarks for future research in artifact-based detection systems.

Science Code : 92432

Key Words : Deepfake detection, deep learning, machine learning, artificial neural networks, GAN detection

Page Number : 81

Supervisor : Prof. Dr. Mustafa ALKAN

## TEŞEKKÜR

Çalışmalarım boyunca desteğini esirgemeyen, tecrübeleri ile her zaman doğru yolu gösteren danışmanım saygıdeğer hocam Prof. Dr. Mustafa ALKAN'a, tezin hazırlanması aşamasında ve diğer çalışmalarda birlikte çalıştığımız değerli Aynur KOÇAK hocama, doktora sürecinde çalışmalarına katkı sağlayan, bu alanda sağladığım ilerlemelere ve gelişmelere katkı sağlayan değerli hocalarım Prof Dr. Türksel KAYA BENSGHİR ve Prof Dr. Ercan Nurcan YILMAZ'a teşekkürlerimi sunarım. Maddi ve manevi desteğini hayatım boyunca hissettiğim aileme ve bu tezin hazırlaması sürecinde çalışma fırsatı bulduğum en başta Traffic Management ekibi olmak üzere tüm Trendyol ailesine teşekkürlerimi sunarım.



## İÇİNDEKİLER

|                                                                   | Sayfa |
|-------------------------------------------------------------------|-------|
| ÖZET.....                                                         | iv    |
| ABSTRACT.....                                                     | v     |
| TEŞEKKÜR.....                                                     | vi    |
| İÇİNDEKİLER.....                                                  | vii   |
| ÇİZELGELERİN LİSTESİ.....                                         | ix    |
| ŞEKİLLERİN LİSTESİ.....                                           | xii   |
| RESİMLERİN LİSTESİ.....                                           | vii   |
| SİMGELER VE KISALTMALAR.....                                      | xv    |
| 1. GİRİŞ.....                                                     | 1     |
| 2. LİTERATÜR TARAMASI.....                                        | 5     |
| 2.1. Deepfake Teriminin Tanımı ve Literatürdeki Yeri.....         | 6     |
| 2.2. Deepfake Oluşturma Yöntemleri.....                           | 6     |
| 2.3. Deepfake Tespit Modelleri ve Yöntemleri.....                 | 8     |
| 2.4. GAN ile Deepfake Oluşturulması.....                          | 11    |
| 2.4.1. GAN modelleri ve yaklaşımları.....                         | 13    |
| 2.4.2. GAN yüz tespit yöntemleri.....                             | 15    |
| 2.4.3. GAN-üreteçli sentetik kalıntılar.....                      | 17    |
| 2.5. Deepfake Veri Setleri ve GAN Üreteçli İmaj Veri Setleri..... | 29    |
| 2.6. GAN Kalıntıları ve Gürültüsü Önleyici Yöntemler.....         | 31    |
| 3. GAN KALINTILARININ İNCELENMESİ.....                            | 33    |
| 3.1. Yapay Sinir Ağları ve Derin Öğrenme.....                     | 33    |
| 3.2. GAN Mimarisi.....                                            | 34    |
| 3.2.1. GAN kalıntıları.....                                       | 39    |

**Sayfa**

|                                                                               |    |
|-------------------------------------------------------------------------------|----|
| 4. METOT VE ANALİZ.....                                                       | 41 |
| 4.1. GAN Kalıntısı Taksonomisinin Çıkarılması.....                            | 41 |
| 4.1.1. Görünmez GAN kalıntıları.....                                          | 43 |
| 4.2. Tanınmamış ve Ön Çalışmasız GAN Tespiti.....                             | 45 |
| 4.3. GAN Kalıntısı Tespit Simülasyonu Geliştirilmesi ve Özellikleri.....      | 46 |
| 4.4. Çoklu GAN Veri Seti.....                                                 | 49 |
| 4.5. Çoklu GAN Simülasyonu ile Kalıntı Algılama Üzerinden Deepfake Tespiti... | 50 |
| 4.6. GAN Simülasyonu Sonuç Analizi.....                                       | 62 |
| 5. SONUÇLAR VE ÖNERİLER.....                                                  | 67 |
| KAYNAKLAR.....                                                                | 69 |
| ÖZGEÇMİŞ.....                                                                 | 81 |

## ÇİZELGELERİN LİSTESİ

| Çizelge                                                                                                                    | Sayfa |
|----------------------------------------------------------------------------------------------------------------------------|-------|
| Çizelge 3.1. GAN mimarisi eşitlikleri.....                                                                                 | 36    |
| Çizelge 4.1. Baz veri seti ve GAN modeli tarafından üretilen veri setleri bilgileri.....                                   | 49    |
| Çizelge 4.2. GAN modeli ve veri seti oluşturma süreleri bilgisi.....                                                       | 50    |
| Çizelge 4.3. InterfaceGAN görselleri için sınıflandırma algoritmaları başarımlarını doğruluk oranları tespit süreleri..... | 58    |
| Çizelge 4.4. ProGAN görselleri için sınıflandırma algoritmaları başarımlarını doğruluk oranları tespit süreleri.....       | 59    |
| Çizelge 4.5. StyleGAN3 görselleri için sınıflandırma algoritmaları başarımlarını doğruluk oranları tespit süreleri.....    | 59    |
| Çizelge 4.6. InterfaceGAN görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri.....       | 60    |
| Çizelge 4.7. ProGAN görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri.....             | 61    |
| Çizelge 4.8. StyleGAN3 görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri.....          | 61    |

## ŞEKİLLERİN LİSTESİ

| Şekil                                                                                                              | Sayfa |
|--------------------------------------------------------------------------------------------------------------------|-------|
| Şekil 2.1. Deepfake oluşturma üzerine yapılan toplam çalışma sayısı grafiği.....                                   | 7     |
| Şekil 2.2. “Deepfake oluşturma” ve “deepfake tespiti” geçen çalışmaların sayısının yıllara göre artış grafiği..... | 9     |
| Şekil 2.3. Özellik tabanlı deepfake tespiti için önerilen taksonomi.....                                           | 10    |
| Şekil 2.4. İmajların ortalama kalıntı ve parmak izi korelasyonu.....                                               | 24    |
| Şekil 2.5. Renk bileşeni ve eş-oluş matrisi üzerinden ikili sınıflandırma yöntemi akışı.                           | 31    |
| Şekil 2.6. Gerçek ve sahte görsellerin ortalama DCT spektral değerleri ve spektral kalıntılar farkı.....           | 26    |
| Şekil 2.7. Spektrum analizini atlatma yöntemi çalışma yöntemi.....                                                 | 31    |
| Şekil 3.1. Örnek yapay sinir ağı modeli.....                                                                       | 33    |
| Şekil 3.2. GAN mimarisi.....                                                                                       | 36    |
| Şekil 3.3. CGAN modeline ait olan üretici-ayrıştırıcı toplam kayıp değeri.....                                     | 37    |
| Şekil 3.4. Çekişmeli üretici ağ çalışmalarının yıllar içindeki sayısı.....                                         | 38    |
| Şekil 4.1. GAN kalıntıları taksonomisi.....                                                                        | 42    |
| Şekil 4.2. CelebA ve InterfaceGAN görsellerinin HOG dönüşümü altındaki 3B veri dağılımı.....                       | 53    |
| Şekil 4.3. CelebA ve ProGAN görsellerinin HOG dönüşümü altındaki 3B veri dağılımı.....                             | 54    |
| Şekil 4.4. CelebA ve StyleGAN3 görsellerinin HOG dönüşümü altındaki 3B veri dağılımı.....                          | 54    |
| Şekil 4.5. CelebA ve InterfaceGAN görsellerinin Wavelet dönüşümü altındaki 3B veri dağılımı.....                   | 55    |
| Şekil 4.6. CelebA ve ProGAN görsellerinin Wavelet dönüşümü altındaki 3B veri dağılımı.....                         | 55    |
| Şekil 4.7. CelebA ve StyleGAN3 görsellerinin Wavelet dönüşümü altındaki 3B veri dağılımı.....                      | 56    |

| Şekil                                                                                            | Sayfa |
|--------------------------------------------------------------------------------------------------|-------|
| Şekil 4.8. CelebA ve InterfaceGAN görsellerinin Fourier dönüşümü altındaki 3B veri dağılımı..... | 56    |
| Şekil 4.9. CelebA ve StyleGAN3 görsellerinin Fourier dönüşümü altındaki 3B veri dağılımı.....    | 57    |
| Şekil 4.10. CelebA ve StyleGAN3 görsellerinin Fourier dönüşümü altındaki 3B veri dağılımı.....   | 57    |



## RESİMLERİN LİSTESİ

| Resim                                                                                                                                                         | Sayfa |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| Resim 2.1. StyleGAN ve versiyonları ile üretilmiş deepfake görselleri.....                                                                                    | 12    |
| Resim 2.2. Farklı GAN modelleri tarafından üretilmiş deepfake imajlar.....                                                                                    | 13    |
| Resim 2.3. Gözlemlenebilir kalıntılara örnek üç GAN-üreteçli deepfake.....                                                                                    | 18    |
| Resim 2.4. Girdi, yeniden oluşturma, GAN kalıntıları ve çıktılar üzerinden kalıntıların karşılaştırılması.....                                                | 19    |
| Resim 2.5. Orijinal görseller (birinci satır), GPNN (ikinci satır) ve SinGAN (üçüncü satır) tarafından üretilen deepfake örnekleri.....                       | 21    |
| Resim 2.6. Görülebilir kalıntılar (birinci satır) ve ikinci satırda sırasıyla ortalama Fourier spektrumu, frekans analizi ve frekans spektrumu çıktıları..... | 22    |
| Resim 2.7. Çeşitli veri setlerine dair frekans analizi sonuçları.....                                                                                         | 27    |
| Resim 2.8. Soldan sağa, gerçek görsel, sentetik görsel, gri ölçekli temel gerçek haritası, ikili (binary) gerçeklik haritası.....                             | 28    |
| Resim 4.1. PGGAN, StarGAN, StyleGAN ve SC-FEGAN tarafından üretilmiş deepfake imajlar.....                                                                    | 43    |
| Resim 4.2. CelebA gerçek yüzler veri seti örnek imajlar.....                                                                                                  | 51    |
| Resim 4.3. ProGAN tarafından üretilmiş deepfake imajlar.....                                                                                                  | 51    |
| Resim 4.4. InterfaceGAN tarafından üretilmiş deepfake imajlar.....                                                                                            | 51    |
| Resim 4.5. StyleGAN3 tarafından üretilmiş deepfake imajlar.....                                                                                               | 51    |
| Resim 4.6. Tüm veri setlerine HOG dönüşümü uygulanması sonucu görünümü.....                                                                                   | 52    |

## SİMGELER VE KISALTMALAR

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

### Simgeler

### Açıklamalar

$\mu$

GAN’da olasılık dağılımını ifade eder

$\gamma$

GAN’da başlangıç dağılımını ifade eder

### Kısaltmalar

### Açıklamalar

**3B**

Üç Boyutlu

**AI**

Artificial Ingelligence

**AUC**

Area Under The Curve

**CGI**

Computer-Generated Imagery

**CNN**

Convolutional Neural Network

**CPU**

Central Processing Unit

**DCT**

Discrete Cosine Transform

**DL**

Deep Learning

**GAN**

Generative Adversarial Networks

**GMM**

Gaussian Mixture Model

**GB**

GigaByte

**GPU**

Graphics Processing Unit

**HOG**

Histogram of Oriented Gradients

**HSV**

Hue Saturation Value

**JPEG**

Joint Photographic Experts Group

**kNN**

K-Nearest Neighbors

**LSTM**

Long Short-Term Memory

**ML**

Machine Learning

**RAM**

Random Access Memory

**RBF**

Radial Basis Function

**RCNN**

Region-based Convolutional Neural Network

**Kısaltmalar****Açıklamalar****RNN**

Recurrent Neural Networks

**RG-RB-GB**

Red Green – Red Blue – Green Blue

**RGB**

Red Green Blue

**SVM**

Support Vector Machines

**TPU**

Tensor Processing Unit

**YCbCr**

Y Luma Bileşeni, CbCr Krominans Bileşeni

**YSA**

Yapay Sinir Ağları



## 1. GİRİŞ

Yapay zekâ, makine öğrenmesi ve derin öğrenme alanlarındaki hızlı ilerlemeler, özellikle görsel medya ve iletişim teknolojilerinde devrim niteliğinde yenilikler getirmiştir. Bu gelişmelerin en çarpıcı örneklerinden biri deepfake ve yüz manipülasyonu teknolojisi. Bu teknolojiler, gerçek kişilerin görüntülerini ve videolarını, onların gerçekte yapmadıkları eylemleri yapar veya söyler gibi göstererek sahte hallerinin sentetik olarak gerçekleştirilme işlemidir. Tez, bu yöntemlerden görsel sahteciliğinde odaklanmakta olup, ses alanındaki sahtecilikleri içermemektedir; deepfake teknolojisinin makine öğrenmesi yöntemleri ile nasıl gerçekleştirildiğini, bu teknolojilerden olan GAN (Generative Adversarial Networks) modellerinin bıraktıkları kalıntılar üzerinden tespitini kapsamlı bir şekilde ele alarak çözüm yollarını derinlemesine incelemektedir. İlgili teknoloji bu alandaki en tehlikeli ve tespit edilmesi güç yöntemlerden birisi olarak dikkat çekmektedir. Bu nedenle bu yönetimin tespitinde ilgili yöntemin kullanıldığı takdirde kalıntılarından sahtecilik tespiti yapılabilmesi araştırılmış; bir simülasyon ile çözüm önerisi sunulmuştur. GAN kalıntısı analizi ile de bu alanda yapılan çalışmalara bir ışık tutularak literatüre kalıntılar üzerinden tespit üzerine bir taksonomi çalışması kazandırılmıştır.

### Problem durumu / Konunun tanımı

Deepfake ve yüz manipülasyonu teknolojilerinin yükselişi, dijital dünyada gerçeğin ve sahtenin ayrımını zorlaştırmıştır. Bu teknolojiler, bireysel hakların ihlali, yanıltıcı bilgi yayılımı ve dijital sahtecilik gibi ciddi sorunlara yol açmaktadır. Bu tez, deepfake oluşturma yöntemlerinden olan GAN modelinin kullanıldığı takdirde insan gözü ile gözlemlenemeyen fakat dijital analizler sonucunda anlaşılan iz/kalıntı bırakan GAN yöntemlerinin tespit edilebilirliğini araştırmakta ve bu alandaki sorunları tanımlamaktadır. GAN modeli oldukça karmaşık ve çalışma prensibi nedeniyle tespit edilmesi güç bir yapıya sahiptir. Öğrenim aşamasında ve çeşitli katmanlarındaki tekrarlar/kullanımlar GAN modellerinin tespit edilmesi güç kalıntılara sebebiyet vermektedir. Literatürde bu kalıntıları sınıflandıran veya doğrudan tanımlayan bir çalışma olmaması ve gelişmiş GAN modellerinde kalıntı - gürültü önleyici yöntemler kullanılması tespit edilebilirliği ciddi seviyede zorlaştırmaktadır. Çeşitli GAN modelleri üzerinden bu kalıntıların ve tekrar nedeniyle oluşan sıklık/genelleme analizi ile GAN tanımlamasının araştırılması önem arz etmektedir. Ek olarak frekans uzayı

üzerinden gerçekleştirilen analizlerin incelenerek diğer kalıntı tespit yöntemleri ile birlikte uyarlanarak tespit doğruluğu ve başarımının katkısının gözlemlenerek incelenmesi bu alandaki çalışmalarındaki konunun anlaşılmasına yardımcı olacaktır.

### Araştırmanın amacı

Bu araştırmanın ana amacı, deepfake ve yüz manipülasyonu teknolojilerinin arkasındaki GAN yöntemlerini ve araçlarını bıraktıkları kalıntıları algılayarak tespit etmeyi amaçlamaktadır. Tez, bu teknolojileri tespit etmek için kullanılan kalıntı tespiti, genelleme yöntemi ve ortak katman kullanımından kaynaklanan izlerden yola çıkarak tespit etmeye odaklanmaktadır. Frekans spektrumuna dönüşüm yöntemleri kullanılarak geliştirilen simülasyon ile deepfake görsellerin tespit edilerek ayrıştırılması hedeflenmiştir. Ayrıca, bu teknolojiler için geliştirilmiş anti tezleri ve yöntemleri incelemektedir. Sunulan simülasyon yöntemi ile de daha önce karşılaşılmamış GAN yöntemleri için bir tespit yöntemi için temel oluşturmayı hedeflemiştir. Buna ek olarak GAN modelinin algılanması için kalıntılar üzerinden bir sınıflandırma çıkartılarak literatüre öncü olması amaçlanmıştır.

### Araştırmanın önemi

Deepfake ve yüz manipülasyonu, dijital çağda karşılaştığımız en önemli zorluklardan biridir. Bu teknolojilerin kullanımı, toplumda güven sorunlarına yol açmakta ve gerçek ile sahte arasındaki sınırları bulanıklaştırmaktadır. Deepfake, dezenformasyon, intikam içerikleri oluşturma, itibar zedeleme, panik ve karmaşıklığa yol açarak karışıklıkların/savaşın seyrini değiştirme gibi sorunlara yol açabilecek kadar tehlikeli bir dijital silah olarak kullanılabilir. Bu tezin önemi, deepfake oluşturulması için kullanılan en önemli yöntemlerden olan GAN'ın tespit edilebilmesi ile bu teknolojilerin olumsuz etkilerini azaltmak için alınabilecek önlemleri kapsamlı bir şekilde tartışmasından kaynaklanmaktadır. Türkçe literatürde GAN kalıntısı üzerine yapılmış kapsamlı ve ayrıntılı bir çalışma olması ile bu alandaki bilgi boşluğunu doldurmayı hedeflemekte ve akademik çalışmalara önemli bir katkı sunmaktadır. GAN kalıntıları ile GAN modeli tanımlanması üzerine bir sınıflandırma çalışması olmaması nedeniyle literatüre kazandırılması bu alandaki çalışmalara yol gösterici olacaktır. GAN tespiti için kullanılmak üzere geliştirilen veri seti sayesinde bu alanda çalışan araştırmacılara yol gösterecek ve halihazırda çeşitli GAN modelleri tarafından üretilen karma veri setleri arayanlar için kolaylık sağlayacaktır. Bu

alanda karma bir veri seti olmaması ve GAN modellerinin geliştirilmesinin güçlüğü nedeniyle bu veri setinin çalışmalarda kullanılması öngörülmüştür.

Çalışmamızın ikinci bölümünde literatürde yer alan deepfake tespiti üzerine yapılan çalışmalar ve bu çalışmaların sınıflandırıldığı başlıklara yer verilmiş, GAN modelleri ile üretilen deepfake'ler hakkında kullanılan yöntemler incelenmiştir. Üçüncü bölümde GAN mimarisi açıklanmış, matematiksel gösterimi ve tarihçesi konusunda bilgiler aktarılmıştır, çalışmamızın konusu olan GAN kalıntıları hakkında bilgiler verilmiştir. Dördüncü bölümde çalışmamızda GAN kalıntıları üzerinden deepfake tespiti gerçekleştirmemizi sağlayan metot aktarılmış, farklı GAN modelleri kullanarak oluşturduğumuz veri seti hakkında bilgiler sunulmuş ve geliştirdiğimiz çoklu GAN tespit modelinin uygulandığı simülasyonun elde ettiği sonuçlar analiz edilmiştir. Beşinci bölümde ise elde ettiğimiz çoklu GAN tespit simülasyonu sonuçları aktarılmış ve diğer çalışmalarla kıyaslamalar yapılarak gelecek çalışmalar konusunda öneriler sunulmuştur.



## 2. LİTERATÜR TARAMASI

Bu bölümde deepfake terimi, tarihi, mevcut deepfake oluşturma ve tespit yöntemleri kısaca incelenmiştir. Önemli deepfake oluşturma yöntemlerinden olan GAN modeli incelenmiş, çeşitli GAN modelleri ve diğer modellere göre öne çıkan özellikleri elde edilmiş, GAN ile üretilen veri setleri hakkında bilgiler derlenmiştir. Tezin konusu olan GAN üreteçli imajlardaki GAN kalıntılarının algılanması ile deepfake tespiti ve GAN modelinin tanımlanması mevcut literatürde araştırılmıştır.

Araştırma yöntemi olarak;

- Yök tez arşivi
- DergiPark
- Google Scholar
- IEEE arşivi
- Scopus
- ScienceDirect
- Cornell Üniversitesi (arxiv) arşivi
- Papers with Code web sitesi
- Dimensions.ai websitesi ve istatistikleri
- İçeriğinde “Thesis” geçen ve tezin anahtar kelimelerini içeren Google araması

kullanılmıştır.

Türkçe literatürde GAN kalıntısı üzerine yapılan sadece bir çalışma yer almaktadır. GAN kalıntıları çeşitli kaynaklarda, “GAN artifacts”, “GAN residue” veya “GAN residual” olarak bahsedilmektedir. Tarama süresince bu kullanım şekilleri göz önünde bulundurulmuş; bu çalışmalar farklı bir konuyu ele aldığı için kapsam dışı olarak nitelendirilmiştir. Çeşitli çalışmalarda anomali olarak da isimlendirilen kalıntılar; bazı çalışmalarda GAN parmak izi (GAN fingerprints) olarak da geçmektedir.

GAN modeli ilk olarak 2014 yılında Goodfellow ve diğerleri (2014) tarafından öne sürülmüştür; öncü çalışmalarının ise 2016 yılında yoğunluk kazandığı ilerleyen bölümlerde ele alınmaktadır. GAN kalıntısı ifadesi ise ilk olarak 2017 yılındaki çalışmalarda görülmüştür. Bu nedenle literatür taraması GAN modelleri için 2016 yılı ve sonrasına,

deepfake çalışmaları için 2017 ve sonrasında, GAN kalıntıları için ise 2017 ve sonrasındaki çalışmalara odaklanmaktadır.

## **2.1. Deepfake Teriminin Tanımı ve Literatürdeki Yeri**

Makine öğrenmesi ve derin öğrenme, yapay zekâ alanında kullanılan otonom hale gelmiş algoritmalara karşılık gelmektedir. Derin öğrenme, makine öğrenmesinin daha gelişmiş bir kolu olarak yapay sinir ağları kullanarak öğrenme sürecini yöneten ve karşılaştığı sorunlara karşı çözüm üreterek ilerleyen bir algoritma/uygulama olarak karşımıza çıkmaktadır.

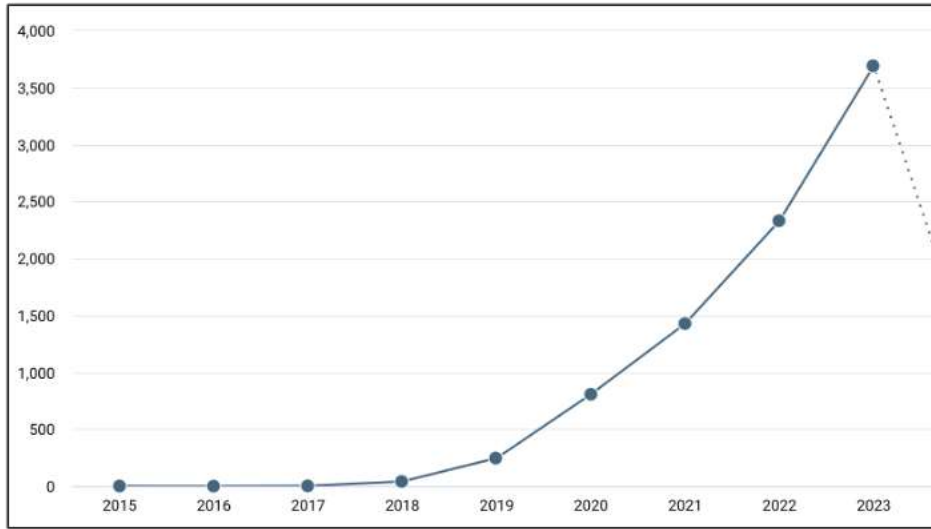
Deepfake, derin öğrenme (deep learning) ve sahtecilik (fake) kelimelerinden türetilmiştir. İlk olarak 2017 yılında Reddit isimli platformda Deep-fakes takma adını kullanan bir hesaptan geldiği (Babayiğit, 2021; Erdoğan, 2023; Chadha, Kumar, Kashyap ve Gupta, 2021) için bu yönde literatürde kabul görmüştür. Deepfake, literatürde yapay öğrenme algoritmaları kullanılarak sahte görsel ve ses anlamında içerik üretilmesine denilmektedir. Deepfake oluşturulması için çeşitli motivasyonlar söz konusudur. Bunlar Sandotra ve Arora (2023) tarafından yapılan çalışmada olumlu ve olumsuz olmak üzere iki sınıfta incelenmiş ve olumlu örneklerinin; eğlence, eğitim, iş ve sanat alanlarında görüldüğü, olumsuz örneklerinin ise; kurumlara yönelik sahtecilik sorunları, kişilik haklarına saldırı, politik çıkarlar ve sahte habercilik/dezenformasyon olarak listelemiştir. Kişilik haklarının ihlali ve politik çıkarlar anlamında deepfake örnekleri özellikle ünlü figürler üzerinde gerçekleştirilen birçok çeşidi yıllar içerisinde görülmüştür. Ses ve görüntüde deepfake yaparak bir kişinin gerçekte söylemediği ifadeleri söyletmek ve yüz ifadesini buna göre değiştirmek ya da bir at fotoğrafından zebra görseli üretmek (Zhu, Park, Isola ve Efros, 2017) gibi örnekleri olarak literatürde ve gerçek dünyada yer bulmaktadır.

## **2.2. Deepfake Oluşturma Yöntemleri**

Bir kişinin fotoğrafındaki yüz hatları, mimikleri, saç, göz vb karakteristik özellikleri değiştirerek farklı bir görünüme dönüştürme ya da farklı bir kişiye benzetme uzun yıllardır Photoshop ve benzeri programlar kullanarak gerçekleştirilen bir süreçtir. Teknolojinin gelişmesi ile son teknoloji cep telefonları ve kameralarda da benzer filtre uygulamaları (Warzel, URL-1, 2019; Ry, 2019; Wirth, 2023) yüz değiştirme veya dönüştürme ve vücut aktarımı gibi görevleri yerine getirecek kaliteye ulaşmıştır. Son yıllarda yapay zeka

uygulamalarının gelişmesi ile bu süreç görece daha zor olan videoda yüz manipülasyonu ve seste sahtecilik yapılması ile daha gerçekçi görünen ve insan algılarıyla tespit edilebilmesi neredeyse imkânsız olan deepfake yapılmasının önünü açmıştır.

Deepfake oluşturma yöntemlerinin yıllar içerisinde literatürde toplam çalışılma sayıları Dimensions.ai websitesinden elde edilen istatistiğe göre Şekil 2.1’de gösterilmektedir. Toplamda 9 752 çalışma listelenmektedir.



Şekil 2.1. Deepfake oluşturma üzerine yapılan toplam çalışma sayısı grafiği

Deepfake oluşturulması sınıflandırma olarak genellikle dört kategoride (Chadha, Kumar, Kashyap ve Gupta, 2021) incelenmektedir. Bunlar; fotoğraf, ses, video ve ses-video birlikte olacak şekilde ayrılmıştır.

Sandotra ve Arora, 2023 tarafından yapılan çalışmada yapay zeka kullanımı ile deepfake tespiti yapılan en güncel ve en kapsamlı çalışmalar arasında yer almaktadır. Deepfake, yüz manipülasyonu yapılarak oluşturulan çalışmaların yoğunlukla odaklandıkları yöntemler bu çalışmada listelenmiştir. Çalışmaya göre kullanılan yüz manipülasyon teknikleri belirtilmiştir. Bunlar;

- Görüntü sentezi
- Kimlik değiştirme
- Dudak senkronizasyonu
- Nitelik/özellik manipülasyonu
- Yeniden canlandırma olarak sınıflandırılmıştır.

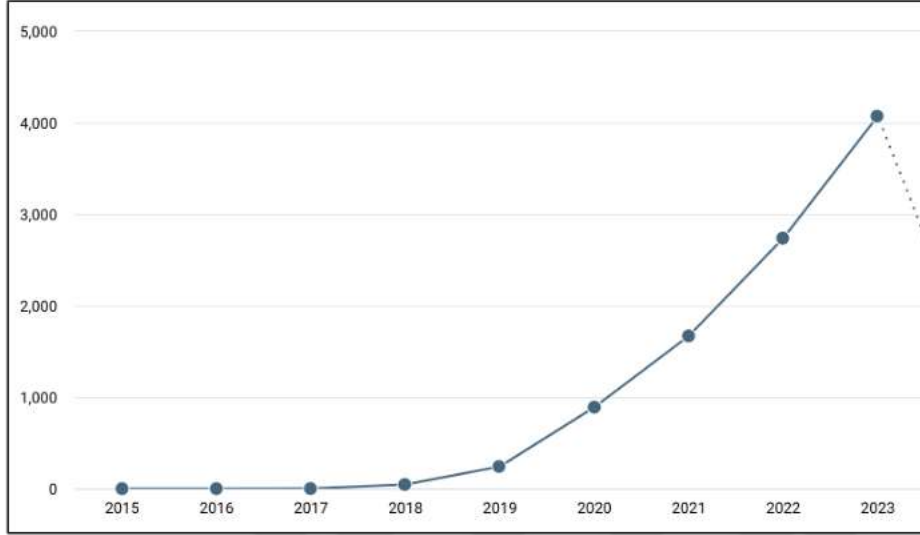
Genellikle kodlayıcı-çözücü (encoder-decoder) ve üretken (generative) ağlar kullanılan (Padmanabhan ve Kishan, 2021) deepfake uygulamalarında kodlayıcı girdi olarak verilen görselden özellik vektörü çıkarımı yaparken çözücü imajın oluşturulmasında görev almaktadır. Ağdaki hata fonksiyonu girdi ve çıktı arasında hesaplama yaparak ağı destekler bu sayede daha iyi sonuçlar alınması yani daha gerçekçi görseller üretilmesi sağlanır.

GAN modeli Goodfellow ve diğerleri (2014) tarafından ortaya atılmıştır. Üretici ve ayrıştırıcı isimli iki farklı derin öğrenme ağı içeren bu modelde öğrenme süreci boyunca ayrıştırıcı, üretici tarafından oluşturulan görselin gerçeki olup olmadığı konusunda belirteç olarak yer alır. Üreticinin oluşturduğu görsele sürekli olarak muhalefet ayrıştırıcı ağ üreticiyi yenmeye çalışır, ayrıştırıcının tespit ettiği görseli tekrardan kodlayan üretici ise ayrıştırıcıyı, daha gerçekçi görsel oluşturarak kandırmaya çalışır. İki farklı derin öğrenme ağının birbirine karşı yarıştığı bir modelde daha iyi sonuçlar ele edinilmesi böylece kaçınılmaz olur. GAN temelli modellerin başarısı eğitildikleri gerçek-sahte görsel veri seti ile de doğru orantılıdır. Üretici ve ayrıştırıcıyı yeterince sayıda veri ile beslemek modelin sonuçlarını yüksek oranda artırmaktadır.

### 2.3. Deepfake Tespit Modelleri ve Yöntemleri

Yapay zekâ alanında gelişmeler ile gerçekçi deepfake'lerin oluşturulması tespit yöntemlerinde de gelişmeyi mecbur kılmıştır. İnsan algısının kavramada güçlük çektiği oldukça gerçekçi görünen deepfake'ler çeşitli açılardan tehdit oluşturmaktadır. Bu nedenle çoğunlukla derin öğrenme yöntemleri kullanılarak gerçekleştirilen deepfake tespiti literatürde sıklıkla çalışılan bir konudur. Günümüzde hiç var olmayan da ya da gerçek bir yüz üzerinden manipülasyon yapılarak sentezlenen imajlar oluşturmak; veri setlerine kolayca ulaşabilmek ve derin öğrenme tekniklerinin gelişmesi ile kolaylaşmıştır (Tolosana, Vera-Rodriguez, Fierrez, Morales ve Ortega-Garcia, 2020). Bu nedenle hem deepfake oluşturulmasında hem de tespitindeki gelişmelerin kaçınılmaz bir gerçek olduğu savunulabilir. ZAO (Sorensen, 2024) ve FaceApp (Warzel, URL-1, 2019; Ry, URL-2, 2019; Wirth, 2023) gibi açık kaynak kodlu, ücretsiz filtreleme uygulamaları bu alanda hiçbir bilgisi olmayan için dahi kolayca sahte fotoğraf ve video oluşturulmasına olanak sağlamıştır.

Akademik çalışma istatistiklerinin elde edilebildiği Dimensions.ai web sitesinden elde edilen istatistiklere göre Şekil 2.2’de 2015’den günümüze kadar yapılan deepfake tespit çalışmalarının toplam artış grafiği gösterilmektedir.

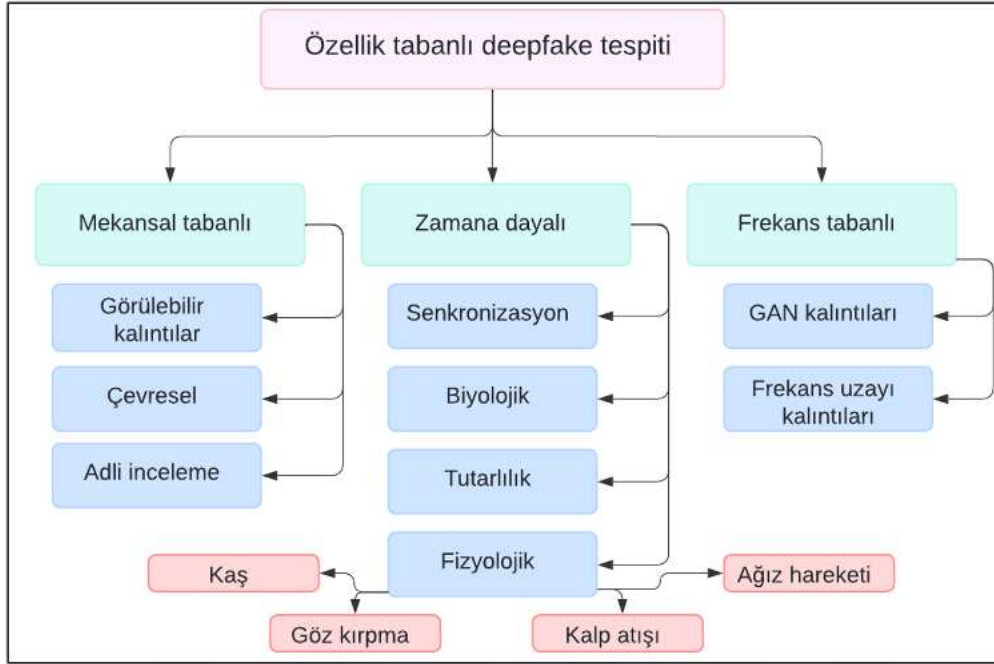


Şekil 2.2. “Deepfake oluşturma” ve “deepfake tespiti” geçen çalışmaların sayısının yıllara göre artış grafiği

Gong ve Li (2024) çalışmalarındaki incelemeler ve Dimensions.ai adresinden alınan istatistiklere göre Şekil 2.1 ve Şekil 2.2 kıyaslandığında deepfake oluşturma yöntemleri üzerine 9 752 kez çalışıldığı, tespit yöntemleri üzerine ise 11 111 çalıştığı görülmekte olup, deepfake tespit yöntemlerinin literatürde %14 daha fazla yer kapsadığı görülmüştür.

Deepfake tespitte çeşitli modeller ve insan karakteristik özellikleri baz alınarak bu özellikler üzerinden sahtecilik ve sentetik üretilmiş içerik tespiti sağlanmaktadır. Bu alanda yapılan özellik bazlı deepfake tespit yöntemlerine dair (Sandotra ve Arora, 2023) çalışması tarafından önerilen bir taksonomi oluşturulmuştur. Tez çalışmasının konusu olan atıklar üzerinden deepfake tespiti bu taksonomide farklı başlıklar altında çeşitli açılardan yer almaktadır.

Sandotra ve Arora (2023) tarafından oluşturulan özellik tabanlı deepfake tespiti taksonomi çalışması Şekil 2.3’te sunulmaktadır.



Şekil 2.3. Özellik tabanlı deepfake tespiti için önerilen taksonomi (Sandotra ve Arora, 2023)

Deepfake tespiti üzerine yapılan öğrenme tabanlı çalışmaların model bazlı sınıflandırmasında ise derin öğrenme çalışmalarının %70, makine öğrenmesi çalışmalarının %20 ve istatistiksel öğrenme çalışmalarının geriye kalan %10'luk seviye içerisinde olduğu (Sandotra ve Arora, 2023) görülmüştür. Diğer bir çalışmaya (Rana, Nobi, Murali ve Sung, 2022) göre ise derin öğrenme ile tespit alanında gerçekleştirilen çalışmaların %78'inin CNN (Oshea ve Nash, 2015), %12'sinin RNN ve LSTM (Sherstinsky, 2020) ve %2'sinin RCNN (Girshick, Donahue, Darrell ve Malik, 2014) üzerine odaklandığı tespit edilmiştir.

Juefei-Xu ve diğerleri (2022) tarafından yapılan çalışmada deepfake tespit yöntemleri, güncel zararlı deepfake oluşumları ve genel çerçeve bir çerçeve çizilmiştir. Buna göre deepfake oluşturulması dört başlık altında incelenmiştir. Bunlar, kimlik değiştirme, yüz ifadesi değiştirme, özellik manipülasyonu ve tüm yüz sentezleme olarak belirtilmiştir. Bazı (Ahmed, 2023; Damer, 2021; Zhang, Liu, Li ve Qi, 2023) çalışmalarda yüz morfolojisi değiştirme ayrı bir sınıflandırma başlığı olarak incelenmektedir. Görseldeki kişiye ait yüzün başkalaştırılması yöntemi ile oluşturulan bu değişim için sinir ağları kullanılan yöntemler kullanıldığı görülmektedir.

Peng, Gao ve Wang (2021) tarafından yapılan çalışmada deepfake oluşturma yöntemleri, sınıflandırılması ve tespit yöntemleri araştırılmıştır. Çalışmanın aktif olarak tuttuğu

veritabanı (URL-3) adresi sayesinde güncel yöntemlere de ulaşılmaktadır. Çeşitli çalışmalarda da deepfake oluşturulmasını (Juefei-Xu ve diğerleri, 2022; Pei ve diğerleri, 2024) çalışmalarında olduğu gibi dört başlık altında sınıflandırılmıştır. Çalışmanın sağladığı veritabanı sayesinde bu sınıflandırmalar altındaki deepfake oluşturma yöntemlerine ve çalışmalarına ulaşmak mümkündür.

## 2.4. GAN ile Deepfake Oluşturulması

Derin öğrenme alanındaki en önemli çalışmalardan olan GAN, mimarısı sayesinde kendi içinde çekişmeli olarak üretim yapmakta ve gerçekçi deepfake oluşturabilmektedir. GAN ile deepfake oluşturulması literatürde sıkça çalışılan bir konu olup ve günümüzde yüzlerce farklı GAN türü üzerinde çalışmalar sürdürülmektedir. (URL-4). Literatürde GAN kullanılarak ya da GAN'dan faydalanılarak yapılan çalışmaların yıllar içerisindeki sayısı katlanarak yükselmiştir. Xia ve diğerleri (2022) tarafından yapılan çalışmaya göre 2017 yılında 3459 olan bu sayı, 2019 yılında 22976, 2021 yılında 32400 olarak ölçülmüştür.

Wang, Guo, Hu, Chang ve Lyu (2022) çalışmasına göre GAN ile deepfake oluşturulması ve oldukça dikkat çekmesi modelin başarımının çok iyi olması nedeniyle oldukça gerçekçi sentetik görseller oluşturabilmesi nedeniyle kaynaklanmaktadır. Resim 2.1'de gösterilen görsellerin tamamı GAN yöntemleri ile üretilmiştir. Gerçekçi duran bu resimdeki deepfake'ler sırasıyla StyleGAN (Karras, Laine ve Aila, 2019), StyleGAN2 (Karras ve diğerleri, 2020), ve StyleGAN3 (Karras ve diğerleri, 2021) ile üretilmiştir. Bu görsellerin diğer GAN metotlarından farkı ise mevcut bir yüz üzerinde değişiklik yapılması üzerine sentezlenerek üretilmesidir; farklı bir yüzden transfer, görselden-görsel ve benzeri bir üretim yöntemi uygulanmadığı aktarılmıştır. Bu yöntem literatürde GAN tabanlı tüm yüz sentezi (GAN-based entire face synthesis) şeklinde ifade edilmiştir.

StyleGAN ve daha güncel versiyonları ile üretilen örnek sentetik görseller Resim 2.1'de sunulmaktadır.



Resim 2.1. StyleGAN ve versiyonları ile üretilmiş deepfake görselleri

Farklı versiyonları ile yıllar içinde gelişen StyleGAN'ın çoğu GAN'ın sahip olduğu çeşitli problemleri çözerek ilerlediği Wang (2022) çalışmasında gözlemlenmiştir. Juefei-Xu ve diğerleri (2022) çalışmasına göre GAN modeli kullanılarak deepfake oluşturulması incelenmiş ve bir taksonomi üretilmiştir. Bu taksonomiye göre GAN kullanılarak yüz manipülasyonu dört sınıfa ayrılmış ve bu alandaki GAN modelleri incelenmiştir. Bu sınıflandırma, tüm yüz sentezi, kimlik değiştirme, ifade değiştirme ve nitelik manipülasyonu olarak sunulmuştur. Buna göre yüz sentezi için DCGAN, WGAN, PGGAN, StyleGAN ve türevleri alanın öncüsü çalışmalar olarak kaydedilmiştir. Kimlik değiştirme için Autoencoder-based, MegaFS, GAN-based, FaceInpainter çalışmaları örnek gösterilmiştir. İfade değiştirme çalışmaları için Face2Face, ICface, A2V ve SVGAN modelleri incelenmiştir. Nitelik manipülasyonu çalışmalarının ise GAN odaklı çalışmalardan StarGAN, AttGAN, STGAN ve TediGAN olmak üzere sınıflandırmıştır. GAN modelleri kullanılarak yapılan çalışmalarda üretilen deepfake'lere örnek olarak Resim 2.2'de dört farklı model tarafından oluşturulmuş deepfake'ler gösterilmiştir. Üretilen deepfake'ler sırasıyla PGGAN (Karras, 2017), StarGAN (Miyato, 2018), StyleGAN (Karras, 2019) ve SC-FEGAN'a (Jo and Park, 2019) aittir.



Resim 2.2. Farklı GAN modelleri tarafından üretilmiş deepfake imajlar

Karras, Aila, Laine ve Lehtinen (2017) çalışmasında yüksek kaliteli deepfake'ler oluşturulması için GAN modeli geliştirilmiştir. PGGAN ismi verilen bu model ile literatürdeki veri setlerindeki düşük çözünürlüğe alternatif yüksek çözünürlüklü yüz görselleri oluşturulması hedeflenmiştir. Görsel kalitesini ilerletmek için “JPEG kalıntı silme” işlemi uygulandığı belirtilmiştir.

Brock, Donahue ve Simonyan (2018) çalışmasında görsel ölçekleyerek kaliteli hale getirme - çözünürlük artırma üzerine çalışmıştır. BigGAN ismi verilen bu çalışmada GAN ağlarının kararlılığının (stabilizasyon) artırılması ile yüksek doğruluklu (fidelity) ve daha fazla çeşitte görsel sentezlenmesinin mümkün olduğu gösterilmiştir.

Ortaç, Doğan, Orman ve Şamlı (2020) çalışmasında 270 ve 290 epoch'a ulaşıldığında gerçekçi bebek yüzleri üretilbildiği kanıtlanmıştır. İlerleyen epoch değerinde GAN modelinin hataya düştüğü; düşük epoch değerlerinde ise gerçekçi olmayan ve görülebilir kalıntılar içeren imajlar elde edildiği paylaşılmıştır.

#### 2.4.1. GAN modelleri ve yaklaşımları

Çeşitli GAN türleri literatürde farklı açılardan çalışılmış ve çalışılmaya devam etmektedir. Bunlardan bazıları yüz ve vücut değiştirme, sahtecilik üzerine geliştirilmiş olup örnekleri sunulmuştur. RSGAN (Natsume, Yatawaga ve Morishima, 2018), yüz ve saç özelliklerini kullanarak yüz değiştirme işlemi uygulayabilen modeldir. StarGAN (Choi ve diğerleri, 2018), yüz özellikleri ve yüz ifadelerini kullanarak yüz değiştirme işlemi uygulayan bir GAN modelidir. GANimation (Pumarola, Agudo, Martinez, Sanfeliu ve Moreno-Noguer, 2020), bir yüzün hatlarını aktarırken aktarılan kimliğini bozmadan sentezleme gerçekleştirebilen GAN modelidir. PGGAN (Karras, Aila, Laine ve Lehtinen, 2017), yüksek kaliteli görseller

oluşturmak için kullanılan ve öğrenme sürecindeki ek katmanları sayesinde gerçekçi deepfake'ler üretebilen GAN modelidir. BiGAN (Donahue, Krahenbühl ve Darrell, 2016), GAN yapısına ek olarak bünyesindeki kodlayıcı aldığı rastgele gürültü değişkenini kullanarak tespit edilmesi zor deepfake üretmek için geliştirilmiştir. CBiGAN (Carrara, Amato, Brombin, Falchi ve Gennaro, 2021), BiGAN'ın genelleme kayıplarını aza indirmek ve daha düşük güç tüketiminde öğrenme sürecini tamamlayabilmek için geliştirilmiştir. WGAN (Arjovsky, Chintala ve Bottou, 2017; Gong, 2017), standart GAN modeline göre optimizasyon sürecinde gelişmiş stabilite ve üreticinin yakınsaması ile kayıp fonksiyonu arasındaki korelasyon nedeniyle daha kaliteli çıktı verdiği belirtilmiştir.

Sağlık alanında tespit anlamında çeşitli açılardan geliştirilen ve kullanılan GAN modelleri literatürde sıkça yer almaktadır. BigBiGAN ve cGAN, korona virüs kaynaklı akciğerlerdeki değişimleri tespit etmek üzere geliştirilmiştir. Spine-GAN, insan omurgasındaki mantıksal segmentasyon üzerine geliştirilmiştir. SeGAN, beyin tümörünü tespit üzerine geliştirilmiştir. AnoGAN, göz retinasındaki lezyonları tespit etmek üzerine geliştirilmiştir. Medikal teşhis koyma alanında geliştirilen diğer GAN modelleri çalışmada kapsamlı bir şekilde işlenmiştir (Xia ve diğerleri 2022).

Endüstriyel alanlarda çeşitli yapı ve malzemelerde hasar, bozulma gibi kusurları tespit etmek üzerine de birçok GAN çeşidi ile çalışma yapılmıştır. DCGAN ile çelik yüzeylerde, güneş panellerinde ve ahşap yüzeylerde kusur tespiti çalışmaları yapılmıştır. WGAN ile düğme, yol çatlakları ve yüzey bozuklukları üzerine literatürde çalışmalar yer almaktadır.

Farklı kaynaklardan elde edilen veriler doğrultusunda görsel sentezleme işlemi gerçekleştirebilen GAN modelleri de mevcuttur. Bunlardan Stack GAN (Zhang ve diğerleri, 2017) ile yazılı içerikten görsel sentezleme ve Discovery GAN (Isola, Zhu, Zhou ve Efros, 2017) ile stil transferi yapılabilmesi söz konusudur. GAN ile görsellerde tamamlama yapılabilmesi, eksik yüklenmeyen parçaların tamamlanması gibi işlemlerin yapılmasının da söz konusu olduğu belirtilmiştir (Saxena, Gupta, Maheshwari ve Maheshwari, 2020). Keleş ve Güvenoğlu (2023) çalışmasında kullanılan yöntemle GAN modeli kullanılarak sanat eserlerinin hasar alan kısımlarındaki eksikleri görüntüleri üzerinden giderme ve tamamlama üzerine çalışılmıştır. Çalışmada mozaik eserlerin görüntüleri üzerinden eksikliklerini

tamamlama üzerine çalışıldığı ve az hasarlı olarak nitelendirilebilecek eserlerde görüntü tamamlamada en yüksek başarı elde edildiği aktarılmıştır.

Jia (2022) çalışmasında GAN modelleri arasındaki farklar araştırılmıştır. GAN, StyleGAN2-ADA ve StyleGAN3 kullanılan çalışmada modeller arasından en gerçekçi görseller oluşturan modelin incelenmesi yapılarak GAN modelleri yarıştırmıştır. Google Colab ortamında geliştirilen projede StyleGAN2-ADA'nın en iyi sonuçlar verdiği aktarılmış ve sonuçları tartışılmıştır. Çalışmaya göre StyleGAN2-ADA'nın en iyi sonucu yaklaşık 5-6 günlük bir eğitim süresince ürettiği 3000 görselin incelenmesi ile elde edildiği aktarılmıştır. Bu çalışmada paylaşılan hiper parametreler kullanılmak üzere kaydedilmiştir.

Ortaç, Doğan, Orman ve Şamlı (2020) çalışmasında 673 renkli 200x200 çözünürlüklü veri setinden GAN modeli kullanılarak gerçekçi görünümde bebek yüzleri üretilmesi başarılmıştır. Çalışmada 210 epoch sonrası bebek yüzü olduğu anlaşılabilir fakat kalıntılar içeren ve sentetik olduğu belli olan görseller üretildiği; 270 ve 290 epoch arasında gerçekçi görseller üretildiği 300 epoch sonrasında ise bazı hatalar alındığı aktarılmıştır.

Daha fazla GAN modeli hakkında detaylı inceleme için çeşitliliği nedeniyle hayvanat bahçesi benzetilmesi yapılan GAN-zoo (URL-5) veritabanı incelenebilir.

#### **2.4.2. GAN yüz tespit yöntemleri**

GAN ile üretilmiş deepfake'lerin tespiti ve GAN kullanılarak geliştirilmiş tespit yöntemleri literatürde oldukça geniş bir kapsama sahiptir. GAN bünyesindeki ayrıştırıcı bu anlamda GAN tarafından görsel sentezleme için üreticinin ürettiği görsellere karşı bu işlemi sürekli olarak uygulamaktadır. Wang, Guo, Hu, Chang ve Lyu (2022) tarafından yapılan ve Kumar ve Sharma (2023) çalışmalarına göre tespit yöntemleri derin öğrenme temelli, fiziksel özelliklerden tanımlama temelli, fizyolojik özellikler temelli ve insan görsel performansı temelli olmak üzere dört grupta incelenmiştir.

Aduwala, Arigala, Desai, Quan ve Eirinaki (2021) tarafından yapılan çalışmada GAN modeli eğitilerek bünyesindeki ayrıştırıcı kullanılmış ve deepfake tespiti için incelenmiştir. Eğitim süresince dörde bir oranında sahte ve gerçek görsel kullanılan bu yöntemde test sırasında ikiye bir oranında görsel kullanılmış ve 50 epoch'a ulaşıldığında %78 tespit elde

edildiği belirtilmiştir. Sonuç olarak GAN ayrıştırıcısının diğer tespit yöntemlerinden daha iyi olmadığı açıkça ifade edilmiştir. Bu durumu eğitildikleri veri seti ile yapılan ayrıştırmada, daha önce hiç görülmemiş bir veri seti ile yapılan deneme nedeniyle kaynaklandığı; bilinen veri seti ve üretici ile birlikte öğrenme durumunda daha iyi sonuçlar elde ettiği belirtilmiş; gerçek dünya örneklerine uygulamada iyi bir tercih olmadığı aktarılmıştır.

Fizyolojik özelliklerden tespit edilmesi üzerine yapılan çalışmalardan StyleGAN ile üretilen görsellerden göz rengi ve iris şeklinden deepfake tespiti yapılabildiği görülmüştür (Yang, Li, Qi ve Lyu, 2019; Matern, Riess ve Stamminger, 2019). Çocuklarda yüz hatlarının daha pürüzsüz geçişlerde olduğu gerçek görsellere karşı sahte görsellerdeki yüz hattı bozuklukları üzerinden tespitin çeşitli GAN modelleri için mümkün olduğu belirtilmiştir (Guo, Hu, Wang, Chang ve Lyu, 2022).

Fiziksel özelliklerden tanımlama temelli tespit yöntemlerinden kamera parametrelerinden, interpolasyon, sıkıştırma özelliği, renk filtresi dizisi, gözlerdeki ışık ile yansılardan tespit ve gözler arasındaki tutarsızlıktan tespit gibi çalışmalar GAN öncesinde çalışılmıştır (Johnson ve Farid, 2007). Bu çalışmalar, Matern, Riess ve Stamminger (2019) ve Karras, Aila, Laine, ve Lehtinen (2017) tarafından yapılan çalışmalara ışık tutmuştur. Çalışmalarda belirtilen tespit yöntemlerinin StyleGAN içine geçerli olduğu belirtilirken StyleGAN2 ile bu tutarsızlıkların ve kalıntıların da önlendiği belirtilmiştir (Wang, Guo, Hu, Chang ve Lyu, 2022).

İnsan algısı görsel tespit performansı üzerine yapılan çalışmalarda ise rastgele sunulan görsellerden sahte ve gerçek olduğu düşünülen görsellerin seçilmesi istenilerek yapılan çalışmalar mevcuttur. (Nickabadi, Fard, Farid, ve Mohammadbagheri, 2022; Nightingale ve Farid, 2022) çalışmasında 400 adet StyleGAN2 ile üretilmiş ve 400 adet gerçek görselden ayırt edebilme performansının katılımcılar arasında %50 oranında olduğu belirtilmiştir. Khodabakhsh, Ramachandra ve Busch (2019) çalışmasına göre ise katılımcıların en çok hatalı sınıflandırma yaptığı kategorideki içeriklerin CGI ve Face GAN temelli yapımlar olduğu belirtilmiştir. Çalışmada hatalı sınıflandırmanın detayları gerçek ve sahte (genuine - fake) arasındaki ayrım hakkında detaylar belirtilmemiştir. Derin öğrenme tabanlı tespit çalışmalarında gerçek görseller ve GAN tabanlı modeller ile gerçekleştirilen deepfake'lerin ayrıştırılması üzerine çalışmalar yapılmıştır. (Do, Na ve Kim (2018) çalışmasında DC-GAN tarafından oluşturulan 64x64 ve PG-GAN tarafından oluşturulan 256x256 ve 1024x1024

boyutlarındaki görseller ile CelebA veri setindeki gerçek görseller kullanılmıştır. Çalışmada “VGG-Face VGG16” yöntemi %80 oranında başarı elde etmiştir. (Mo, Chen ve Luo (2018) tarafından yapılan ve CNN kullanılarak deepfake tespiti gerçekleştirilen bir diğer çalışmada ise %99.4 oranında başarı sağlandığı belirtilmiştir. Marra, Saltori, Boato ve Verdoliva (2019) çalışmasında CycleGAN, ProGAN ve StarGAN tarafından üretilen sahte görsellerin tespiti üzerine çalışılmıştır. 256x256 boyutlarında kırılmış görseller kullanılan çalışmada ProGAN için %80, diğer modeller için %90 ve üzeri tespit başarımı elde edildiği belirtilmiştir. Belirli bir GAN modelindeki çalışması tamamlandığında diğer GAN modelleri tarafından üretilmiş görseller ile sınanan bu yapıda başarı oranlarının tatmin edici seviyede olduğu belirtilmiştir.

GAN bünyesindeki ayrıştırıcı sayesinde deepfake tespiti de mümkündür. Kumar ve Sharma (2023) çalışmalarında Deep Convolution GAN yöntemi kullanılarak açık kaynak olarak paylaşılan deepfake meydan okuma yarışmalarında (challenge) tespit oranını artırmak amacıyla geliştirilmiştir. Veri setinin görece sınırlı sayıda kalmasına karşılık GAN bünyesindeki gürültü girdisinden faydalanılarak veri çeşitliliği sağlanmış ve tespit doğruluğu anlamında iyi değerler elde edildiği belirtilmiştir.

#### **2.4.3. GAN-Üreteçli sentetik kalıntılar**

GAN üreteçli deepfake imajlar üzerine çeşitli tespit çalışmaları mevcuttur. Deepfake tespitinde yoğunlukla CNN ve türevi algoritmalar kullanılmaktadır. GAN kalıntıları üzerine tespit çalışmaları genellikle çeşitli GAN modelindeki katmanlardan kaynaklı görsellerde olan ortak kalıntılar üzerinden tespit, görsellerde oluşan tekrar ve sıklık ya da GAN modelinini yaptığı genelleme kaynaklı olduğu belirtilmektedir (Zhang, Karaman ve Chang, 2019; Shrivastava ve diğerleri, 2016). GAN kalıntısı üzerine Türkçe literatürde sadece bir çalışma (Altun ve Talu, 2021) yer almakta olup; mekân düzenleme ve obje silme alanındaki çalışmadaki (Ntavelis, Romero, Kastanis, Van Gool, ve Timofte, 2020) kalıntılar üzerine kullanılmıştır. İlgili çalışmadaki bu ifade kalıntı yerine ara katman olarak kullanılan yardımcı katman niteliğindedir.

GAN kalıntılarının tespiti ile deepfake görselin GAN olup olmadığını tespit ve tanımlama anlamındaki en iyi sınıflandırma çalışma tarafından yapılmıştır (Gragnaniello, Cozzolino,

Marra, Poggi ve Verdoliva, 2021). Çalışma bu anlamda GAN kalıntısını üç kategoriye ayırmıştır. Bunlar;

- Mekansal alandaki özellikler,
- Frekans uzayındaki özellikler,
- Genelleme kaynaklı özellikler olarak listelenmiştir.

Çeşitli kaynaklarda GAN kalıntılarını bir veya daha fazla kategori altında incelendiği görülmüştür. Wang, Guo, Hu, Chang ve Lyu (2022) çalışması bu sınıflandırmayı görülebilir kalıntılar ve görülmez kalıntılar olarak ayırmış ve gözlemlenemeyen kalıntıları üç başlık altında ele alınmıştır. Görülmez kalıntılar; Fourier spektrumu, frekans uzay analizi (frequency) ve frekans spektrumu altında incelenmiştir.

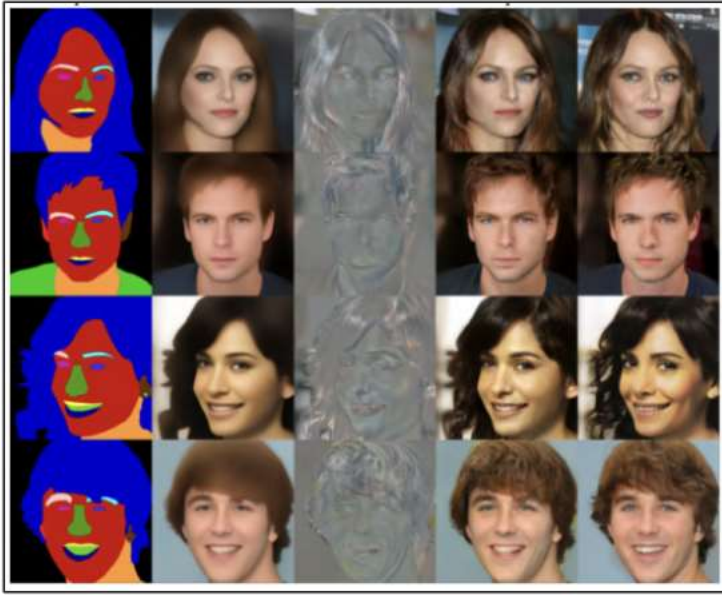
Yang, Li, Qi ve Lyu (2019) çalışmasında GAN üreticilerine sağlanan rastgele gürültü girdisinin yüz hatlarını tasvirlerde iyi bir parametre olduğunu fakat farklı yüz bileşenleri özelinde kısıtlamalara takılması nedeniyle konumlandırmada hatalar yapabildiğini ve tespit edilebilirliği artırdığını belirtmiştir. PGGAN'ın bu konuda tespit edilebilir olduğunu, bazı örneklerinde görülebilir kalıntılar nedeniyle tespitin fiziksel özelliklerden yapılabileceği gösterilmiştir. Buna karşılık STGAN'ın diğer GAN metotlarının aksine gürültü düzeyinden sentezlenmesi bu yöntemle tespiti konusunda uygulanabilir olmadığı anlamına gelebileceği bahsedilmiştir. Çalışmanın kalıntı ve anormallik olarak nitelendirdiği gözlemlenebilir kalıntılar Resim 2.3'te gösterilmektedir.



Resim 2.3. Gözlemlenebilir kalıntılara örnek üç GAN-üreteçli deepfake (Yang, Li, Qi ve Lyu, 2019)

Suri, Meshry, Davis ve Shrivastava (2024) çalışmasında kalıntılar için “residual” ifadesi kullanılmış ve kalıntıları Spectral yoğunluk açısından incelemiştir. GAN kalıntılarını

CelebAMask-HQ veri seti kullanarak elde etmeyi başaran ve inceleyen çalışma, görselden-görsele dönüşüm sağlanması ile oluşan yeniden yapılandırma kalıntılarını elde ederek inceleme yapmıştır. Yeniden oluşturma ile oluşan GAN kalıntıları çalışmada gösterildiği üzere Resim 2.4’de sunulmuştur. Çalışma, GAN artıklarını yakalayarak yerel varyasyonları yüksek frekansta ele almaktadır. CelebAMask-HQ veri seti için test kümesine karşılık gelen tüm sentezlenmiş çıktıların ortalamasını RGB uzayında inceleyerek sonuç elde etmektedir.



Resim 2.4. Girdi, yeniden oluşturma, GAN kalıntıları ve çıktılar üzerinden kalıntıların karşılaştırılması

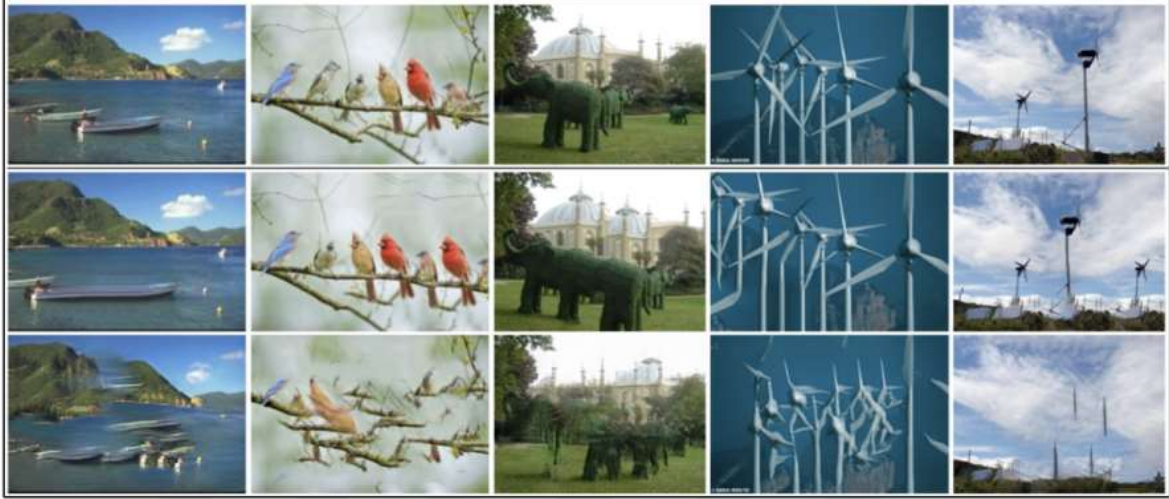
Çalışma, frekans spektrumunda yaptığı ölçümler ile yeniden oluşturulmuş sentetik görselde daha yüksek şiddette düşük frekans bilgisi olduğunu tespit etmiştir. Kalıntıların yüksek frekansı daha çok çektiği belirtilmiştir.

Bau ve diğerleri (2018) çalışması GAN sentezli görsellerde kalıntılara sebep olan noktaları araştırmıştır. GAN ile üretilen çalışma odası görsellerine odaklanarak problemi saptamayı hedefleyen çalışma, öge, obje ve olay yeri açısından GAN sentezli görseller üzerinde çalışmış ve GAN modellerinin öğrendiği çeşitli konseptler nedeniyle genelleme yaptığını belirtmiştir. Buna göre GAN modellerinin, konferans odasında koltuklar var olmalıdır ve yatak odasının penceresi olmalıdır gibi düzenleri öğrendiğini ve bu düzenlerin istenmediği durumlarda GAN’ların zorlandığı ya da GAN’lardan istenilenin elde edilmesinin güçleştiği aktarılmıştır.

Literatürde GAN tarafından üretilmiş görselleri tespit etmek üzerine yapılmış çalışmada ise bu renk geçişlerindeki tutarsızlık, simetrik tutarsızlıklar, yapaylık belirtisi (fingerprint), spektrum analizinde sapma şeklinde belirtilmiştir (Gragnaniello, Cozzolino, Marra, Poggi ve Verdoliva, 2021). Literatürde GAN kalıntılarına odaklanan diğer bir çalışma ise bu kalıntıları, bölgesel yapaylığının analizi, harmanlama-geçiş bölgesi kısımlarındaki keskinlik analizi, doku analizi ve frekans düzeyindeki yapaylıklar ile tespit etmeyi incelemiştir. GAN çıkarım (inference) odaklı kalıntılara odaklanan literatürdeki diğer bir çalışmada ise GAN ile üretilmiş yüksek çözünürlüklü görsellerdeki kalıntıların karakteristiğine odaklandığı ve bu kalıntıların diğer çalışmalardaki gibi öğrenim aşamasında cezalandırma yöntemi ile giderilmesinden farklı bir yöntemle tespit edilip giderilebileceği savunulmuştur (Xie ve diğerleri, 2023). Literatürde doğrudan GAN üreteçli yüzlere odaklanan bir çalışmada ise (Wang, Guo, Hu, Chang ve Lyu, 2022), renk bozuklukları kalıntıları, Fourier spektrum kalıntısına, frekans analizi ve frekans spektrumu kalıntılarına odaklanarak tespit çalışması yürütülmüştür. Bir diğer çalışmada ise (Alkishri, Widyarto ve Yousif, 2024) belirli GAN modellerinin stabil olarak aynı izleri (fingerprint) bıraktığını ve bu izlerin silinebilmesi ile tespitin zorlaşabileceği tezi öne sürülmüştür. Renk ipuçlarından yola çıkarak GAN kalıntısı algılama (McCloskey ve Albright, 2018) ve saturasyon ipuçlarından yola çıkarak GAN kalıntısı algılama (McCloskey ve Albright, 2019) sentetik üretilen görsellerde foton algısı olmaması nedeniyle saturasyon analizi sonucunda gerçeklik algısının anlaşılabilmesi belirtilmiştir. Bu açıdan GAN kalıntısı ile tespit için RGB renk analizi ve saturasyon analizi gerçekleştirerek basit bir adli analiz ile dahi aşırı maruz kalan piksellerin sıklığı üzerinden GAN üreteçli görsel tespiti gerçekleştirebileceği dile getirilmiştir.

Granot, Feinstein, Shocher, Bagon ve Irani (2022) yaptıkları ve GPNN olarak isimlendirdikleri çalışmalarında tekli görsel oluşturmada GAN modellerinin uzun eğitim sürecine ihtiyaç duydukları belirtilmiştir. Klasik yöntemlerden en yakın komşu (nearest neighbors) algoritması kullanarak nitelik ve nicelik anlamında kaliteli görseller oluşturulduğu; tek bir görselden dahi gerçekliği ayırt edilemez içerikler üretebildiği ve saatler yerine saniyeler içerisinde çalışabildiği aktarılmıştır. Çalışmada yapılan karşılaştırma karşılaştırmada görülebilir kalıntılar ele alınmış ve karşılaştırmada SinGAN modeli ile oluşturulan görseller kıyaslanmıştır. Kıyaslamaya göre algoritmanın SinGAN'a (Shaham, Dekel ve Michaeli 2019) göre daha az ya kalıntı bıraktığı ve üretilen görsellerde de daha az

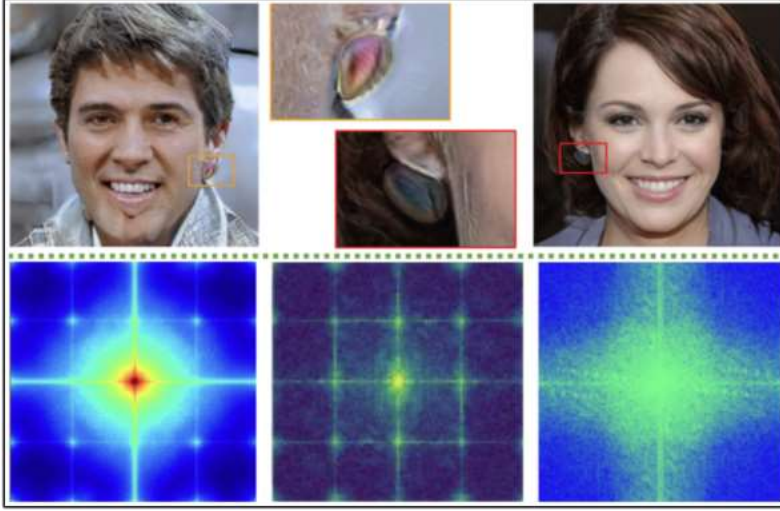
kalıntı ya da hiç görülebilir kalıntı bulunmadığı paylaşılmıştır. Resim 2.5’te bir örneği olarak aktarılmaktadır.



Resim 2.5. Orijinal görseller (birinci satır), GPNN (ikinci satır) ve SinGAN (üçüncü satır) tarafından üretilen deepfake örnekleri

Juefei ve diğerleri (2022) tarafından yapılan çalışmada GAN kalıntıları geniş bir çerçevede incelenmiştir. Deepfake tespit yöntemlerinden frekans tabanlı altında incelenen GAN kalıntıları; uzaysal tabanlı kalıntılar altında belirgin kalıntı izleri ve imaj adli bilişim inceleme yöntemleri altında; biyolojik sinyal tespit yöntemleri altında ise görsel tutarsızlıklar altında ele alınmıştır. İncelenen çalışmalarda bir GAN katmanı olan üst-örnekleme (upsampling) kaynaklı kalıntılar, istatistiksel, yönlendirilmiş degrade ve blob açısından frekans uzayında tespit edilebilen kalıntılar, piksel eş-oluşluk matrisi ile tespit edilebilen kalıntılar ve uzay-mekan imaj incelemesi ve faz spektrumu birlikte çalışılan kalıntı tespit yöntemleri ele alınmıştır. Masi, Killekar, Mascarenhas, Gurudatt ve AbdAlmageed (2020) tarafından yapılan çalışmayı da inceleyen bu çalışmada sentetik yüz tespit etmede renk uzayı ve frekans uzayını birlikte ele alan çalışmanın sonuçları da paylaşılmıştır. Farklı veri setlerine karşı oldukça başarılı sonuçlar elde edildiği aktarılmıştır.

Wang, Guo, Hu, Chang ve Lyu (2022) tarafından yapılan çalışmada frekans spektrumu ve frekans analizi ve Fourier dönüşümleri yapılarak görünmez kalıntıların tespiti üzerine inceleme yapılmıştır. Yapılan incelemeye göre analizler sonucunda frekans uzayındaki dağılımlar görselleştirilmiştir; Resim 2.6’da gösterilmektedir.



Resim 2.6. Görülebilir kalıntılar (birinci satır) ve ikinci satırda sırasıyla ortalama Fourier spektrumu, frekans analizi ve frekans spektrumu çıktıları

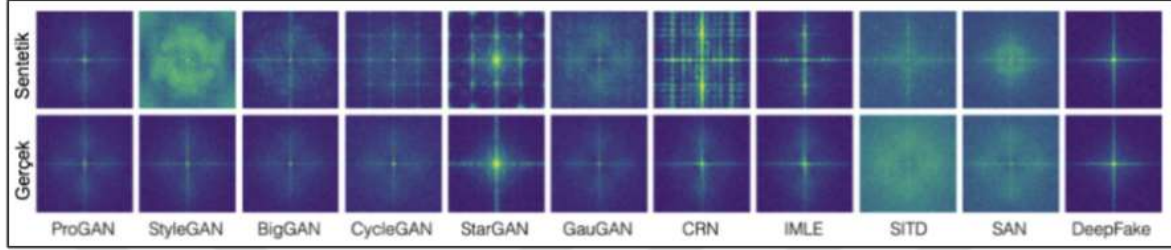
Görülebilir kalıntılar üzerine yapılan çalışmaları da listeleyen bu çalışma (Wang, Guo, Hu, Chang ve Lyu, 2022) insan görüsü performansı üzerinden gerçek-sahte çalışmalarını da değerlendirmiştir.

Gragnaniello, Cozzolino, Marra, Poggi ve Verdoliva (2021) tarafından yapılan çalışmada GAN tespit yöntemleri incelenmiş; eğitim ve test süreci arasındaki farklar, sıkıştırma ve boyutlandırma/ölçeklendirme gibi uygulamalar nedeniyle GAN tespiti için gerçek dünyada kullanılabilecek bir uygulama olmadığı aktarılmıştır.

Tan, Wen, Chen, Zeng ve Yang (2021) tarafından yapılan çalışmada StyleGAN tarafından üretilen görsellerdeki kalıntılar incelenmiş ve bunların giderilmesi üzerine çalışılmıştır. Kalıntıları uzamsal-mekansal açıdan inceleyen çalışmada 128x128 boyutundaki sentetik görseller incelenmiştir. Kalıntıların rastgele belirlediğini belirtilen çalışma birkaç piksellik bozulmanın dahi gözle görülebilir olduğunu göstermiş ve piksel örneği normalleştirme (pixel-instance normalization) kullanarak bu kalıntıları düzelttiğini göstermiştir.

Hackney (2023) çalışmasında kalıntıları, genellikle eğitim sürecindeki veri setinde alan ve bir düzen içinde olan noktalar, çizgiler veya dalgalardan kaynaklanan kalite bozuklukları şeklinde tanımlamıştır. Örnek olarak CatGAN modeli ile üretilen 128x128 RGB uzayındaki görsellerde yer alan dama tahtası desenli (checkerboard pattern) kalıntıları göstermiştir. Bu durumunda 16 görselden dokuzunda olduğunu saptamıştır.

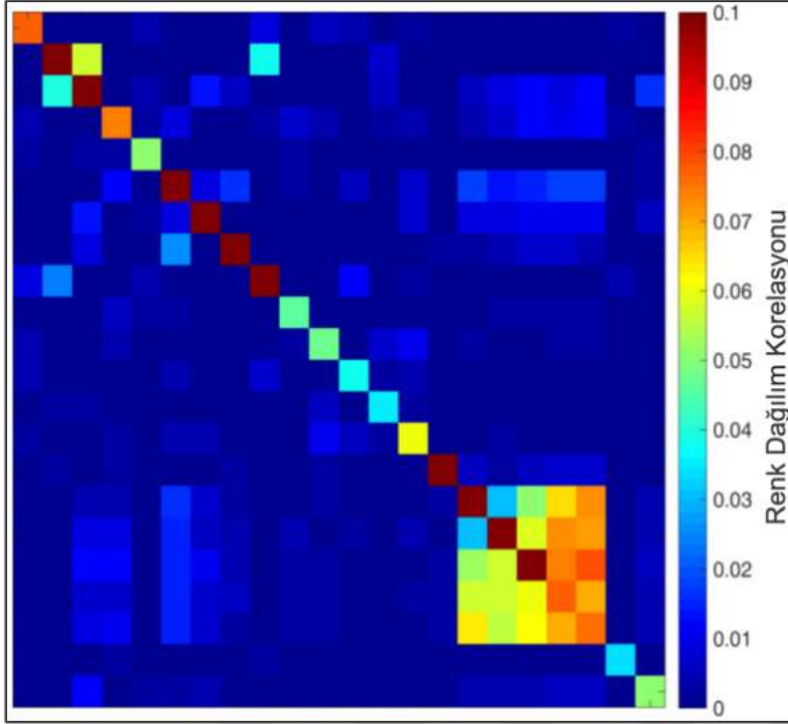
Wang, Wang, Zhang, Owens ve Efros (2020) tarafından yapılan çalışmada 11 farklı GAN modeli tarafından üretilen görseller kıyaslanarak incelenmiştir. ProGAN'ın en gerçekçi sentetik içerikleri üretildiğini belirten çalışma bu modellerin ters evrişimli (deconvolutional) katmana sahip olması nedeniyle dama tahtası desenli kalıntılar bıraktığını tespit etmiştir. Analiz sonuçlar Resim 2.7'de paylaşılmaktadır.



Resim 2.7. Çeşitli veri setlerine dair frekans analizi sonuçları

Her bir veri setine uygulanana frekans analizi sonucunda yüksek geçişli filtrelenenmiş ortalama spektruma ulaşılmıştır. Elde edilen sonuçlara göre BigGAN ve ProGAN'da diğerlerine göre daha az olmak üzere periyodik desenler görüldüğü aktarılmıştır (Wang, Wang, Zhang, Owens ve Efros, 2020).

Marra, Gragnaniello, Verdoliva ve Poggi (2019) çalışmasında yapılan ve parmak izi kalıntı araştırması korelasyonu Şekil 2.4'te verilmiştir. Şekil 2.4'te yer alan renk dağılım korelasyonu Nikon kamera, CycleGAN'la üretilmiş sentetik imajları, ProGAN ile üretilmiş sentetik imajları ve StarGAN ile üretilmiş sentetik imajları göstermektedir. Çalışma GAN olsun veya olmasın aynı veri seti ile oluşturulan görsellerde renk korelasyonu ile tespit edilebilmenin söz konusu olduğu aktarılmıştır (Marra, Gragnaniello, Verdoliva ve Poggi, 2019).



Şekil 2.4. İmajların ortalama kalıntı ve parmak izi korelasyonu

Jeong ve diğerleri (2022) tarafından yapılan çalışmada ikili yüksek geçişli filtreler kullanılarak frekans seviyesi (Fourier) kalıntıları büyütülerek (magnify) algılanmış ve hem piksel artıkları hem de frekans uzayındaki artıklar kullanılarak tespit deepfake tespiti gerçekleştirilmiştir.

Giudice, Guarnera, ve Battiato (2021) tarafından yapılan ve DCT parmak izi çıkarımı ile sahtenin yakalanması (Capture the Fake by DCT Fingerprint Extraction) çalışmasında GAN modellerinin frekans spektrumu incelenmesi ile fark edilmesi mümkün eşsiz bir parmak izi bıraktığı ve bununla birlikte sentetik içeriklerin tespit edilebileceği gösterilmiştir. Farklı sıkıştırma seviyelerinde 3.000 adet farklı çözünürlükte görsel üzerinden yapılan analizlerde StarGAN, StyleGAN ve StyleGAN2 ile üretilen sentetik içerikler için tespit işlemi gerçekleştirilebildiği belirtilmiştir.

Li ve Lyu (2018) tarafından yapılan sentetik yüz tespiti çalışmasında afin dönüşümü (ölçekleme, döndürme ve kesme) işlemi sırasında gerçekleştirilen işlemin yüz çevresindeki kalıntılarının algılanması ile sentetik yüz tespiti gerçekleştirilmiştir.

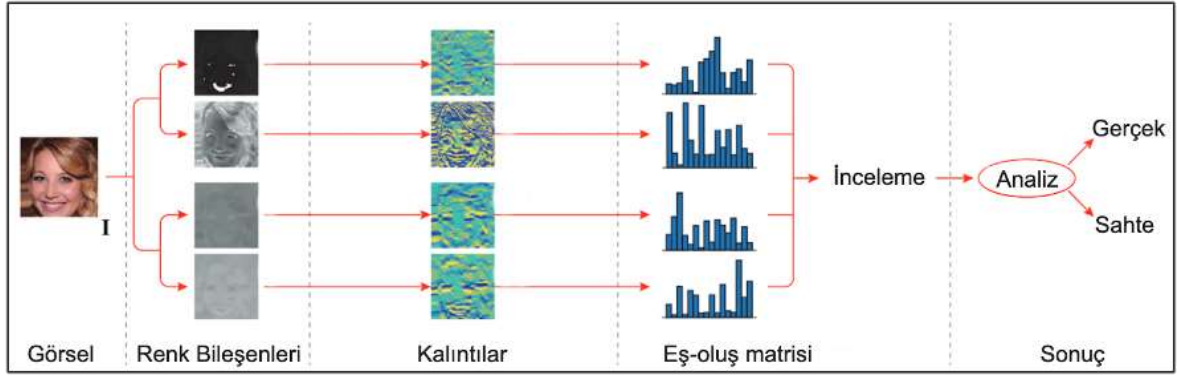
Barni, Kallas, Nowroozi ve Tondi (2020) tarafından yapılan çalışmada yüksek kaliteli sentetik görsellerde görülebilir kalıntıların olmadığı durumlarda tespitin dijital yöntemlerle yapılması gerektiği; spektral banttaki tutarsızlıkların incelenerek, bantlar arası eş-oluş matrisi ile sentetik görsel tespiti gerçekleştirilebileceği gösterilmiştir. CNN sınıflandırma algoritması kullanılan çalışmada piksel eş-oluş matrisi de kullanılmıştır.

Le ve Woo (2021) tarafından yapılan çalışmada Gaussian Mixture Model (GMM) ve Support Vector Machine (SVM) sınıflandırma yöntemleri kullanarak frekans uzayındaki istatistiksel özellikler kullanarak tespit hedeflemiştir. Fake Head Talker, StyleGAN ve StarGAN veri setleri üzerinden yapılan çalışmada GMM ve SVM sonuçları karşılaştırılmıştır. Eğitilmemiş SVM sınıflandırmasında %80 oranında tespit sonucu alınabildiği aktarılmıştır.

Dzanic, Shah ve Witherden (2020) tarafından yapılan çalışmada çeşitli GAN modelleri tarafından üretilen sentetik görseller ve gerçek görseller arasında yüksek frekans altında farklılıklar araştırılmıştır. Fourier spektrumunun yüksek frekanslardaki bozulmasını (decay) modelleyip gerçek görüntülerin daha yüksek bozulma oranları ve şiddet farklılığı gösterirken; sentetik görsellerin daha düşük şiddette olduğu tespit edilmiştir. Fakat, sıkıştırma algoritmaları kullanarak oluşturulan görsellerde bu farklılıkların azaldığı ve ayırt edilemez seviyelere geldiği aktarılmıştır.

Zhou, Han, Morariu ve Davis (2018) yaptıkları çalışmada RGB akışı ve gürültü (noise) akışı üzerinden manipülasyon tespiti gerçekleştirmiştir. Stenografi yöntemleri kullanılarak geliştirilen yöntemde steganaliz zengin modeli filtresi uygulayarak gürültü değerine ulaşılmıştır. Bu şekilde gerçek ve oynanmış (tampered) görsel arasındaki farklılıklar tespit edilerek manipülasyon algılanması hedeflenmiştir.

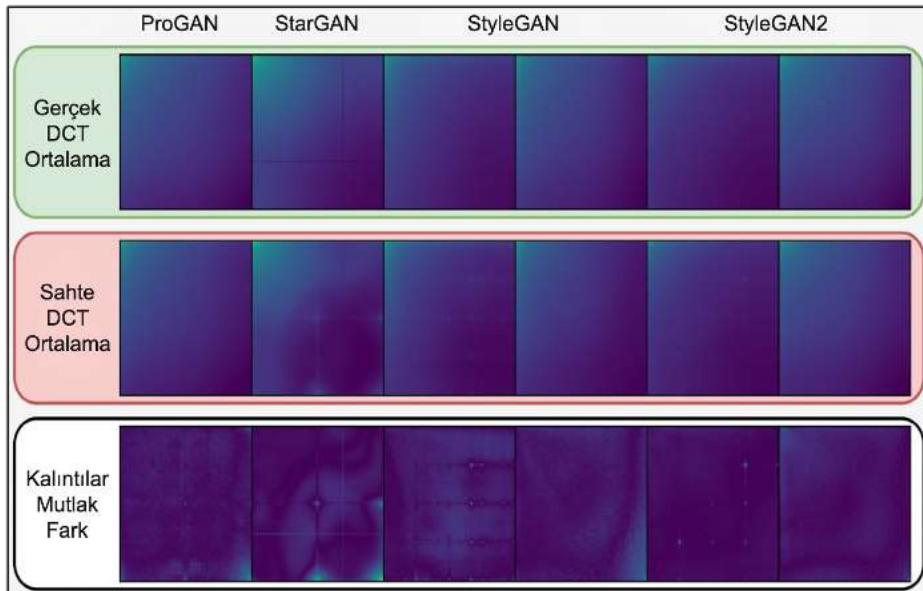
Li, Li, Tan ve Huang (2020) tarafından yapılan çalışmada renk bileşenlerindeki uyumsuzluk-tutarsızlıktan kaynaklanan anomalileri istatistiksel yöntemler ve eş-oluşum matrisi ile inceleyerek tespit gerçekleştirilmiştir. Derin öğrenme üreteçli olarak GAN modellerinden daha genel bir çerçevede sentetik görsel tespitini hedefleyen çalışma çeşitli GAN modellerini tanımlamada %50 gibi bir sonuç elde etmiştir. Çalışmanın çalışma-akış diyagramı Şekil 2.5'te gösterilmektedir.



Şekil 2.5. Renk bileşeni ve eş-oluş matrisi üzerinden ikili sınıflandırma yöntemi akışı

İkilik sistemde sınıflandırma (binary classifier) yöntemi uygulayan çalışma (Li, Li, Tan ve Huang, 2020) ayrıca RGB uzayında görülmeyen kalıntıların RGB, HSV ve YCbCr uzayında incelenmesini sağlayarak algılama yöntemini genişletmiştir.

Khoo, Phan ve Lim (2022) çalışmasında farklı GAN modelleri tarafından üretilen DCT spektral artıkları sunulmuştur. ForenSynths veri seti kullanılarak karşılaştırılması yapılan GAN modellerinin DCT analizinde yüksek frekans bileşenleri arasında belirgin bozukluklar-uyumsuzluklar olduğu belirtilmiştir. Frekans uzayında gömülü imzalar olarak isimlendirilen kalıntıların Şekil 2.6’da gösterilmektedir.



Şekil 2.6. Gerçek ve sahte görsellerin ortalama DCT spektral değerleri ve spektral kalıntılar farkı

Çalışmada ProGAN ile oluşturulan sentetik görselin Fourier dönüşümü altında diğer modellere göre belirgin kalıntıları olmadığı belirtilmiştir.

Wu, Xie, Gao ve Xiao (2020) tarafından yapılan çalışmada diğer çalışmaların aksine uzamsal özellikler, steganaliz inceleme yöntemi ve zamansal-geçici özellikler alanı birleştirilerek bir çalışma yürütülmüştür. Uzamsal özellikler CNN, yüksek seviyeli görüntü içeriği bastırma yöntemi ve zamansal-geçici özellikler LSTM tarafından çıkarılmıştır. FF++ veri seti kullanılan çalışmada steganaliz yönteminin önceki çalışmalara göre daha iyi sonuçlar verdiği aktarılmıştır.

Chhabra, Thakral, Mittal, Vatsa ve Singh (2023) tarafından yapılan çalışmada “görülebilirlik matrisi” ismi ile ifade edilen yöntem geliştirilmiş ve saklı kalıntılar üzerinden sahte içerik tespit etmeyi hedeflemiştir. Sıkıştırma nedeniyle bilgi kaybına uğrayan içeriklere karşı da çalıştığı aktarılmıştır. Sonuçlara göre %69 ile %73 arasında AUC skoru elde edildiği belirtilmiştir.

Gamazo Tejero (2020) tarafından yapılan çalışmada StyleGAN ile üretilmiş 256x256 boyutlarında 20 000 görsel üzerinden GAN üreticilerinin düşük filtre ile çalışması nedeniyle verilen görsellerin yüksek frekansta orjinal görsellere göre daha çok sapma gösterdiği aktarılmıştır. Bu sapmanın ve diğer bulguların GAN ezberlemesi (memorization) ile ilintili olduğu sonucuna ulaşılmıştır.

Chai, Bau, Lim ve Isola (2020) tarafından yapılan çalışmada GAN modellerinin genellemeleri üzerine çalışılmış ve GAN modelleri arasındaki ortak noktaların sentetik içerik tespitinde kullanılabilirliği araştırılmıştır. Bölgesel kalıntılar üzerine yoğunlaşan çalışma saç dokusu, arkaplan, ağız ve yüz bölgesinde incelemeler üzerine sınıflandırmanın parametrelerinde iyileştirmeler yaparak geliştirmiş ve tespit iyileştirmesi yapmıştır. Isı haritaları oluşturma tekniğini kullanarak olası manipülasyon bölgelerini tespit edip kalıntıları ölçekleyerek algılamaya çalışmıştır.

Zhu, Wang, Salloum, Chen ve Kuo (2022) tarafından yapılan çalışma GAN’ların yüksek kaliteli gerçek görsellerden üretilen içeriklerde yüksek frekans bileşenlerini oluşturamadığı varsayımı ile yola çıkılmıştır. Pikselin bölgesel komşu algoritması kullanılarak geliştirilen ve PixelHop olarak isimlendirilen filtreler kullanılan çalışmada ilişkili kanallar belirlenmiştir. Yerel düzenlere göre belirlenen çoklu PixelHop sayesinde aykırı kanallar

algılanarak tespit sağlanmıştır. GAN kalıntılarının detaylı olarak incelendiği çalışmada evrişimsel katmanlardan kaynaklı üst örneklemeden doğan bozulma ve asimetrik karakteristikler gibi kalıntılar incelenmiştir.

Kalıntı benzerliğinden deepfake tespiti gerçekleştirmeyi hedefleyen çalışmada (Li, He, Li, Wang ve Zhang, 2021) farklı GAN modelleri ve farklı veri setlerinden oluşan tespit güçlüğü gidermek hedeflenmiştir. Önerilen çözüm için 1024 x 1024 çözünürlüğünde 5000 adet PGGAN ve StyleGAN ile üretilmiş kullanılan çalışmada JPEG dönüşümü ve bulanıklaştırma yöntemleri uygulanarak testler gerçekleştirilmiştir.

Huang ve diğerleri tarafından yapılan ve FakeLocator ismi verilen çalışmada GAN tabanlı yüz sahtecilikleri tespiti üzerine çalışılmıştır. GAN modeli farketmeksizin sentetik içerik tespit etmeyi hedefleyen çalışma gri ölçekleme ve ikili temel gerçek haritası analizi ile tespit etmeyi hedeflemektedir. JPEG sıkıştırması, bulanıklık, gürültü ve düşük çözünürlüğe dönüştürme gibi işlemlerin uygulandığı çalışmada düşük çözünürlüğe dönüştürüp orijinal hale tekrar çevirmek dışındaki tüm işlemlerde başarılı sonuçlar alındığı belirtilmiştir. Kelleştirme, saçları sarıya dönüştürme ve gülücük ekleme üzerine yaptıkları sentetik üretimler sonucu elde edilen gri ölçekleme ve ikili gerçeklik haritası işlemleri sonucunda elde edilen görsel haritası Resim 2.8’de verilmiştir.



Resim 2.8. Soldan sağa, gerçek görsel, sentetik görsel, gri ölçekli temel gerçek haritası, ikili (binary) gerçeklik haritası

Çalışma bölgesel değişikliklerle sentetik görsel oluşturan uygulamalara karşı GAN agnostik bir çözüm üretmeyi hedeflemiştir. Test edilen GAN modellerinde başarılı sonuçlar verdiği görülmektedir.

## 2.5. Deepfake Veri Setleri ve GAN Üreteçli İmaj Veri Setleri

Deepfake veri setleri literatürde en yoğun olarak çalışılan bu alandaki konulardan birisidir. Farklı çözünürlüklerde ve içeriklerde birçok veri seti üzerinde çalışmalar yapılmaktadır. Veri setlerinde ağırlık ünlü figürlerin farklı zaman ve konumlardaki çekimleri baz alınarak oluşturulmuştur. Bunun sebebi ise özellikle youtube.com gibi çeşitli video kaynaklarında figürlerin farklı açılardan, farklı giyimlerde ve farklı zamanlarda dolayısıyla da farklı saç stili, farklı yüz ifadesi, farklı gözlük kullanımı açılarından çeşitlilik sağlanması oldukça mümkündür.

Ağ öğrenimi ya da model testleri için kullanılan deepfake veri setlerinde en çok kullanılan çalışmalar çıkarılmıştır. Bunlar; CelebA (Liu, Wange ve Tang, 2015), Celeb-DF (Li, Yang, Sun, Qi ve Lyu, 2020), UADFV (Yang, Li ve Lyu, 2019), DF-TIMIT (Korshunov ve Marcel, 2018), FaceForensics++ (Rossler ve diğerleri, 2019), Google DeepFake detection dataset (Dufour ve diğerleri, URL-6, 2019), FaceBook DeepFake detection challenge (Dolhansky, Howes, Pflaum, Baram ve Ferrer, 2019), FakeAVCeleb (Khalid, Tariq, Kim, Woo, 2021) ve HFM (Lee, Tariq ve Shin, 2020) olarak listelenmiştir. Heidari, Jafari Navimipour, Dag ve Unal (2024) çalışmasına göre %24 gibi bir oranla en çok FaceForensic veri seti kullanıldığı tespit edilmiştir. Diğer bir çalışmaya göre ise FaceForensic'in toplamda 52 ile çalışmada kullanıldığı en yakın çalışma olan 23 kere tercih edilen Celeb-DF'ye iki kattan daha fazla tercih edildiği belirtilmiştir.

Önceki çalışmalarda kullanılan Deepfake ve GAN tabanlı veri setleri taranmıştır. Bu veri setleri göre;

- CycleGAN veri seti (URL-7) çeşitli kategorilerde CycleGAN ile sentezlenmiş görseller içermektedir.
- StarGAN veri seti (URL-8) yüz görsellerinin farklı renk, yaş, cinsiyet ve duygular içeren sürümlerinden oluşmaktadır.
- generated.photos StyleGAN veri seti (URL-9) 100 bin görsel içermektedir.
- CycleGAN (URL-10) veri seti at ve zebra sentetik görselleri içermektedir.

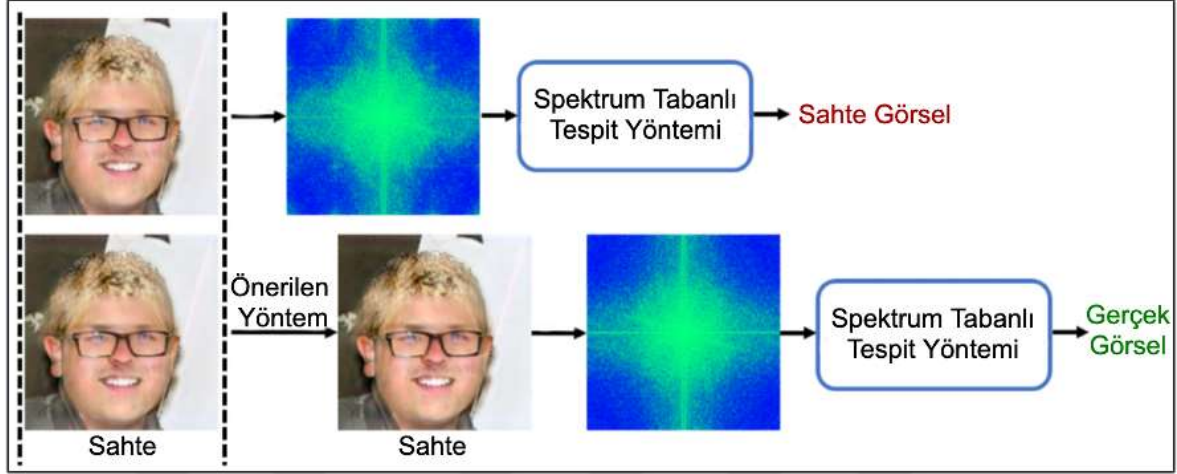
- Real-ESRGAN (URL-11) veri seti video ve görseller içermektedir.
- iFakeFaceDB veri seti (URL-12) StyleGAN ile üretilmiş ve GANprintR yöntemi kullanılarak dönüştürülmüştür.
- The DeepFake Detection Challenge (DFDC) veri seti (URL-13) Facebook tarafından yapılan bir yarışma için üretilmiştir. Kısmen GAN üretici kullanıldığı belirtilmektedir.
- Diverce Fake Face Dataset - DFFD (URL-14) içeriğinde GAN tabanlı sentetik görsellerin de yer aldığı veri setidir.
- 100K-Faces veri seti (URL-15) StyleGAN tarafından üretilmiştir.
- ProgressiveGAN veri seti olan CelebA-HQ (URL-16) çeşitli ortamlara ait 100 bin imaj içermektedir. Veri setinde görseller farklı çözünürlüklerde hazır olarak sunulmaktadır.
- BigDatasetGAN veri seti (URL-17) bünyesinde BigGAN ve Synthesizing ImageNet with Pixel-wise Annotations
- DCGAN veri seti (URL-18) DCGAN ile üretilen sentetik görseller içermektedir.
- Least Squares GAN ile oluşturulmuş veri seti LSUN (URL-19) mutfak, yatak odası ve araba gibi kategorilerde görseller içermektedir.
- BigGAN tarafından üretilen BigGAN-Deep veri seti (URL-20) 100 milyon adet farklı kategorilerde görseller içermektedir.
- StackGAN tarafından üretilen veri seti (URL-21) kuş ve çiçek görselleri içermektedir.

Ulaşılan veri setlerinin çeşitli deepfake tespit yöntemleri için kullanılan veri setlerinden GAN modelleri kullanılarak oluşturulduğu görülmektedir. Çoğu GAN modeli üzerine yapılan çalışmalar bu veri setlerinden üretilen imajları ilgili GAN modeli ile üreterek çalışmıştır. Bu nedenle GAN dışında bir yöntem içermeyen, bir veya birden fazla GAN modeli kullanılarak türetilmiş veri setleri literatürde yeterince bulunmamaktadır.

## 2.6. GAN Kalıntıları ve Gürültüsü Önleyici Yöntemler

Çeşitli çalışmalarda GAN kalıntılarının önlemek ve parmak izi analizi üzerinden yapılan analizi atlatmak amacıyla yöntemler geliştirildiği görülmüştür. Bazı çalışmalarda geliştirilen

yöntemlerin en iyi tespit yöntemlerinden olan spektrum analizini dahi atlatabildiği ifade edilmiştir. Bir örneği Şekil 2.7’de gösterilmektedir (Dong, Kumar ve Liu, 2022).



Şekil 2.7. Spektrum analizini atlama yöntemi çalışma yöntemi (Dong, Kumar ve Liu, 2022)

SpectralGAN (Dong, Kumar ve Liu, 2022) çalışmasının bulguları, sahte görüntü spektrumlarının analizi temelinde çalışan dedektörlerin, GAN tarafından üretilen sahte görüntüleri tespit etmede sağlam olmayabileceğini göstermektedir. Çünkü bu tür spektrumlarda bulunan artefaktlar, kötü niyetli saldırganlar tarafından kolayca giderilebilir ve bu işlemden sonra ilgili görüntüler görsel olarak çok benzer görünebilir. Bu nedenle, topluluğumuz için giderek artan zorluklar oluşturan gelişmiş GAN tabanlı modellerden sahte görüntüleri güvenilir bir şekilde tespit edebilecek sağlam sahte görüntü dedektörleri geliştirmek için önemli çalışmalar gerekmektedir (Dong, Kumar ve Liu, 2022).

Wesselkamp, Rieck, Arp ve Quiring (2022) çalışmalarında spektrum uzayındaki kalıntıları silerek deepfake tespitini önlemeye yönelik bir geliştirme sunmaktadır. Önemli olmayan ayrıntılar üzerinden manipülasyon-oynama gerçekleştirilerek yapılan bu işlem sonrasında GAN modellerine ait parmak izlerinin silinebildiği belirtmiştir. Çeşitli GAN modelleri ile test edilen bu yöntem için spektrum tabanlı tespit yöntemlerini atlama hedef alınmış uzamsal-mekansal kalıntılar üzerinden tespit gerçekleştiren sınıflandırıcı uygulamalara karşı efektif olmayabileceği belirtilmiştir.

Huang ve diğerleri (2020) tarafından geliştirilen FakePolisher yöntemi ile GAN modeli bilgisine sahip olmadan yüzeysel yeniden yapılandırma denilen görselin tekrardan oluşturulması ve kalıntılarının giderilmesi yöntemi sunulmuştur. Mevcut deepfake tespiti yapabilen yöntemleri atlatabildiğini belirten bu yöntem ile GAN modellerinin zayıflıkları

arařtırılmıř ve giderilmeye alıřılmıřtır. GAN modelleri tarafından oluřturulan kalıntıları gidermeye ynelik geliřtirilen bu alıřmada deepfake'lere karřı daha geliřmiř yntemlerin geliřtirilmesi gerektięi vurgulanmıřtır.

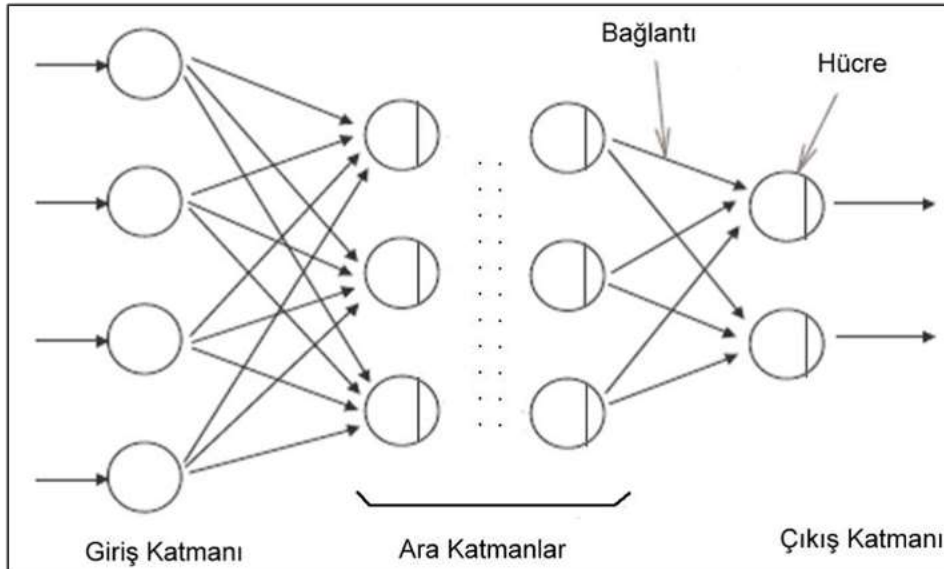


### 3. GAN KALINTILARININ İNCELENMESİ

Bu bölümde bir derin öğrenme yöntemi olan GAN mimarisi incelenmiş ve GAN kullanılarak gerçekleştirilen deepfake'lerin tespiti için GAN kalıntıları hakkında bilgiler belirtilmiştir. Çeşitli GAN modellerindeki kalıntılar nedeniyle sahtecilik tespiti için yöntemler ve bu yöntemlere karşı ilgili GAN kalıntılarını deepfake üretimi sırasında ortadan kaldırmak, sonradan temizlemek veya GAN genellemelerini en aza indirmek için üretilmiş yöntemler incelenmiştir.

#### 3.1. Yapay Sinir Ağları ve Derin Öğrenme

Yapay sinir ağları insan beyninin yapısı temel alınarak geliştirilmiş, nöronların birbirleri arasındaki iletişim mekanizması (sinaptik bağ) kurması ile oluşturduğu ağın bilgisayar programları ile benzetimine denilmektedir. Yapay sinir ağlarındaki yapay kelimesi ilgili ağın dijital olarak oluşturulduğunu karşılık gelir. Bu ağlar katmanlar halinde, her katmanda belirli sayıda nöronun yer aldığı bir yapı şeklindedir ve veri saklama, veri işleme, hafızaya alma, girdilere göre hesaplamalar yapma ve değişken ağırlıklarda karar mekanizmasına katkıda bulunma gibi çeşitli açılardan görev alırlar. Şekil 3.1'de bir yapay sinir ağı modeli (Öztemel, 2006) gösterilmiştir.



Şekil 3.1. Örnek yapay sinir ağı modeli (Öztemel, 2006)

Yapay sinir ağıları sayesinde ses algılama ve tanıma, işlenen görüntüden cisim tespiti, doğal dil işleme (natural language processing) ve öneri altyapı sistemi kurma gibi süreçler oluşturulabilmektedir. Yapay sinir ağlarının yapısında bulunan kayıp fonksiyonu ve ağırlık dağıtımı sayesinde ise ağ gelişir ve çıktıları sürekli olarak iyileşir hale gelir. Bu da yapay sinir ağlarının oldukça başarılı sonuçlar elde etmesine neden olur.

Derin öğrenme, yapay sinir ağlarının çok katmanlı hale getirilerek öğrenme sürecinde kullanılması ile gerçekleştirilir. Giriş katmanı, gizli katmanlar ve çıktı katmanı gibi çok katmanlı bir yapıda katmanlar sırasıyla elde ettikleri veriler üzerinde hesaplamalar gerçekleştirir ve sonraki katmana iletir. Nöronlar arasında olduğu gibi katmanlar arasında da bağlantılar mevcuttur ve çıktıdan elde edilen verilere göre katmanlar arasındaki ağırlıklar sürekli güncellenir. Öğrenim sürecinde bu değerlerin sürekli olarak güncellemesi sayesinde en iyi verim elde edilen ağırlıklara göre ağ eğitilir. Öğrenim süreci sonunda ise ağ istenilen çıktı için tespit edilmesi talep edilen en iyi sonuçlar elde etmeyi hedefler. Derin öğrenme/yapay zeka alanındaki gelişmelerle birlikte üretken ağlar kavramı ortaya çıkmıştır. Üretken ağ modelleri sayesinde çeşitli girdiler ile birlikte istenilen çıktı elde edilebilmektedir. Örnekleri günümüzde text-to-image, image-to-image sentezleme şeklinde yer almaktadır. Bu alandaki en önemli model ise GAN ve türleri olarak literatürde yer almakta; gerçek dünya uygulamaları görülmektedir.

### 3.2. GAN Mimarisi

GAN (generative adversarial networks) modeli ilk olarak 2014 yılında Goodfellow ve diğerleri (2014) tarafından ortaya atılmıştır. GAN, çekişmeli öğrenme yöntemi ile öğrenerek yeni veri üretmek için geliştirilmiş derin öğrenme yöntemidir (Xia ve diğerleri, 2022). GAN'ın bünyesinde iki farklı yapay sinir ağı bulunmaktadır. Bu ağlar birbirlerine sürekli olarak karşılık gelmektedir. Bu ağlardan üreteç olarak görev alan yapay sinir ağı ürettiği içerik sürekli olarak ayrıştırıcı tarafından test edilir. GAN'ın en yoğun olarak kullanıldığı alan olan görsel sahteciliği, deepfake açısından üretcin oluşturduğu görsellere sürekli olarak muhalefet eden ayrıştırıcı ilgili görselin sahte olduğunu belirtmesi durumunda üretecin tekrardan kendini iyileştirmesi gerektiği bilgisi ile ağı besler. Bu sayede üreteç yapay sinir ağı, ayrıştırıcı sinir ağını yenilgiye uğrayana kadar bu süreci sürdürür. Ayrıştırıcı sinir ağı ise hem gerçek görsellerden eğitildiği hem de üretcin sağladığı görsellerin sahte ve gerçek algısını sağladığı sürece kendini beslemeye - eğitmeye devam eder. Bu sırada üretcin sinir

ağı ise; rastgele girdi, girdiyi anlamlı hale getiren üreteç sinir ağı, ayrıştırıcı ağının çıktısı ve ayrıştırıcının tespit etmesi halinde oluşan üretici kayıp hesaplaması tarafından beslenir.

GAN mimarisinde üreteç ve ayrıştırıcının formülü ve matematiksel gösterimi, Kamar, Üre ve Ünal (2022), Yılmaz ve Korn (2023) ve Wang (2020) tarafından yapılan çalışmalarda yayınlanmıştır. GAN modelinde ayrıştırıcı ve üreteç modelin öğrenim aşamasındaki kayıp hesaplamasını Binary Cross Entropy (ikilik çapraz çeşitlilik fonksiyonu) ile hesaplar. Ayrıştırıcı sınıflandırma yaparken ikilik (binary) sistemde sonuç üretir. Tahmin ettiği sonuç ile gerçek değer arasındaki fark Binary Cross Entropy olarak hesaplanır ve formülü aşağıdaki gösterilmiştir.

Tanımlar;

- G: Üretici
- D: Ayrıştırıcı
- z: Gürültü değeri
- y: ikilik sistemde gerçek değeri
- p: ikilik sistemde tahmin edilen değeri
- $\mu$ : olasılık dağılımı
- $\gamma$ : başlangıç dağılımı
- E: dağılımla ilgili beklenen değer
- G(z): sentetik içerik
- D(x): İkilik sistemde ayrıştırıcı olasılık değeri

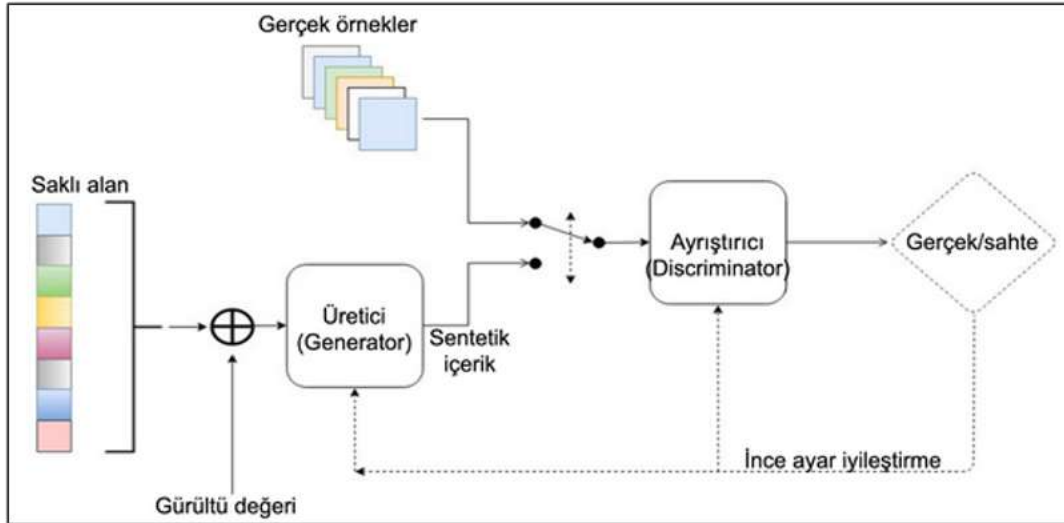
Çizelge 3.1’de gösterilen eşitliklere göre üretici sürekli olarak karşısında yazan eşitliği düşürmeyi hedefler; ürettiği içeriğin tespit edilmesini zorlaştırmayı amaçlar. Buna karşılık ayrıştırıcı Çizelge 3.1’de karşısında yazan eşitliği sürekli olarak artırmayı hedefler ve üretici tarafından oluşturulmuş sentetik içeriği tespit etmeyi amaçlar.

Çizelge 3.1. GAN mimarisi eşitlikleri

| Tanım                | Eşitlik                                                                                                        |
|----------------------|----------------------------------------------------------------------------------------------------------------|
| Üretici              | $\log(1-D(G(z)))$                                                                                              |
| Ayrıştırıcı          | $\log D(x) + \log(1-D(G(z)))$                                                                                  |
| Birleşik Kayıp       | $\min_G \max_D V(D, G) = \min_G \max_D [E_{x \sim \mu} [\log(D(x))] + E_{z \sim \gamma} [\log(1 - D(G(z)))] ]$ |
| Çapraz Eşitlik (BCE) | $- [y * \log(p) + (1 - y) * \log(1 - p)]$                                                                      |

GAN mimarisi eşitlikleri paylaşılan Çizelge 3.1’de üretici ve ayrıştırıcı sinir ağlarının matematiksel hedefleri sunulmuştur. Buna göre üretici ve ayrıştırıcı arasındaki matematiksel yakınsama GAN modelinin yapısındaki çekişmeli olma prensibini açıklamaktadır.

Dış dünyadan aldığı gerçek örnekler üzerinden sentetik görseller üreten ve ürettiği sentetik görselleri bünyesindeki ayrıştırıcı ile tespit ederek her iterasyonda iyileştirmeye giden GAN mimarisi Şekil 3.2’de gösterilmektedir (Chaudhari, Agrawal ve Kotecha, 2020).

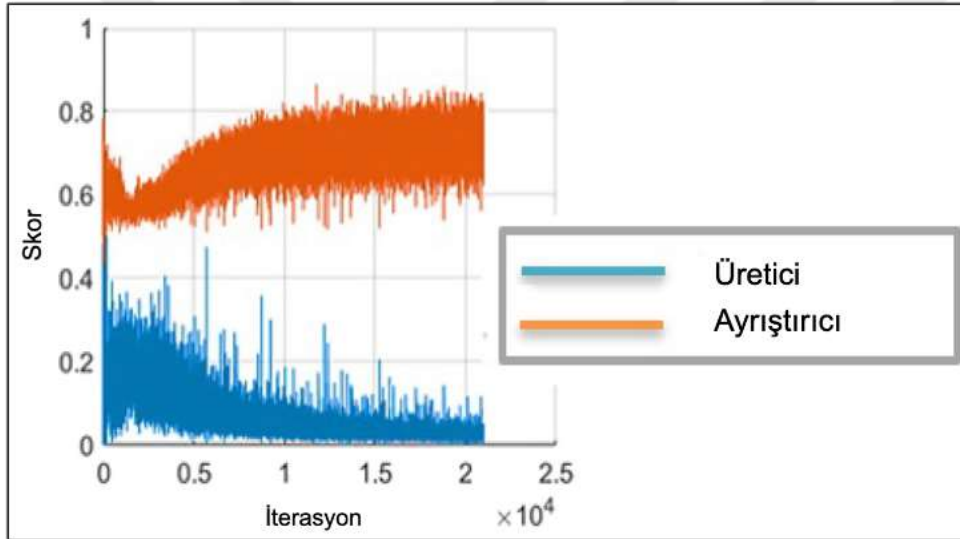


Şekil 3.2. GAN Mimarisi

Saklı alan (latent space) veya gizli alan olarak ifade edilen girdi kısmı GAN modelinde doğrudan gözlemlenebilir olmayan veya insan tarafından anlaşılır olmayan girdiyi ifade eder (Hackney 2023). Saklı alanın amacı veriyi daha düşük boyutlu uzaya indirgemek ve sinir

ağlarının veriyi daha verimli işlemesine olanak sağlamaktır. GAN modelindeki üretici sinir ağı bu saklı alanı veriyi doğru dağıtabilmek için kullanır ve gerçekçiliği artırmayı hedefler.

Üretici-ayrıştırıcı çekişmesi süreci, oyun teorisine göre zero-sum oyunu olarak adlandırılır (Mohebbi Moghaddam diğerleri, 2023). CGAN modeline ait bir çalışmaya ait üretici ve ayrıştırıcı arasındaki tahmin-tespit skor tabelası örneği gösterilmiştir. GAN modellerinde öğrenme derecesi, optimizasyon parametreleri ve veri kümesi uzayı (latent space) gibi değerler için hiper parametre (hyperparameters) ifadesi kullanılmaktadır. Hiper parametreler doğru tanımlanmadığı ve kalibre edilmediği takdirde GAN modeli öğrenim aşamasında çökebilir ya da anlamsız veriler üretebilir.

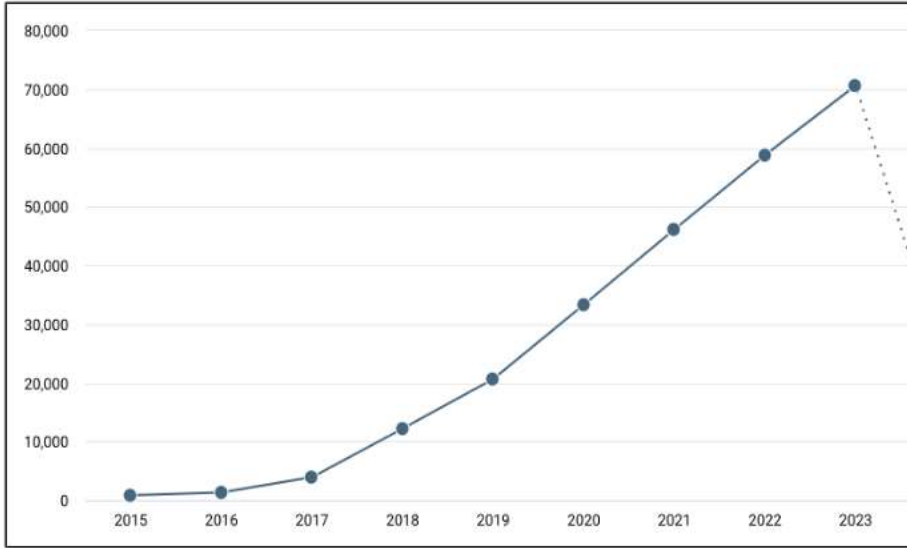


Şekil 3.3. CGAN modeline ait olan üretici-ayrıştırıcı toplam kayıp değeri (Elaraby, Barakat ve Rezk, 2022)

Üreticinin toplam kaybı iterasyon arttıkça düşmekte; ayrıştırıcının toplam kaybı artmaktadır. Kayıp fonksiyonu ve birleşik kayıp Çizelge 2.1’de gösterilmiştir. Bu değerlere göre geri yayılım algoritması (backpropagation) ile her tekrarda (epoch) ağlar, ağırlık (weight) ve kayıp (loss) değerlerini günceller ve birbiriyle yarışarak gelişir. 500 epoch boyunca eğitilen ağda üreticinin oluşturduğu sentetik görselleri tespit etmede ayrıştırıcının hata oranını sürekli arttığı; üreticinin ise kayıp oranını azalttığı görülmektedir (Elaraby, Barakat ve Rezk, 2022). Bunun sonucunda üretici sıfır kayıplı min-maks oyunun kazananı olacaktır. İyi eğitilmiş GAN modellerinde oluşturulan deepfake’lerin kalitesi diğerlerine göre daha iyi seviyede yer alır. GAN mimarisinde ayrıştırıcının üretici ile birlikte öğrenim aşamasında senkron olması diğer üretici sinir ağı modellerine göre daha iyi bir noktaya taşımaktadır.

Goodfellow ve arkadaşları bunu çalışmalarında “*Synchronizing the discriminator with the generator*” şeklinde ifade ederek önemli bir kazanımı dile getirmişlerdir (Goodfellow ve diğerleri, 2014).

Üreteci sinir ağları alanındaki gelişmeler ışığında yıllar içerisinde çeşitli GAN modelleri ve uygulamaları ortaya çıkmıştır. Ortaya çıktığından beri oldukça geniş bir çerçevede incelenen GAN, Dimensions.ai web sitesinden elde edilen istatistiklere göre çıktığı günden bu yana üzerinde 279 855 çalışmada geçtiği ve üzerinde çalışıldığı görülmektedir. Çalışma sayısının yıllar içindeki artışı Şekil 3.4’te gösterilmiştir.



Şekil 3.4. Çekişmeli üretici ağ çalışmalarının yıllar içindeki sayısı

Temelleri 2014 yılında atılan GAN’ın 2016 yılındaki görselden-görsel (image-to-image) üretimi uygulamaları ile çizimden gerçekçi görseller (sketch-to-image) üretimi, gündüz fotoğrafından geceye dönüşüm gerçekleştirilebilmesi mümkün olmuştur. GAN modelindeki gelişmeler ve hızlı öğrenim süreci sayesinde ilerleyen uygulamalarda StackGAN gibi öncü yazıdan-görsel (text-to-image) üretim uygulamaları yine 2016 içerisinde ortaya çıkmıştır. Gelişmiş uygulamaları ile öğrenim aşamasını daha iyileştiren ve kayıpları daha aza indiren DCGAN, WGAN, BiGGAN, StyleGAN ve sürümleri gibi modellere karşılık CT-GAN, Wasserstein GAN gibi modeller ile üç boyutlu (3D) görsel verisi üzerinde çalışmanın mümkün olduğu gösterilmiştir. Günümüzde ise GAN modelleri video işleme de yapabilmekte video kalitesi artırmada, geliştirmede ve eski kayıtları iyileştirmede kullanılabilmektedir. Yüksek dinamizm içeren ve yüksek kare / saniye oranına sahip

gerçekçi videoları sıfırdan oluşturan örnekler geliştirilmiştir (Reesink, 2020; Kench ve Cooper, 2021; Brooks ve diğerleri, 2022; Dundar, Gao, Tao ve Catanzaro, 2023).

### 3.2.1. GAN kalıntıları

GAN tarafından üretilen sahte görseller GAN bünyesindeki üreteç sinir ağı ile çeşitli manipülasyon teknikleri uygulanarak üretilmektedir. Üreteç bu noktada girdi olarak  $z$  (noise) gürültü girdisi almaktadır. Bu girdi kaynaklı benzetim oranı ağ çalışması sürecinde sürekli olarak değişmektedir. Sıfır ve bir arasında bir değer olan bu gürültü girdisi, birbirine yakın veya aynı olduğu durumlarda benzer sentetik görseller üretilmesine neden olmaktadır. Gürültü girdisine ek olarak çoğu GAN modelinde yer alan üst örnekleme yöntemi de görseller üzerinde kalıntılar ya da benzerlikler bırakabilmektedir. Özellikle üretilen görsellerdeki aynı bölgelerde aynı değişikliklerin uygulanması tekrarlama/genelleme analizi yapılarak GAN tespiti yapılabilmesine olanak sağlar. GAN kalıntıları, modelin sürekli olarak aynı ya da benzer görsel üretmesi nedeniyle isimlendirilen modelin çökmesi (mode collapse) ya da üretici kayıp değerinin sıfıra yakınsaması - istenilen çıktı yerine anlamsız çıktılar üretmesi ile isimlendirilen yakınsama arızası (convergence failure) nedeniyle kaynaklanmamaktadır. Sentetik üretimi başarılı olarak nitelendirilebilecek görseller üzerinde çeşitli izler, görsel hatalar veya dijital olarak incelendiğinde anlaşılabilecek aynı GAN tarafından üretildiğinde aynı düzendeki (pattern) değerler olarak nitelendirilebilir.

GAN kalıntıları genellikle frekans tabanlı tespit yöntemleri altında çalışmakta olup; GAN genellemesi ve belirgin-gözlemlenebilir kalıntılar olarak da incelenmektedir. GAN genellemesi kabiliyeti olarak yoğunlukla video deepfake'lerinde yüz ve kafa hareketlerinin sürekli olarak belirgin bir düzende ve düzlemde olması, konuşma şeklinin benzer olması gibi çeşitli karakteristik özelliklerdir. Spesifik GAN modeline has bu durum için farklı modeller tarafından oluşturulan imajlarda tespit mümkün olmayabilir. Aynı GAN modelinde ise ağı eğitimi durumuna göre bu genelleme söz konusu olmayabilir. Bu durum GAN'ın öğrenim aşamasındaki veri seti büyüklüğüne göre değişebilmektedir.

Görülebilir GAN kalıntıları insan gözü performansı ile tespit edilebilir seviyede olmaktadır. Fakat, görünmez kalıntılar için çeşitli adli bilişim yöntemleri gerekmektedir. Bu konudaki incelemeler ve çalışmalar ikinci bölümde detayları ile verilmiştir. Bu alandaki çalışmalar; insan görüşü performansı üzerinden anket yapılması, GAN parmak izi analizi, uzamsal

kalıntılar, anomali tespiti, frekans uzayında inceleme, eş-oluş matrisi çıkararak analiz gibi yöntemler kullanılmaktadır.



## 4. METOT VE ANALİZ

Bu bölümde önerilen yöntem ve genel akış için gerekli veri setleri, araçlar ve metotlar hakkında bilgiler verilmektedir. Literatürde yer alan GAN kalıntısı tespiti için kullanılan tanımlamalar, araçlar, yöntemler ikinci bölümde incelenmiştir. Daha sonra ortak ve öne çıkan özellikleri tespit edilen bu araçların yeni ve güncel GAN kalıntısı üzerinden deepfake tespiti yöntemleri ile birlikte geliştirilmesi yapılarak sonuçlar analiz için kaydedilmiştir.

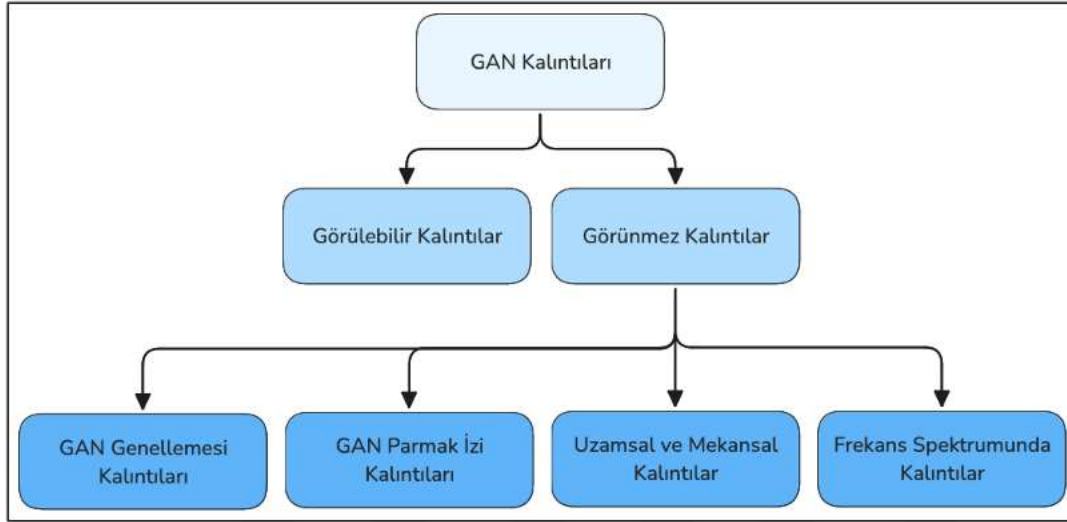
Öncelikle görece literatürde az işlenen bir konu olması nedeniyle literatüre katkı sağlayacağını düşündüğümüz GAN kalıntısı için bir taksonomi çalışması yapılmıştır. Taksonomiye göre literatürdeki çalışmalar için bir sınıflandırma sağlanmıştır. Daha sonra, GAN kalıntılarını imajların renk korelasyonu üzerinden çıkarım yapmak üzere frekans spektrum analizi ile inceleyerek tespit eden simülasyon ile belirlenmiş GAN modelleri tarafından üretilmiş deepfake'ler simülasyon ile test edilmiştir. Simülasyonda kullanılmak üzere üç farklı GAN modeli belirlenmiş ve bu GAN modelleri kullanılarak oluşturulan özgün veri seti oluşturulması çalışmanın ve bu alanda eksik olduğunu düşündüğümüz çoklu GAN modeli olma özelliğini taşımaktadır. Simülasyon özellikleri literatürdeki çeşitli yöntemlerden faydalanılarak seçilmiş, GAN modelleri ise yüz görsellerine odaklanan ve deepfake oluşturulması alanındaki öncü modellerden özenle seçilmiştir.

### 4.1. GAN Kalıntıları Taksonomisinin Çıkarılması

GAN Kalıntısı üzerinden deepfake tespiti ya da GAN modeli tanımlaması ile ilgili ikinci bölümde yapılan literatür taramasında sık çalışılan bir alan olmadığı tespit edilmiştir. Bunun üzerinde yapılan çalışmaları sınıflandırmak için bir taksonomi çıkarılması ihtiyacı doğmuştur.

GAN modellerinin iç yapısı, eğitim aşamasındaki veri çeşitliliği, veri seti büyüklüğü, veri seti kalitesi, veri setindeki görsellerin kalitesi, görsel çözünürlüğü gibi birçok parametreden kaynaklanabilen GAN kalıntıları genel olarak insan gözüyle ayırt edilebilir ve ayırt edilemez olarak iki kategoride sınıflandırılabilir. İnsan gözü tarafından ayırt edilemez (invisible) fakat çeşitli yöntemlerle dijital analizi gerçekleştirildiği takdirde tespiti mümkün olan kalıntılar için çalışmalarda farklı sınıflandırma yöntemleri uygulanmıştır. Bu sınıflandırmayı

yaptığımız taramalar ışığında daha genel bir çerçeveye oturarak Şekil 4.1’de görülen taksonomiye elde ettik. Buna göre, GAN izlerinden analiz, GAN genelleme ve frekans uzayı analizi ile gerçekleştirilen çalışmaları insan gözüyle gözlemlenemeyen kalıntıların altında konumlandırılmıştır. İncelenen literatür taraması sonucunda elde edilen ve çalışmamız kapsamında öne sürülen taksonomi Şekil 4.1’de gösterilmektedir.



Şekil 4.1. GAN kalıntıları taksonomisi

Oluşturulan taksonomide yer alan her sınıflandırma için örnekler sunulmuştur. Buna göre tez, esas olarak görünmez kalıntılara odaklanmaktadır. Çalışmanın ilerleyen kısımlarında gözlemlenebilir kalıntılar üzerinde herhangi bir çalışma yürütmemiştir. Literatürde bu anlamda çeşitli anket çalışmaları ve inceleme çalışmaları görülmektedir (Khodabakhsh Ramachandra ve Busch, 2019). Görünmez kalıntılar için geliştirilen simülasyonun görülebilir çalışmaları da derin öğrenme yöntemleri ve dijital inceleme açısından kapsamı beklenmektedir. Matern, Riess ve Stamminger (2019) çalışması gözle görülebilir GAN kalıntılarını; gözlerde şekilsel bozukluk, diş çizimindeki hatalar ve yüz renk geçişlerinde bozukluk gibi sıralamıştır.

GAN tarafından üretilmiş deepfake’lerdeki görülebilir kalıntılara örnek olarak; PGGAN, StarGAN, StyleGAN ve SC-FEGAN tarafından üretilmiş imajlar Resim 4.1’de gösterilmektedir. İnsan gözü tarafından algılanabilecek seviyede kalıntılar içeren bu deepfake’ler açıkça belli olması nedeniyle sifra yakın seviyede risk oluşturduğu söylenebilir.



Resim 4.1. PGGAN, StarGAN, StyleGAN ve SC-FEGAN tarafından üretilmiş deepfake görseller (Juefei-Xu ve diğerleri, 2022)

Gözle görülebilir kalıntılara sahip deepfake'ler oluşturan bu modellerin gözle ayırt edilemeyecek örnekleri ikinci bölümde gösterilmiştir. İlgili GAN modellerinin bu örneklerdeki kalıntıları öğrenim aşamasındaki veri setinin düşük kaliteli görseller içermesinden kaynaklanmaktadır (Juefei-Xu ve diğerleri, 2022).

İnsan gözü tarafından ayırt edilmesi mümkün olmayan görünmez GAN kalıntıları dört kategoriye ayrılmıştır. Bu sınıflandırma bir sonraki bölümde detaylarıyla analiz edilmiş ve örneklendirilmiştir.

#### 4.1.1. Görünmez GAN kalıntıları

İnsan gözüyle ayırt edilmesi mümkün olmayan fakat dijital yöntemler ve analizler sonucunda ortaya çıkarılarak sahte-sentetik olduğu anlaşılabilir GAN tarafından üretilmiş imajlar dört sınıfa ayrılmıştır. Bunlar;

- GAN genellemesi sonucu oluşan kalıntılar,
- GAN parmak izi kalıntıları,
- Uzamsal-mekansal kalıntılar,
- Frekans uzayı kalıntıları olarak sınıflandırılmıştır.

GAN genellemesi (GAN generalization) ile oluşan kalıntılar; öğrenim ve test hatalarının birbirine yakınsaması nedeniyle oluşan GAN tarafından dağıtımın (distribution) öğrenemediği durumlarda gerçekleşen kalıntılar ile oluşmaktadır. Örnekleri Arora, Ge, Liang, Ma ve Zhang (2017) çalışmasında sunulmuştur; *“One issue raised by our analysis is that the current GANs objectives cannot even enforce that the synthetic distribution has high diversity. Furthermore the issue cannot be fixed by simply providing the discriminator with more training examples. Possibly some other change to the GANs setup are needed.”* Şeklinde ifade ederek, yaptıkları analizler sonucunda mevcut GAN’ın iç yapısının sentetik üretimde dağıtımın çok çeşitli olması gibi bir görev üstlenmediğini belirtmektedir. Bu durumda ayrıştırıcıya daha fazla öğrenim verisi sunmak ile çözülemeyeceğini; GAN yapısında değişiklik gerekliliğini vurgulamıştır. Yang ve Weinan (2022) Çalışması ise bu durumu ayrıştırıcının çeşitlilik eksikliğini tespit edememesinin aksine ayrıştırıcının öğrenim sürecini iyi genellemesinin bir avantaj olarak nitelendirilebileceği belirtilmiştir.

GAN parmak izi kalıntıları, (GAN fingerprints) GAN tarafından üretilen sentetik imajlarda yer alan ve model bazlı veya imaj bazlı tespit edilebilecek seviyede olan kalıntılardır. Her GAN modelinin kendine ait çeşitli karakteristik özellikleri mevcuttur. Bunlar, eğitim seti verisi dağılımı, ağ yapısı, hiper parametreleri, optimizasyon fonksiyonları olabilir. Üretici ve ayrıştırıcı arasındaki başlangıç değerleri model ağırlıkları nedeniyle her eğitim ile aynı değerlere yakınsama yapar – benzeşim ortaya çıkar. Bu nedenle, iyi eğitilen GAN modelleri dahi kendi içerisinde benzer performans sunar. Yüksek kaliteli görsel üretimini hedefleyen GAN modelleri benzerlik nedeniyle mode agnostik olmak üzere eşsiz parmak izi kalıntılarına neden olmaktadır. Yu, Davis ve Fritz (2019) çalışmasında bu durumu *“In a specifically designed case, an image fingerprint can be an RGB image the same size as the original image.”* Olarak ifade eder ve bir görüntü parmak izi, orijinal görüntüyle aynı boyutta bir RGB görüntüsü olabileceği yorumu ile parmak izi kalıntısını doğruladığını ifade eder. Marra, Gragnaniello, Verdoliva ve Poggi 2019 çalışmasında ise mimari ve eğitim seti farketmeksizin birçok GAN’ın oluşturduğu sentetik imajlarda parmak izi olarak adlandırılacak ayırt edici işaretler bıraktığını tespit etmiştir. Ding, Thakur ve Li (2021) çalışması ise GAN parmak izinin GAN modeline göre eşsiz olarak tespit edilebildiğini ve elde edilen parmak izinin eğitim süresince kullanılan veri setine bağlı olduğunu da belirtmektedir. Uzamsal-mekansal (spatial domain features) kalıntılar GAN modelleri tarafından üretilen sentetik görsellerdeki modelin çeşitli kısıtlara takılması nedeniyle limitlendiği noktalardan kaynaklanmaktadır. GAN modellerinin eğitim verisinde olmasına

rağmen satüre görseller üretmediği veya geleneksel fotoğraf makineleri gibi kısa pozlama özelliklerini sentetik imajlara yansıtmadığı durumlar söz konusu olabilmektedir. McCloskey ve Albright (2019) Çalışmasında kameraların genellikle sature olduğu ve kısa-uzun pozlamanın belli olduğu bölgelere sahip olduğu belirtilirken GAN sentetik imajlarında yer almadığı tespit edilmiştir. Nataraj ve diğerleri (2019) tarafından yapılan çalışmada RGB renk piksellerinin eş-oluş matrisi üzerinden yapılan analiz üzerine tespit gerçekleştirerek GAN tespiti gerçekleştirmiştir. Li, H., Li, B., Tan ve Huang (2020) çalışmasında krominans renk bileşenlerinden elde edilen değerler ışığında sentetik imajları tespit etmeyi amaçlamıştır. Buna ek olarak, GAN modellerinin genel olarak RGB renk uzayında sentezleme yaptığını; RGB, HSV, and YCbCr renk uzaylarında incelendiğinde elde edilen kalıntılardan istatistiksel yöntemler kullanarak tespitin mümkün olduğunu savunmuştur.

Frekans uzayı (frequency spectrum features) kalıntıları sentetik imajların frekans uzayındaki dönüşümü yapılarak incelenmesi ile edilen çıktılar olarak göze çarpmaktadır. Çeşitli çalışmalar sentetik imajları gerçeklerinden ayırt etmek için bu yöntemi kullanmış ve farklı dönüşüm yöntemlerini öne sürmüştür. Bu çalışmaların öncülerinden olan Frank ve diğerleri tarafından (2020) yapılan ve çalışmasında Lewis ve diğerleri (2020) tarafından yapılan çalışmalarda DCT (Discrete Cosine Transform) dönüşümü ile elde edilen çıktılar sayesinde; gerçek ve sentetik imajları ayırt edebildikleri ve hangi GAN modeli tarafından üretildiğini algılayabildiğini belirtilmiştir. Le ve Woo (2021) çalışmasında ise Fourier dönüşümü kullanılarak deepfake tespiti hedeflenmiştir. Li ve diğerleri (2020) çalışmasında ise frekans uzayında filigran ekleme yöntemi kullanılmış ve tekrar incelendiğinde tespit edilmesine güvenilen bu yöntem ile fikri mülkiyet haklarında koruma katmanı geliştirilmesi amaçlanmıştır.

#### **4.2. Tanınmamış ve Ön Çalışmasız GAN Tespiti**

Belirli GAN modeli tarafından üretilen deepfake imajları (ya da veri setlerini) yapay sinir ağına tanıtarak öğrenme yoluyla tespit gerçekleştirmek literatürde en sık çalışılan konulardandır. Fakat, daha önce derin öğrenme ağı tarafından hiç görülmemiş bir GAN modeli tarafından üretilen ve özellikle yüksek çözünürlüklü ve kaliteli deepfake'lerin tespiti oldukça zordur. Çalışmada farklı GAN modelleri arasındaki yapıların tespit yöntemleri tarafından algılanmasının başarımı test edilmiştir. Buna göre amaçlanan belirli bir GAN modeli ile oluşturulmuş veri setinin sinir ağına sunulması ve sonrasında başka bir GAN

modeli tarafından oluşturulmuş imajları ayırt etmesini beklemektir. Literatürde de bu alanda çeşitli kaygıların belirtildiği ikinci bölümde “unseen GAN”, “unknown GAN model” veya “new GAN architecture” tespit yöntemleri sunulmuştur. Bu yöntemler deepfake saldırıları için daha önemlidir çünkü deepfake imajının hangi GAN modeli ile üretildiğini tespit açısından bilinmez bir konudur.

Tanınmamış ve ön çalışmasız GAN modeli tespiti için simülasyona tanıtılan bir veya daha fazla GAN modelinden farklı bir GAN modeli ile üretilmiş sentetik imajları girdi olarak sağlayarak tespit etmesini beklemekteyiz. Bu sayede farklı GAN modelleri arasında ortak özellikler ve katmanlardan kaynaklanan kalıntıların tespitinin araştırılması hedeflenmektedir.

#### **4.3. GAN Kalıntısı Tespit Simülasyonu Geliştirilmesi ve Özellikleri**

GAN kalıntılarını inceleyerek imajların sahteliğinin tespiti ve sentezlendikleri GAN modelini anlamak için geliştirilmesi hedeflenen simülasyon için gereksinimler literatürdeki bilgiler ışığında toplanmıştır. Buna göre en iyi başarı elde eden çözümler temel alınarak gereksinim analizi oluşturulmuş ve simülasyon geliştirilmiştir. Çalışmamızda GAN kalıntıları üzerinden tespit gerçekleştirilmesi üzerine çalışılan altı farklı çalışmadan (Shrivastava ve diğerleri, 2016; Jenni ve Favaro, 2018; Marra, Gragnaniello, Verdoliva ve Poggi 2019; Yu, Davis ve Fritz, 2019; Zhang, Karaman ve Chang, 2019; Jeong, Kim, Kim, Ro ve Choi 2021; Amerini, Anagnostopoulos, Maiano ve Celsi, 2021) esinlenilmiştir.

Önerilen simülasyon dört aşamada çalışmaktadır; girdi katmanı, özellik çıkarımı, veri ölçekleme ve sınıflandırma ile sonuç üretimi sırasıyla çalışmaktadır. Frekans spektrumu altındaki doğal olmayan renk korelasyonu altında oluşan anomali - kalıntıların tespiti üzerine çalışan bu simülasyon sonucunda girdi olarak verilen üç farklı GAN modeli, üç farklı frekans spektrumunda dönüşüm yöntemi ve 12 farklı sınıflandırma algoritması kullanılması nedeniyle her iterasyonda dokuz farklı sonuç elde edilmektedir.

Simülasyonda, GAN modelleri olarak, ProGAN, InterfaceGAN ve StyleGAN3; dönüşüm metotları olarak Fourier, Wavelet ve HOG; sınıflandırma algoritmaları olarak ise Gaussian Mixture Model (GMM), Gradient Boosting, XGBoost, ADA Boost, Random Forest, Yapay Sinir Ağları, karar destek makinaları (SVM), k-en yakın komşu (KNNs), Logistic

Regression, Naive Bayes ve Extra Trees (Extremely Randomized Trees) algoritmaları uyarlanmış ve kullanılmıştır. Tüm sürecin tek bir çalışma süresinde ilerletildiği simülasyonda doğruluk, duyarlılık, geri çağırma ve fl değerleri hesaplanmıştır.

### Test ortamı

Simülasyon, Jupyter Notebooks ortamını bulut ortamına taşıyan Google Colab (Collaboratory) ortamında çalışacak şekilde tasarlanmış, çalışma kolaylığı ve hız açısından GPU kullanımı uygun bulunmuştur. Simülasyona sunulacak veri setlerinin GAN modelleri tarafından oluşturulması aşaması, simülasyonun eğitimi ve analiz aşamalarında Colab Pro+ hesabı kullanılmıştır. Colab, çeşitli derin öğrenme çalışmalarında kullanılmış ve GPU, TPU, Python çalışma ortamı (runtime), RAM ve disk sayesinde bu alandaki çalışmalara kolaylıklar sağlamıştır (Kanani ve Padole, 2019; Canesche, Bragança, Neto, Nacif ve Ferreira, 2021; Gunawan ve diğerleri, 2020; Jia, 2022). Çalışma ortamına sağladığı özelliklere ek olarak, bu alanda kullanılan veri setleri hazır bir şekilde Colab ortamında mevcuttur; bu veri setlerini kullanmak bir satırlık kod ile içeri alma işlemi uygulamak mümkündür. Python kodu ile geliştirilen simülasyon ile üç farklı açıdan inceleme yapılmıştır; spektrum analizi, RGB renk uzayı korelasyonu ve istatistiksel değerler üzerinden analiz şeklindedir.

### GAN modelleri

Simülasyona girdi olarak sağlanacak imajlar için çeşitli GAN modelleri belirlenmiş ve bu GAN modelleri tarafından üretilen imajlar sunulmuştur. Bunun için belirlenen GAN modelleri özellikle yüz görsellerine odaklanan StyleGAN3, InterfaceGAN ve ProGAN olarak belirlenmiştir. Birden fazla GAN modeli ile görsel üretilerek ile çoklu GAN modeli üzerinden deepfake tespiti başarımları karşılaştırmasına olanak sağlanmıştır.

### Özellik çıkarımı ve renk farkı hesaplaması

Çalışmamızın deepfake tespiti kısmını oluşturan ilk aşamada imajlar üzerinden veri elde etmek için özellik çıkarımı işlemi uygulanmıştır. Bu aşamada elde edilen veriler daha sonra ölçeklendirilerek önemi ve ağırlığına göre sınıflandırma algoritmasına sunulmak kaydıyla işlenmektedir. Çalışmamızda özellik çıkarımı için HOG algoritması tercih edilmiştir. Yüz

tanımlama uygulamalarında da kullanılan HOG görüntüyü küçük bölgelere ayırarak gradyan hesaplaması uygulayarak kenar bölgelerin histogramlarını çıkartarak nesne tespiti ve algılama üzerine çalışmaktadır. HOG, kontrast ve aydınlatma değişikliklerine karşı dirençli olup yüksek doğruluk için bölgesel yoğunluk tespiti üzerinden ilerlemesi sayesinde etkili bir yöntemdir.

Özellik çıkarımı yaparken sunulan görsellerin renk farkı hesaplaması göz önünde bulundurulur. Deepfake imajlarındaki özellikle yumuşatma veya harmanlama gibi tekniklerin uygulanması nedeniyle doğal renk kanalları arasındaki ilişkide hatalar yapabilmekte ve farklara neden olmaktadır. Bu farkları bir görüntünün kırmızı (red), yeşil (green) ve mavi (blue) kanallarında HOG dönüşümü uygulanması ile frekans spektrumunda analiz ederek tanımlayabilmek mümkün hale gelmektedir. HOG dönüşüm yöntemi görsellerdeki nesnelerin kenar bölgelerini belirgin kısımlarını algılayarak özellik çıkarımı gerçekleştirmektedir. Bu yöntem ilk olarak yüz algılamada kullanıldığı için deepfake tespitinde de kullanılması uygun görülmüştür. Özellik çıkarımı ve renk farkı hesaplaması sonrası elde edilen verilerin incelenebilmesi için ölçeklendirme yapılmaktadır. Bunun için standardizasyon ve normalizasyon işlemi uygulanır. Çalışmamızda, renk kanalları arasındaki farkları tespit edebilmek için kullanılan değerler;

- Ortalama fark
- Minimum değer
- Maksimum değer
- Maksimum ve minimum arasındaki fark
- Korelasyonsuz davranış için;  $1 - \text{korelasyon}(K, G)$  değeri elde edilmiştir.

#### Baz veri seti

Belirlenen GAN modelleri için doğrudan ulaşılabilen veri setleri bulunmadığından Colab ortamında doğrudan erişim sağlanabilen CelebA veri seti kullanılmıştır. Farklı yaşlardan, farklı etnik gruplardan ve farklı arka planlar içeren yüz fotoğraflarında gözlük, güneş gözlüğü ve şapka gibi çeşitli aksesuarların çeşitliliği de mevcut olduğundan kullanılması uygun bulunmuştur. Bu veri seti simülasyona tespit aşamasında ve karşılaştırma sürecinde gerçek olduğunu belirtmek için simülasyona sunulmuştur.

### Sınıflandırma algoritmaları

Özellik çıkarımı ve HOG frekans uzayındaki dönüşüm ile edilen veriler ölçeklendirme ile daha açıklanabilir ve sınıflandırmaya hazır hale gelmektedir. Bu aşamada sınıflandırma anlamında ikilik sistem (binary classification) kullanılmaktadır. İkilik sistemde elde edilen sonuç true-false, gerçek-sahte veya doğru-sentetik çıktısını sağlamaktadır. Çalışmamızda sınıflandırma algoritması olarak farklı açılardan başarımları test edebileceğimiz öngörüsü ile Gaussian Mixture Model (GMM), Gradient Boosting, XGBoost, ADA Boost, Random Forest, Yapay Sinir Ağları (YSA), Support Vector Machine (SVM) ve RBF çekirdekli bir modeli (SVM RBF), K-Nearest Neighbors (kNN), Logistic Regression, Naive Bayes ve Extra Trees (Extremely Randomized Trees) algoritmaları tercih edilmiştir.

#### **4.4. Çoklu GAN Veri Seti**

Çalışmamızda kullanmak üzere üç farklı GAN modelinden toplamda 5000 yüz görseli içeren sentetik imajlar üretilmiştir. Çalışmamızda CelebA veri seti baz alınmıştır ve yüz görselleri için önceden eğitilmiş (pre-trained) model kullanılmıştır. Kullanılan modeller, baz alınan veri seti ve üretilen veri setine dair görsel sayısı bilgileri Çizelge 4.1’de gösterilmektedir. Doğru sonuçlar elde edilmesi kıyaslamamanın geçerli olması adına üretilen ve mevcut tüm görseller için JPEG formatı ve 256x256 ölçüsü kullanılmıştır.

Çizelge 4.1. Baz veri seti ve GAN modeli tarafından üretilen veri setleri bilgileri

| Gerçeklik | Model        | Adet  | Kaynak               |
|-----------|--------------|-------|----------------------|
| Gerçek    | CelebA       | 5 000 | CelebA veri seti     |
| Sahte     | StyleGAN3    | 5 000 | Tarafımızca üretildi |
| Sahte     | InterfaceGAN | 5 000 | Tarafımızca üretildi |
| Sahte     | ProGAN       | 5 000 | Tarafımızca üretildi |

Çalışmada gerçek-sahte tespitinde kullanılmak üzere oluşturulan özgün veri seti içerisinde toplam 20 bin adet yüz imajı yer almaktadır. Her bir GAN modeli ile üretilen sahte veri seti 5 bin adet olup; karşılaştırılan CelebA veri seti de 5 bin gerçek yüz görsellerinden oluşmaktadır. Üretim aşamasında Colab Pro+ kullanılmıştır ve veri seti büyüklüğü nedeniyle

51 GB kullanımına olanak sağlayan yüksek hafıza (High RAM) kullanılarak üretim gerçekleştirilmiştir. Üretilen görseller kullanılmak üzere Drive üzerine kaydedilmiştir. Üretim sırasında önceden eğitilmiş model kullanılmasına karşı üretim sürecinin farklı sürelerde üretildiği görülmüştür. Üretim sürecine dair bilgiler Çizelge 4.2’de yer almaktadır.

Çizelge 4.2. GAN modeli ve veri seti oluşturma süreleri bilgisi

| GAN modeli   | Adet  | Toplam süre | Ortalama süre       |
|--------------|-------|-------------|---------------------|
| StyleGAN3    | 5 000 | 15 dakika   | 5,56 görsel/saniye  |
| InterfaceGAN | 5 000 | 35 dakika   | 2,38 görsel/saniye  |
| ProGAN       | 5 000 | 6 dakika    | 13,89 görsel/saniye |

Belirlenen GAN modeli tarafından oluşturulan veri setlerinin toplam ve ortalama süreleri Çizelge 4.2’de gösterilmektedir. Buna göre en hızlı oluşturulan veri set modeli ProGAN ve en yavaş oluşturulan veri seti modeli InterfaceGAN olmuştur.

Gerçek görsellerle karşılaştırma yapmak için yine CelebA içerisinden rastgele 5 bin yüz görseli seçilmiştir. Rastgele seçim için Linux komutlarında “shuf” kullanılmıştır ve komutta belirtildiği (`find /celeba -type f | shuf -n 5000 | xargs -I{} mv {} /random_picked_celeba`) gibi bir seçim yapılması sağlanmıştır. Üretilen görseller kendi içinde ayrı dizinlerde konumlandırılmış tespit testleri için gerekli sınıflandırma yapısı oluşturulmuştur.

#### 4.5. Çoklu GAN Simülasyonu ile Kalıntı Algılama Üzerinden Deepfake Tespiti

Çalışma kapsamında üretilen veri setleri ve gerçek yüzlerden oluşan veri seti arasında inceleme yapılarak karşılaştırma yapılmıştır. Buna göre gerçek veri setinde görülen kalıntıların deepfake görsellerde yer alıp almadığı, sıklığı ve frekans uzayındaki görüntüsü incelenmiştir. Bu alanda yapılan çalışmalar ikinci bölümde detayları ile paylaşılmıştır.

Öncelikli olarak üretilen veri setlerinde uzamsal-mekansal kalıntılar incelenmiştir. Veri seti kalitesinin anlaşılabilmesi için bu anlamda gözleme dayalı bir inceleme uygun bulunmuştur. Çalışmada kullanılan gerçek yüzler içeren veri seti CelebA’dan örnek görseller Resim 4.2’de gösterilmektedir.



Resim 4.2. CelebA gerçek yüzler veri seti örnek imajlar

ProGAN ile üretilen sentetik görsellere dair örnekler Resim 4.3'te; InterfaceGAN ile üretilen sentetik görsellere dair örnekler Resim 4.4'te ve StyleGAN3 ile üretilmiş sentetik görsellere dair örnekler Resim 4.5'te gösterilmiştir.



Resim 4.3. ProGAN tarafından üretilmiş deepfake imajlar



Resim 4.4. InterfaceGAN tarafından üretilmiş deepfake imajlar

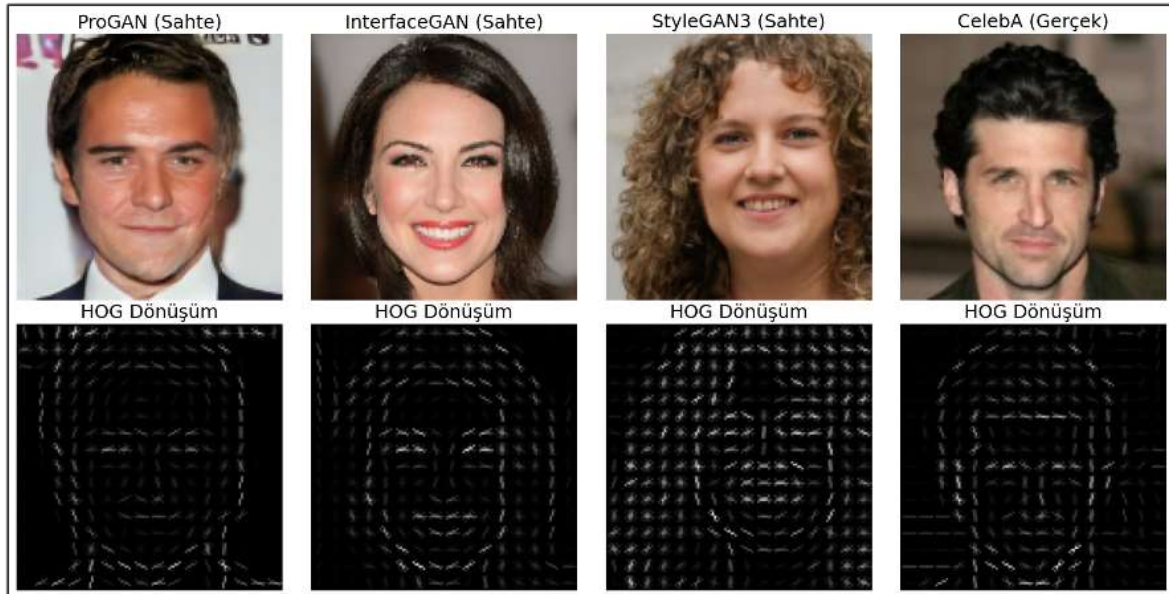


Resim 4.5. StyleGAN3 tarafından üretilmiş deepfake imajlar

GAN kalıntıları taksonomisine göre oluşturduğumuz sınıflandırma açısından uzamsal-mekansal açıdan yapılan analizlere göre StyleGAN3 en iyi deepfake üretici olduğu açıkça

görülmektedir. Üretilen veri setindeki çoğu imajın kalıntısız bir şekilde üretildiği açıkça tespit edilmiştir. Diğer yandan ProGAN ve InterfaceGAN ile üretilen görsellerde çeşitli sorunların yer aldığı söylenebilir. Resim 4.3’de ProGAN tarafından üretilen görsellerde yakınlaştırma yapıldığında özellikle saç bölgesinde çeşitli çizim hataları olduğu tespit edilebilmektedir. Resim 4.3’de yer alan ikinci yüz görselinde ise gözlük çerçevesinde çeşitli kalıntılar seçilebilmektedir. InterfaceGAN ile oluşturulan veri seti de oldukça gerçekçi deepfake oluşturabilmektedir. Resim 4.3’e yer alan görseller oldukça gerçekçi durmaktadır fakat; birinci görseldeki göz boyutundaki ölçü farklılığı, üçüncü görseldeki göz rengi anomalisi ve dördüncü görseldeki saç kısmında yer alan çizim sorunları kalıntı tespiti anlamında tespit edilebilmektedir.

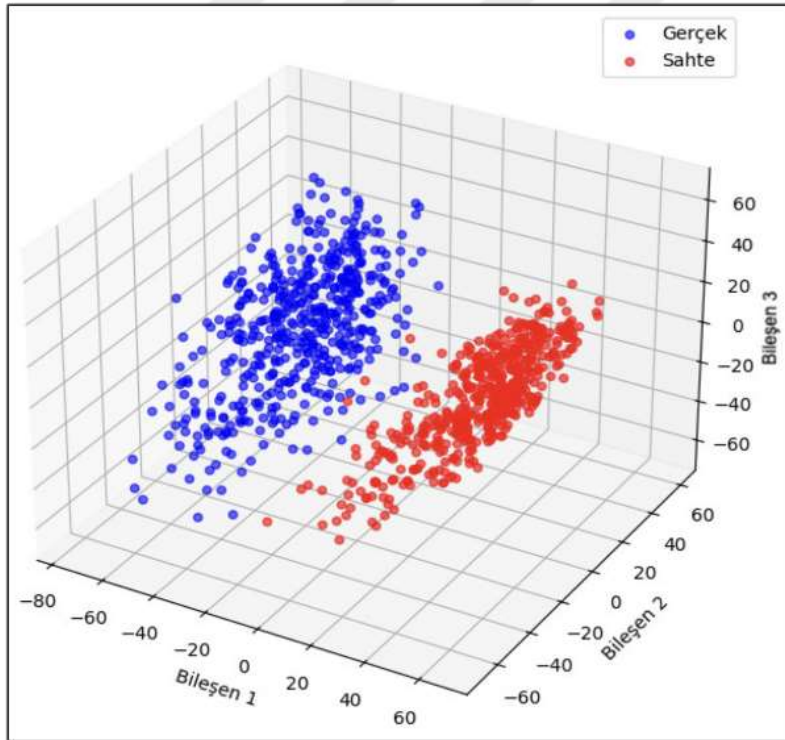
Frekans uzayında inceleme üzerine çalışan simülasyonda her sentetik veri seti gerçek veri seti ile birlikte kıyaslanmıştır. Buna göre frekans uzayındaki hesaplamalar karşısında alınan değerler ve görselleştirmeleri incelenmiştir. Resim 4.6’da sırasıyla ProGAN, InterfaceGAN, StyleGAN3 ve CelebA veri setlerinden alınan görsellerin HOG dönüşümü kullanılarak yapılan spektrum analiz karşılaştırılması örnek olarak sunulmuştur. Yüz hatları kenar noktalarına odaklanan bu dönüşümü modeli için yüz algılama ve ardından hatların keskinliği üzerinden yapılan hesaplamalar üzerine karşılaştırma yaparak simülasyonun tespit aşamasına ilerlemesi sağlanmıştır.



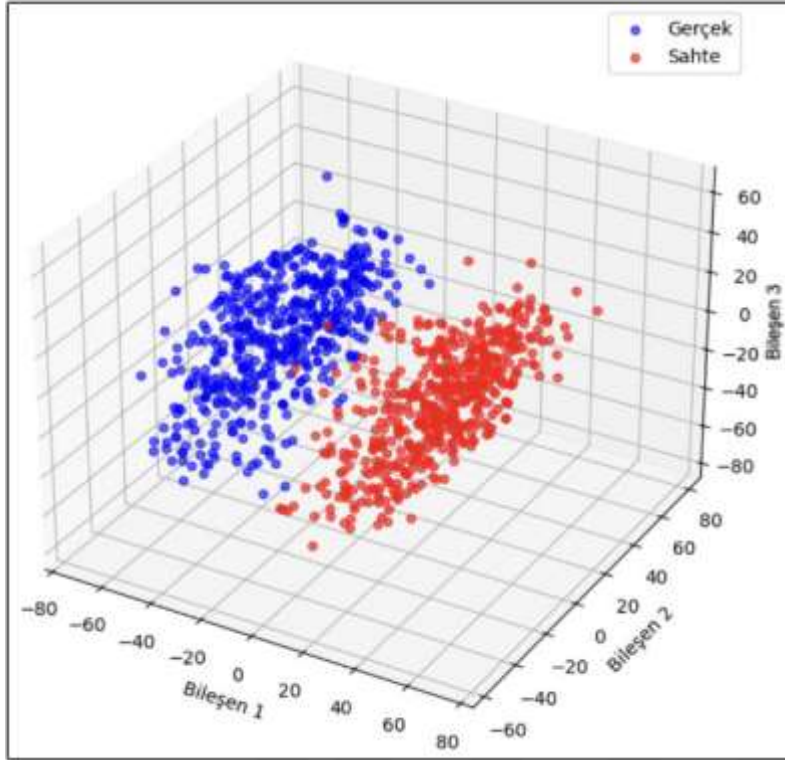
Resim 4.6. Tüm veri setlerine HOG dönüşümü uygulanması sonucu görünümü

Simülasyona sunulan sentetik ve gerçek görsellerin özellik çıkarımı aşamasından sonra 3B uzayda incelenerek anlamlı hale getirilmesi sağlanmıştır. Bunun için temel bileşen analizi yöntemi kullanılarak ilgili 3B uzayda elde edilen veriler görselleştirilmiştir. Aynı veriler sınıflandırma algoritmaları tarafından yorumlanarak ikilik sistemde gerçek ve sahte olarak ayırt edilmektedir.

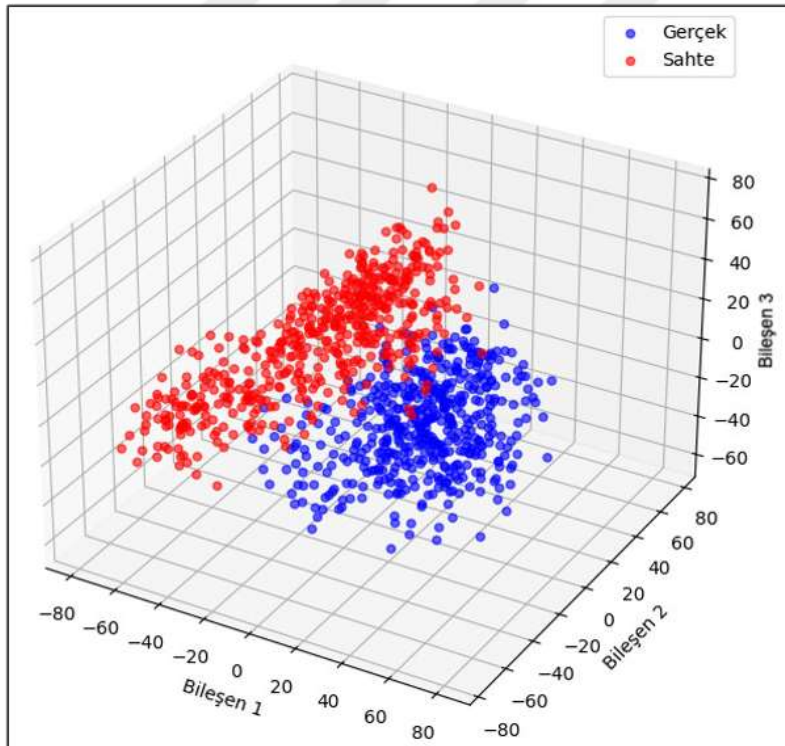
Kullandığımız veri setlerinden InterfaceGAN, ProGAN ve StyleGAN3 görselleri ile CelebA görsellerinin simülasyona sunulması ardından HOG dönüşümü altında yapılan analizin üç boyutlu görselleştirilmesi sırasıyla Şekil 4.2, Şekil 4.3 ve Şekil 4.4'te gösterilmektedir. InterfaceGAN, ProGAN ve StyleGAN3 görselleri ile CelebA görsellerinin simülasyona sunulması ardından Wavelet dönüşümü altında yapılan analizin üç boyutlu görselleştirilmesi sırasıyla Şekil 4.5, Şekil 4.6 ve Şekil 4.7'te sunulmaktadır. Kullandığımız veri setlerinden InterfaceGAN, ProGAN ve StyleGAN3 görselleri ile CelebA görsellerinin simülasyona sunulması ardından Fourier dönüşümü altında yapılan analizin üç boyutlu görselleştirilmesi sırasıyla Şekil 4.8, Şekil 4.9 ve Şekil 4.10'de gösterilmektedir.



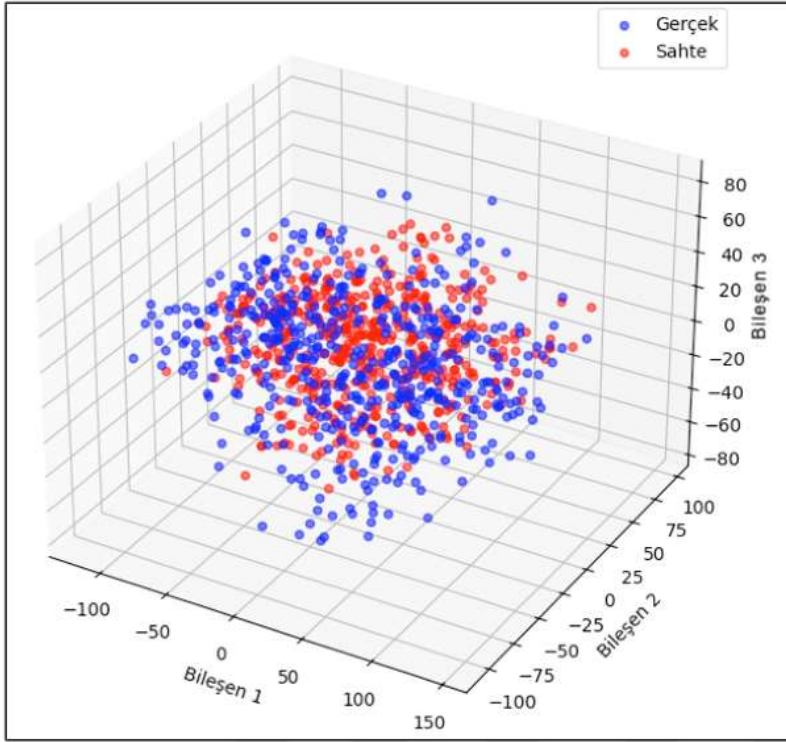
Şekil 4.2. CelebA ve InterfaceGAN görsellerinin HOG dönüşümü altındaki 3B veri dağılımı



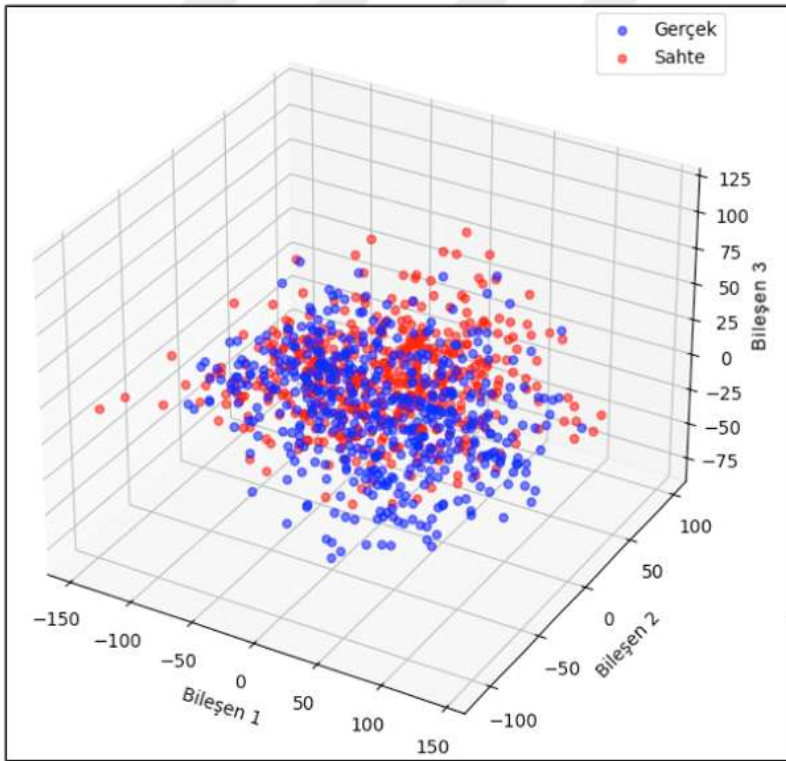
Şekil 4.3. CelebA ve ProGAN görsellerinin HOG dönüşümü altındaki 3B veri dağılımı



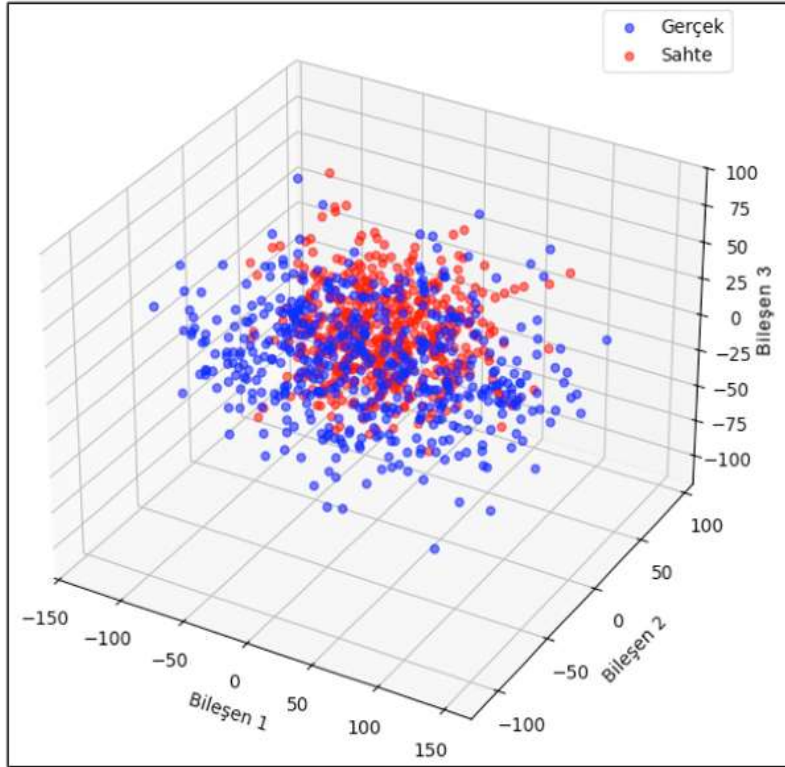
Şekil 4.4. CelebA ve StyleGAN3 görsellerinin HOG dönüşümü altındaki 3B veri dağılımı



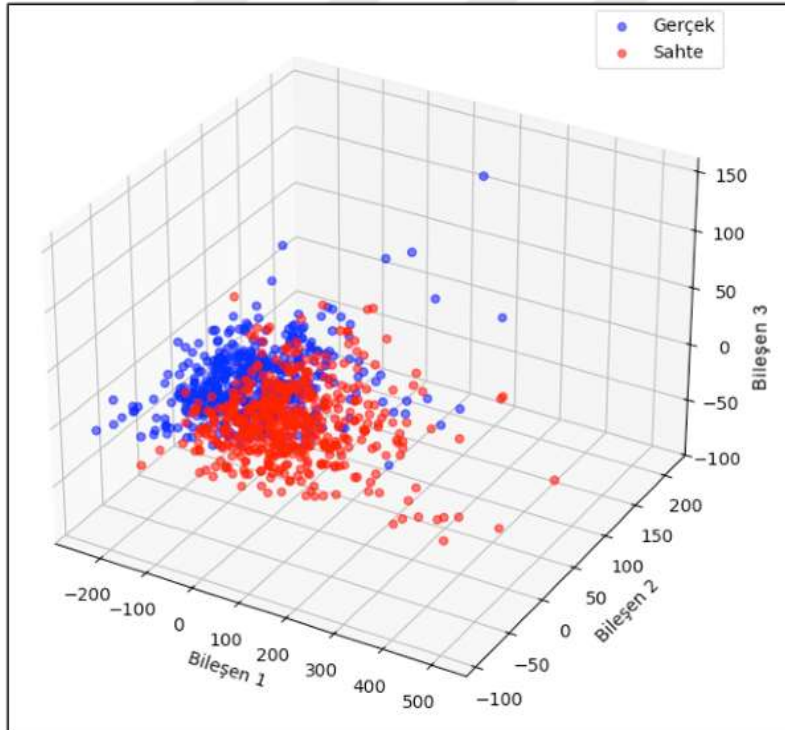
Şekil 4.5. CelebA ve InterfaceGAN görsellerinin Wavelet dönüşümü altındaki 3B veri dağılımı



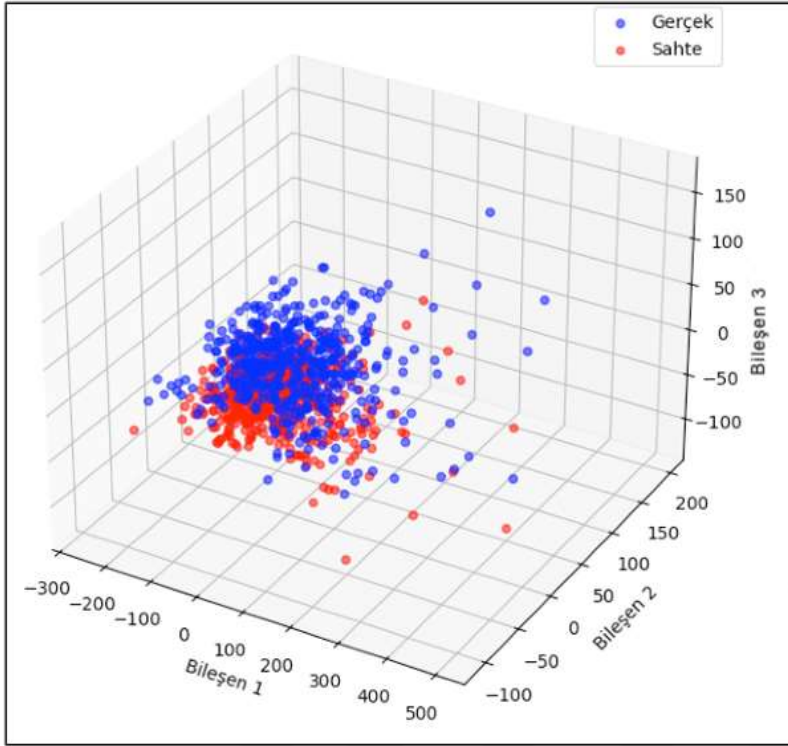
Şekil 4.6. CelebA ve ProGAN görsellerinin Wavelet dönüşümü altındaki 3B veri dağılımı



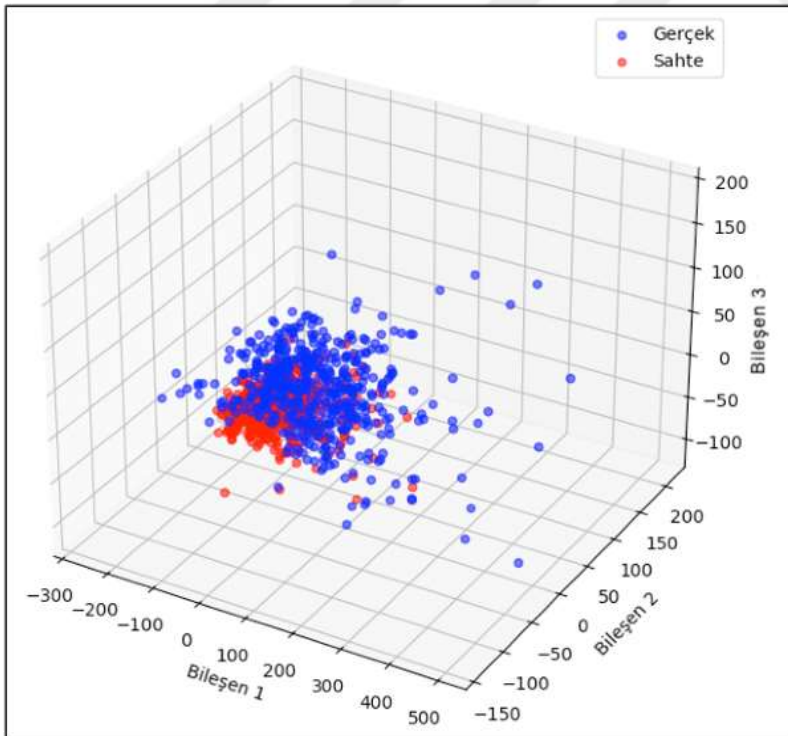
Şekil 4.7. CelebA ve StyleGAN3 görsellerinin Wavelet dönüşümü altındaki 3B veri dağılımı



Şekil 4.8. CelebA ve InterfaceGAN görsellerinin Fourier dönüşümü altındaki 3B veri dağılımı



Şekil 4.9. CelebA ve ProGAN görsellerinin Fourier dönüşümü altındaki 3B veri dağılımı



Şekil 4.10. CelebA ve StyleGAN3 görsellerinin Fourier dönüşümü altındaki 3B veri dağılımı

Tüm veri setleri ve dönüşüm işlemlerinin uygulanmasının ardından elde edilen değerlerin ölçeklendirilerek anlaşılabilir hale getirilmesi ile elde edilen değerler sınıflandırma algoritmaları tarafından analiz edilmiştir. Sınıflandırma algoritmalarının tespit süresince geçirdikleri zamana karşılık başarımlarının gösteren veriler InterfaceGAN, ProGAN ve StyleGAN3 için sırasıyla Çizelge 4.3, Çizelge 4.4 ve Çizelge 4.5’te sunulmuştur.

Çizelge 4.3. InterfaceGAN görselleri için sınıflandırma algoritmaları başarımlarının doğruluk oranları tespit süreleri

| Metot               | Fourier  |              | Wavelet  |              | HOG      |              |
|---------------------|----------|--------------|----------|--------------|----------|--------------|
| Sınıflandırıcı      | Doğruluk | Süre(saniye) | Doğruluk | Süre(saniye) | Doğruluk | Süre(saniye) |
| Naive Bayes         | 0,9211   | 0,03         | 0,9578   | 0,03         | 0,9888   | 0,03         |
| Logistic Regression | 0,9847   | 0,11         | 1        | 0,08         | 1        | 0,06         |
| kNN                 | 0,9825   | 0,87         | 0,9903   | 0,76         | 0,9986   | 0,78         |
| Gradient Boost      | 0,9947   | 33,45        | 0,9996   | 32,95        | 1        | 33,55        |
| Ada Boost           | 0,9894   | 10,19        | 0,9979   | 10,23        | 1        | 10,09        |
| XGBoost             | 0,9954   | 1,75         | 0,9991   | 0,38         | 1        | 1,54         |
| Random Forest       | 0,969    | 5,54         | 0,9845   | 5,4          | 0,9948   | 5,45         |
| Extra Trees         | 0,9765   | 0,63         | 0,9839   | 0,64         | 0,9963   | 0,62         |
| SVM                 | 0,9916   | 1,21         | 0,9997   | 1,3          | 0,9999   | 0,54         |
| SVM RBF             | 1        | 670,29       | 1        | 681,27       | 1        | 660,09       |
| YSA                 | 1        | 7,8          | 1        | 7,32         | 1        | 7,55         |
| GMM                 | 0,9741   | 0,72         | 0,5055   | 1,97         | 0,9978   | 0,63         |

SVM RBF ve YSA algoritmaları InterfaceGAN ile üretilen deepfake’lerin tamamını tespit edebilmiştir. Buna karşılık Naive Bayes algoritması ise diğer algoritmalara kıyasla en hızlı sürede tespit işlemi gerçekleştirebilmiştir. Tespit oranı ve geçen süre göz önünde bulundurulduğunda YSA algoritmasının InterfaceGAN ile oluşturulmuş deepfake görselleri tespit etmede oldukça başarılı olduğu görülmektedir.

Çizelge 4.4. ProGAN görselleri için sınıflandırma algoritmaları başarımlarının doğruluk oranları tespit süreleri

| Metot               | Fourier  |              | Wavelet  |              | HOG      |              |
|---------------------|----------|--------------|----------|--------------|----------|--------------|
| Sınıflandırıcı      | Doğruluk | Süre(saniye) | Doğruluk | Süre(saniye) | Doğruluk | Süre(saniye) |
| Naive Bayes         | 0,8817   | 0,03         | 0,909    | 0,04         | 0,969    | 0,04         |
| Logistic Regression | 0,8975   | 0,1          | 0,9885   | 0,12         | 0,9991   | 0,07         |
| kNN                 | 0,9131   | 0,88         | 0,9716   | 0,76         | 0,9979   | 0,74         |
| Gradient Boost      | 0,9483   | 33,43        | 0,996    | 33,25        | 0,9998   | 33,6         |
| Ada Boost           | 0,9021   | 10,13        | 0,9827   | 10,18        | 1        | 10,21        |
| XGBoost             | 0,9453   | 1,96         | 0,9931   | 0,34         | 0,9999   | 0,33         |
| Random Forest       | 0,882    | 5,43         | 0,974    | 5,38         | 0,9841   | 5,51         |
| Extra Trees         | 0,9066   | 0,66         | 0,9803   | 0,66         | 0,9942   | 0,67         |
| SVM                 | 0,9525   | 3,95         | 0,9975   | 1,89         | 0,9997   | 0,74         |
| SVM RBF             | 1        | 439,38       | 1        | 441,86       | 1        | 439,51       |
| YSA                 | 0,9929   | 6,8          | 0,9989   | 8,04         | 1        | 7,48         |
| GMM                 | 0,181    | 1,14         | 0,4806   | 2,45         | 0,997    | 0,74         |

Çizelge 4.5. StyleGAN3 görselleri için sınıflandırma algoritmaları başarımlarının doğruluk oranları tespit süreleri

| Metot               | Fourier  |              | Wavelet  |              | HOG      |              |
|---------------------|----------|--------------|----------|--------------|----------|--------------|
| Sınıflandırıcı      | Doğruluk | Süre(saniye) | Doğruluk | Süre(saniye) | Doğruluk | Süre(saniye) |
| Naive Bayes         | 0,8884   | 0,03         | 0,8715   | 0,03         | 0,9535   | 0,03         |
| Logistic Regression | 0,9103   | 0,08         | 0,9666   | 0,11         | 0,9892   | 0,09         |
| kNN                 | 0,9073   | 0,74         | 0,9636   | 0,75         | 0,9986   | 0,76         |
| Gradient Boost      | 0,9562   | 33,27        | 0,9902   | 33,48        | 0,9992   | 33,38        |
| Ada Boost           | 0,9212   | 10,24        | 0,9644   | 12,52        | 0,9965   | 10,19        |
| XGBoost             | 0,9525   | 0,36         | 0,9887   | 1,79         | 0,9982   | 0,33         |

Çizelge 4.5. (devam) StyleGAN3 görselleri için sınıflandırma algoritmaları başarımlarını doğruluk oranları tespit süreleri

|               |        |        |        |        |        |        |
|---------------|--------|--------|--------|--------|--------|--------|
| Random Forest | 0,908  | 5,56   | 0,9627 | 5,47   | 0,9868 | 5,49   |
| Extra Trees   | 0,9249 | 0,65   | 0,9593 | 0,73   | 0,9869 | 0,66   |
| SVM           | 0,9633 | 3,39   | 0,9958 | 2,14   | 0,9983 | 0,97   |
| SVM RBF       | 1      | 435,39 | 1      | 444,22 | 1      | 441,12 |
| YSA           | 0,9954 | 7,38   | 1      | 7,42   | 1      | 7,69   |
| GMM           | 0,1773 | 0,85   | 0,3066 | 2,49   | 0,6711 | 1,99   |

Özellik çıkarımı ve frekans uzayında incelemek üzere dönüşümler uygulanarak elde edilen veriler üzerinde yapılan analizler sonucu yapay sinir ağları (neural networks) sınıflandırılmasında tüm gerçek ve sahte görselleri bütün veri setleri için başarıyla tespit etmiştir. Bu nedenle doğruluk oranına ek olarak hassaslık, geri çağırma ve f1 skoru değerleri hesaplanmıştır.

Çizelge 4.6, Çizelge 4.7 ve Çizelge 4.8 sırasıyla InterfaceGAN, ProGAN ve StyleGAN3 için geri çağırma, hassasiyet ve f1 skoru değerlerini göstermektedir.

Çizelge 4.6. InterfaceGAN görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri

| Metot               | Fourier      |           |          | Wavelet      |           |          | HOG          |           |          |
|---------------------|--------------|-----------|----------|--------------|-----------|----------|--------------|-----------|----------|
| Sınıflandırıcı      | Geri Çağırma | Hassaslık | F1 Skoru | Geri Çağırma | Hassaslık | F1 Skoru | Geri Çağırma | Hassaslık | F1 Skoru |
| Naive Bayes         | 0,9211       | 0,9246    | 0,9209   | 0,9578       | 0,959     | 0,9578   | 0,9888       | 0,9889    | 0,9888   |
| Logistic Regression | 0,9847       | 0,9847    | 0,9847   | 1            | 1         | 1        | 1            | 1         | 1        |
| kNN                 | 0,9825       | 0,9826    | 0,9825   | 0,9903       | 0,9903    | 0,9903   | 0,9986       | 0,9986    | 0,9986   |
| Gradient Boost      | 0,9947       | 0,9947    | 0,9947   | 0,9996       | 0,9996    | 0,9996   | 1            | 1         | 1        |
| Ada Boost           | 0,9894       | 0,9894    | 0,9894   | 0,9979       | 0,9979    | 0,9979   | 1            | 1         | 1        |
| XGBoost             | 0,9954       | 0,9954    | 0,9954   | 0,9991       | 0,9991    | 0,9991   | 1            | 1         | 1        |
| Random Forest       | 0,969        | 0,9693    | 0,969    | 0,9845       | 0,9848    | 0,9845   | 0,9948       | 0,9948    | 0,9948   |
| Extra Trees         | 0,9765       | 0,9766    | 0,9765   | 0,9839       | 0,9842    | 0,9839   | 0,9963       | 0,9963    | 0,9963   |
| SVM                 | 0,9916       | 0,9916    | 0,9916   | 0,9997       | 0,9997    | 0,9997   | 0,9999       | 0,9999    | 0,9999   |

Çizelge 4.6. (devam) InterfaceGAN görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri

|         |        |        |        |        |        |        |        |        |        |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| SVM RBF | 1      | 1      | 1      | 1      | 1      | 1      | 1      | 1      | 1      |
| YSA     | 1      | 1      | 1      | 1      | 1      | 1      | 1      | 1      | 1      |
| GMM     | 0,9741 | 0,9748 | 0,9741 | 0,5055 | 0,5055 | 0,5047 | 0,9978 | 0,9978 | 0,9978 |

Çizelge 4.7. ProGAN görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri

| Metot               | Fourier      |           |          | Wavelet      |           |          | HOG          |           |          |
|---------------------|--------------|-----------|----------|--------------|-----------|----------|--------------|-----------|----------|
| Sınıflandırıcı      | Geri Çağırma | Hassaslık | F1 Skoru | Geri Çağırma | Hassaslık | F1 Skoru | Geri Çağırma | Hassaslık | F1 Skoru |
| Naive Bayes         | 0,8817       | 0,8825    | 0,8816   | 0,909        | 0,909     | 0,909    | 0,969        | 0,9695    | 0,969    |
| Logistic Regression | 0,8975       | 0,8975    | 0,8975   | 0,9885       | 0,9885    | 0,9885   | 0,9991       | 0,9991    | 0,9991   |
| kNN                 | 0,9131       | 0,9143    | 0,913    | 0,9716       | 0,9716    | 0,9716   | 0,9979       | 0,9979    | 0,9979   |
| Gradient Boost      | 0,9483       | 0,9489    | 0,9483   | 0,996        | 0,996     | 0,996    | 0,9998       | 0,9998    | 0,9998   |
| Ada Boost           | 0,9021       | 0,9023    | 0,9021   | 0,9827       | 0,9827    | 0,9827   | 1            | 1         | 1        |
| XGBoost             | 0,9453       | 0,946     | 0,9453   | 0,9931       | 0,9931    | 0,9931   | 0,9999       | 0,9999    | 0,9999   |
| Random Forest       | 0,882        | 0,8875    | 0,8816   | 0,974        | 0,974     | 0,974    | 0,9841       | 0,9843    | 0,9841   |
| Extra Trees         | 0,9066       | 0,9083    | 0,9065   | 0,9803       | 0,9803    | 0,9803   | 0,9942       | 0,9942    | 0,9942   |
| SVM                 | 0,9525       | 0,9527    | 0,9525   | 0,9975       | 0,9975    | 0,9975   | 0,9997       | 0,9997    | 0,9997   |
| SVM RBF             | 1            | 1         | 1        | 1            | 1         | 1        | 1            | 1         | 1        |
| YSA                 | 0,9929       | 0,9929    | 0,9929   | 0,9989       | 0,9989    | 0,9989   | 1            | 1         | 1        |
| GMM                 | 0,181        | 0,1372    | 0,1555   | 0,4806       | 0,4805    | 0,4799   | 0,997        | 0,997     | 0,997    |

Çizelge 4.8. StyleGAN3 görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri

| Metot               | Fourier      |           |          | Wavelet      |           |          | HOG          |           |          |
|---------------------|--------------|-----------|----------|--------------|-----------|----------|--------------|-----------|----------|
| Sınıflandırıcı      | Geri Çağırma | Hassaslık | F1 Skoru | Geri Çağırma | Hassaslık | F1 Skoru | Geri Çağırma | Hassaslık | F1 Skoru |
| Naive Bayes         | 0,8884       | 0,8884    | 0,8884   | 0,8715       | 0,8731    | 0,8714   | 0,9535       | 0,9538    | 0,9535   |
| Logistic Regression | 0,9103       | 0,9105    | 0,9103   | 0,9666       | 0,9666    | 0,9666   | 0,9892       | 0,9892    | 0,9892   |
| kNN                 | 0,9073       | 0,9165    | 0,9068   | 0,9636       | 0,9646    | 0,9636   | 0,9986       | 0,9986    | 0,9986   |

Çizelge 4.8. (devam) StyleGAN3 görselleri sınıflandırma algoritmaları geri çağırma, hassasiyet ve F1 skoru değerleri

|                |        |        |        |        |        |        |        |        |        |
|----------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Gradient Boost | 0,9562 | 0,9566 | 0,9562 | 0,9902 | 0,9902 | 0,9902 | 0,9992 | 0,9992 | 0,9992 |
| Ada Boost      | 0,9212 | 0,9216 | 0,9212 | 0,9644 | 0,9644 | 0,9644 | 0,9965 | 0,9965 | 0,9965 |
| XGBoost        | 0,9525 | 0,9534 | 0,9525 | 0,9887 | 0,9887 | 0,9887 | 0,9982 | 0,9982 | 0,9982 |
| Random Forest  | 0,908  | 0,911  | 0,9078 | 0,9627 | 0,9628 | 0,9627 | 0,9868 | 0,9868 | 0,9868 |
| Extra Trees    | 0,9249 | 0,9283 | 0,9248 | 0,9593 | 0,9599 | 0,9593 | 0,9869 | 0,9869 | 0,9869 |
| SVM            | 0,9633 | 0,9636 | 0,9633 | 0,9958 | 0,9958 | 0,9958 | 0,9983 | 0,9983 | 0,9983 |
| SVM RBF        | 1      | 1      | 1      | 1      | 1      | 1      | 1      | 1      | 1      |
| YSA            | 0,9954 | 0,9954 | 0,9954 | 1      | 1      | 1      | 1      | 1      | 1      |
| GMM            | 0,1773 | 0,1318 | 0,1511 | 0,3066 | 0,2188 | 0,2479 | 0,6711 | 0,801  | 0,6313 |

#### 4.6. GAN Simülasyonu Sonuç Analizi

Özgün veri seti ile test edilen simülasyon sentetik ve gerçek görseller HOG, Wavelet ve Fourier dönüşüm yöntemleri ile frekans uzayına dönüştürülmüş ve sınıflandırma algoritmalarına sunulan analiz sonuçları paylaşılmıştır. Veri seti ilk olarak insan gözü performansı ile değerlendirilerek çeşitli görsellerdeki bariz kalıntılar tespit edilmiş ve bu kalıntılar not edilmiştir. Bu anlamda StyleGAN3 modelinin en gerçekçi deepfake oluşturduğu açıkça görülmüştür. Geliştirilen simülasyon ile görsellerin üç farklı açıdan analiz edilmiştir. Bunlar; spektrum analizi, RGB renk uzayı korelasyonu ile analiz ve gerçek-sentetik görsellerin istatistiksel değerleri üzerinden analiz edilmiştir. Spektrum analizine göre elde edilen kıyaslamada örnek görseller üzerinden özellik çıkarımı metodu ile HOG, Wavelet ve Fourier dönüşümü uygulanarak frekans spektrumu analizi yapılmıştır. Frekans spektrum üzerinden yapılan analiz sonucunda spesifik GAN modeli ile üretilen görsellerin hep benzer şiddette olduğu görülmüştür. İndirgenmiş spektrum üzerinden yapılan analiz sonucunda ise InterfaceGAN ve ProGAN modelinde grafiğin oldukça doğrusal olduğu görülmüştür. Bu analiz yöntemi ile StyleGAN3 üzerinden bir sonuca varamayacağı düşünülmektedir. RGB renk uzayı üzerinden yapılan analiz sonucunda ise üç modelin de bu anlamda kalıntılarının görülebildiği tespit edilmiştir. RGB renk kanalı frekansının gerçek görseller ve sentetik görseller arasında çok net bir şekilde ayrıştığı üç GAN modeli için de görülmektedir. Bu anlamda ProGAN'ın renk kanalı frekansının ve RG-RB-GB ikili renk korelasyonları açısından gerçek görsellere daha yakın olduğu tespit edilmiştir.

GAN tespit yöntemi olarak simülasyondan elde edilen istatistikler paylaşılmıştır. Buna göre;

- Naive Bayes algoritması yaklaşık 0,03 saniyelik süresiyle en hızlı çalışan algoritmadır. Bu yöntem, diğer sınıflandırma algoritmalarına göre en az iki kat daha hızlı çalışmakta ve StyleGAN3, ProGAN ve InterfaceGAN veri setleri için sırasıyla 0,89, 0,91 ve 0,97 gibi oldukça yüksek ortalama doğruluk oranları sunmaktadır.
- Lojistik Regresyon sınıflandırma algoritması, Naive Bayes'e kıyasla iki ile üç kat daha yavaş çalışmasına rağmen, özellikle Wavelet dönüşümü kullanıldığında daha yüksek doğruluk oranları sağlamaktadır.
- kNN sınıflandırma algoritması, kullanılan veri setleri ve algoritma akışı açısından uygun bir sınıflandırıcı tercihi olmadığı görülmüştür. Lojistik Regresyon'dan yedi ile on kat arasında daha fazla işlem süresine ihtiyaç duymasına rağmen genellikle daha düşük doğruluk göstermektedir. Sadece ProGAN veri seti ile Fourier dönüşümünün birlikte kullanıldığı durumda, çok az bir farkla daha iyi performans sergilemektedir.
- Daha yüksek doğruluk oranları elde etmek için, belirli bir işlem süresi göze alındığında Gradient Boosting uygun bir sınıflandırma algoritması olarak öne çıkmaktadır. Ortalama yaklaşık yarım dakikalık işlem süresi ile Naive Bayes'ten yaklaşık bin kat daha yavaş olmasına rağmen, doğruluk oranlarında 7 ile 12 puan arasında daha iyi sonuç aldığı görülmüştür. Naive Bayes'in halihazırda en az %87 doğruluk sunduğu göz önünde bulundurulduğunda geçirilen süre kıyaslamasının anlamsız olabileceği söylenebilir.
- AdaBoost, Gradient Boosting'e kıyasla yalnızca üçte biri kadar işlem süresi gerektirir; ancak, performansı veri setine ve kullanılan dönüşüme bağlıdır. Gradient Boosting sınıflandırma algoritmasına kıyasla benzer doğruluk oranları sunarken çok daha verimli çalışan bir diğer yöntem ise XGBoost sınıflandırma algoritmasıdır; bu yöntem 30 ila 100 kat daha hızlıdır.
- Random Forest ortalama olarak beş saniyeden fazla işlem süresi gerektirmektedir ancak doğruluk oranı açısından Lojistik Regresyon'un gerisinde kalmaktadır. Extra Trees algoritması, Random Forest'a göre on kat daha hızlı çalışmakta ve benzer ya da daha iyi doğruluk değerleri sunmaktadır. Ancak, neredeyse benzer sonuçlara rağmen Lojistik Regresyon'un verimliliğine ulaşamamaktadır.
- SVM, Gradient Boosting'e kıyasla 10 ila 50 kat daha hızlı çalışmakta ve StyleGAN3 ile ProGAN veri setleri için Fourier ve Wavelet dönüşümleri uygulandığında daha yüksek doğruluk oranları göstermektedir. Diğer veri kümesi ve dönüşüm kombinasyonlarında ise sadece az bir farkla daha düşük başarımleri sergilemektedir.

- RBF çekirdekli SVM, Gradient Boosting ve AdaBoost algoritmaları, incelenen sınıflandırma algoritmaları arasında en yavaş çalışan üç yöntemdir. RBF çekirdekli SVM, StyleGAN3 ve ProGAN veri setleri için sınıflandırmayı yedi dakika gibi bir sürede, InterfaceGAN için ise on bir dakikada tamamlamaktadır.
- Yapay Sinir Ağları ve ince ayar yapılmış SVM algoritması genellikle %100'e yakın doğruluk oranları elde etmektedir. Ancak, RBF çekirdekli SVM'lerin parametrelerinin optimize edilmesi uzun süreler tüketmektedir ve diğer tüm yöntemlere göre birkaç kat daha uzun sürede çalışmaktadır. Buna karşın, yapay sinir ağları yalnızca 6 ila 8 saniyelik işlem süresi ile yüksek ve başarılı seviyede doğruluk sağlamaktadır. Deneylerimizde, tüm veri setleri ve dönüşümler için en iyi SVM parametreleri 'C': 10 ve 'gamma': 0,001 olarak belirlenmiştir.
- GMM, StyleGAN3 veri seti üzerinde tüm dönüşümlerle zayıf performans göstermiştir. ProGAN veri seti için yalnızca HOG dönüşümü ile anlamlı sonuçlar üretmiştir. InterfaceGAN veri seti için ise GMM hem HOG hem de Fourier dönüşüm yöntemleri iyi performans sergilemiştir.
- GMM sınıflandırma algoritmasının StyleGAN3 ile üretilen deepfake görsellerin tespitinde zorlandığı, %67 oranında tespit seviyesi ile en zayıf sınıflandırma algoritması olduğu görülmüştür.
- InterfaceGAN ile yapılan simülasyon sürecinde her bir veri seti için en uzun çalışma zamanı kaydedilmiştir. Ortalama süre olarak da en uzun süre bu veri seti için harcanmıştır.
- Gradient Boosting, StyleGAN3 ve ProGAN veri setleri üzerinde HOG dönüşümleri ile aynı ya da biraz daha iyi performans göstermektedir.
- Geçirilen süre ve tespit oranı karşılaştırılması sağlaması yapıldığında SVM sınıflandırma yönteminin en efektif çözüm olduğu göze çarpmıştır.

Analizler sonucunda elde edilen değerlere göre InterfaceGAN ve ProGAN modelleri ile üretilen sentetik görsellerin tespit edilebildiği fakat StyleGAN3 ile oluşturulan sentetik görsellerin çeşitli algoritmalar karşısından tespit oranlarında düşük değerler elde edildiği gözlemlenmiştir. StyleGAN ve StyleGAN2 gibi başarılı olduğu belirtilen modellerin daha gelişmiş bir sürümü olan StyleGAN3, GAN modelleri arasında sentetik görsel oluşturma hem insan gözü ile ayırt etmede hem de dijital yöntemlerle incelendiğinde dahi tespit edilmesi zor deepfake'ler oluşturduğu söylenebilir. Tüm kıyaslamalar göz önünde bulundurulduğunda insan algısının başarı oranının %50 civarında olduğu (Nickabadi,

Fard, Farid, ve Mohammadbagheri, 2022), bu seviyede veya daha az değerlerde elde edilen sonuçların hatalı olduğu değerlendirilebilir. Bu nedenle HOG dönüşüm yönteminin StyleGAN3 için uygun bir model olmadığı değerlendirilebilir. Elde ettiğimiz sonuçlara göre RealForensic (Beckman, Hilsman ve Eisert, 2023) çalışmasında edilen ortalama tespit sonuçları olan %75 ile %83 arasındaki tespit oranına kıyasla %20 ile %30 arasında daha iyi sonuçlar elde ettiğimiz görülmüştür. PGGAN ve DC-GAN kullanılarak oluşturulan, farklı çözünürlük ve boyutlarda görseller içeren VGG-Face veri seti kullanılarak derin öğrenme yöntemleri uygulanarak tespit yapılan çalışmada %80 tespit oranı elde edilmiştir (Do, Na ve Kim, 2018). Simülasyondaki en kötü sonucumuz olan StyleGAN3 veri seti için elde ettiğimiz GMM sınıflandırma algoritmasının %67 tespit performansı dahi hesaplama katıldığında ortalama %89 seviyesinde tespit edilebilirlik ile ilgili çalışmadan yaklaşık %10 daha iyi başarımlar elde ettiğimiz görülmüştür. Marra, Saltori, Boato ve Verdoliva (2019) tarafından farklı GAN modellerinin çalışmanın ilerleyen aşamalarında dahil edilmesi ile ayrıştırma hedefleyen ve ProGAN tarafından üretilen sahte görüntülerin tespiti için %80, CycleGAN ve StarGAN için ise %90 ve üzeri tespit başarımları elde eden çalışmalardan oldukça iyi sonuçlar aldığımız görülmektedir.

InterfaceGAN, ProGAN ve StyleGAN3 kullanarak oluşturduğumuz veri setinin HOG, Wavelet ve Fourier dönüşüm metodu kullanılarak frekans uzayında analizi sonucunda çeşitli algoritmalar tarafından incelenmesi ile tespit çalışması (Say, Alkan ve Koçak, 2025) yabancı dilde literatüre kazandırılmak adına yayımlanmıştır. Gerçekleştirdiğimiz daha kapsamlı çalışma sonucunda HOG dönüşüm yönteminin deepfake tespiti konusunda diğer yöntemlerden daha iyi olduğu, YSA ve SVM algoritmalarının dönüşüm yöntemleri ile ikilik sistemde en iyi sınıflandırıcı olarak öne çıktığı görülmüştür. Çalışma incelendiğinde özellikle SVM algoritmasının diğer sınıflandırma yöntemlerine göre neredeyse 400 kat yavaş kalmasına karşılık %100'e yakın bir tespit sonucu elde edebildiği görülmüştür. Zaman kısıtlamasının olmadığı veya göreceli olarak kabul edilebilir olduğu durumlarda SVM algoritması tercih edilebilir bir algoritma olarak öne çıkmaktadır.



## 5. SONUÇLAR VE ÖNERİLER

Bu çalışmada, farklı GAN modelleriyle üretilmiş sentetik yüz görsellerinin tespit edilebilirliğini değerlendirmek amacıyla kapsamlı bir analiz ve simülasyon süreci gerçekleştirilmiştir. İlk olarak, GAN kalıntılarına odaklanan literatür çalışmaları incelenmiş ve bu alanda literatürdeki açığı kapatacak bir taksonomi oluşturulmuştur. Taksonomiye göre GAN kalıntıları, görünür ve görünmez olmak üzere iki ana başlık altında sınıflandırılmış, görünmez kalıntılar ise frekans uzayı, renk kanalı farklılıkları, istatistiksel anormallikler ve sinyal işleme temelli özellikler olmak üzere dört alt başlığa ayrılmıştır. Bu sınıflandırma temel alınarak, çalışmamızda görünmez kalıntıların frekans ve renk uzayı düzeyinde incelendiği bir tespit simülasyonu geliştirilmiştir. Ek olarak bu simülasyonda kullanılmak üzere farklı GAN modelleri kullanılarak veri seti oluşturulmuştur. Çoklu GAN modeli kullanılan bu simülasyon alanındaki nadir örneklerden olmak üzere öne çıkmaktadır.

Simülasyonda; InterfaceGAN, ProGAN ve StyleGAN3 olmak üzere üç farklı GAN modeliyle üretilmiş toplam 15 000 sahte yüz görseli, bunlarla kıyaslama yapmak üzere ise CelebA veri setinden aynı sayıda gerçek görsel kullanılmıştır. Tüm görseller, HOG, Fourier ve Wavelet dönüşüm yöntemleri ile frekans uzayına aktarılmış, analiz edilmiş ve bu özellikler 12 farklı sınıflandırma algoritmasına sunulmuştur.

Simülasyon sonuçlarından öne çıkan bulgulara göre;

- Yapay sinir ağları, tüm GAN modelleri için %100 tespit başarısı ile en yüksek doğruluğa ulaşmıştır.
- SVM algoritması, ortalama %99 doğruluk oranı ile ikinci en başarılı yöntem olurken, aynı zamanda en hızlı çalışan algoritma olmuştur.
- GMM algoritması, özellikle StyleGAN3 görsellerinde yalnızca %67 doğruluk oranı ile en düşük performansı göstermiştir.
- InterfaceGAN ile yapılan simülasyonlarda, tüm veri setleri içinde en uzun işlem süresi bu modelde kaydedilmiştir.

Spektrum analizi sonucunda, InterfaceGAN ve ProGAN ile üretilen görsellerin frekans dağılımlarının daha belirgin ve doğrusal olduğu gözlemlenmiştir. Ancak StyleGAN3 görsellerinde bu yöntemle anlamlı bir ayırım yapılamamıştır. RGB renk uzayı analizi, üç

GAN modelinde de kalıntıların izlenebildiğini, ancak ProGAN modelinin renk kanalı frekansı ve RG-RB-GB korelasyonları açısından gerçek görsellere daha yakın çıktığını göstermiştir. İnsan gözüyle yapılan değerlendirmelerde de StyleGAN3, en gerçekçi görselleri üreten model olarak öne çıkmış; bariz kalıntılar daha az tespit edilmiştir. Bu durum hem görsel kalite hem de dijital iz bırakmama açısından StyleGAN3'ün daha gelişmiş olduğunu ortaya koymaktadır.

Sonuç olarak, InterfaceGAN ve ProGAN modelleriyle üretilen deepfake görseller hem spektrum hem de renk analizleri üzerinden yüksek doğrulukla tespit edilebilirken; StyleGAN3, mevcut analiz ve sınıflandırma yöntemleri karşısında daha dirençli olduğu için tespit edilmesi daha zordur. Bu durum, GAN tabanlı deepfake üretim teknolojisinin ne denli hızlı evrildiğini ve özellikle gelişmiş GAN modelleri karşısında mevcut tespit sistemlerinin yetersiz kalabileceğini göstermektedir.

StyleGAN3 ve benzeri gelişmiş GAN modelleri üzerine odaklanan daha derinlemesine analiz yöntemleri geliştirilmesi gerektiği çıkarılmıştır. Tespit algoritmalarında frekans-temelli özellikler ile renk uzayı analizlerinin birlikte kullanılması doğru bir yaklaşım olarak öne çıkmaktadır. Bu alanda yapılabilecek daha fazla çalışmanın gelişen deepfake oluşturma yöntemlerine karşı avantaj sağlayacağı söylenebilir. Özellikle spektrum analizine dayalı tekniklerin diğer yöntemlerle birlikte kullanılarak hibrit yapılar oluşturulması GAN üreteçli modellerin kalıntıları üzerinden tespit edilebilirliğini kolaylaştıracaktır.

## KAYNAKLAR

- Aduwala, S. A., Arigala, M., Desai, S., Quan, H. J., & Eirinaki, M. (2021, August). Deepfake detection using GAN discriminators. *IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, 69-77.
- Ahmed, S. R. A. (2023). *An enhanced convolutional neural network for detecting deepfake videos*, Doctoral Dissertation, Karabuk University Computer Engineering, Karabuk.
- Alkishri, W., Widyarto, S., & Yousif, J. H. (2024). Evaluating the effectiveness of a GAN fingerprint removal approach in fooling deepfake face detection. *Journal of Internet Services and Information Security*, 14(1), 85-103.
- Altun, S., ve Talu, M. F. (2021). Görüntüden görüntüye dönüşüm ve görüntü sentezleme yapan üretici çekişmeli ağların incelenmesi. *Avrupa Bilim ve Teknoloji Dergisi*, 53-60.
- Amerini, I., Anagnostopoulos, A., Maiano, L., & Celsi, L. R. (2021). Deep learning for multimedia forensics. *Foundations and Trends in Computer Graphics and Vision*, 12(4), 309-457.
- Arjovsky, M., Chintala, S. ve Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research*, 214-223.
- Arora, S., Ge, R., Liang, Y., Ma, T., & Zhang, Y. (2017, July). Generalization and equilibrium in generative adversarial nets (gans). *In International Conference on Machine Learning*, 224-232.
- Babayiğit, B. (2021). Deepfake'in ceza hukuku bakımından değerlendirilmesi ve de lege feranda öneriler. *Ankara Hacı Bayram Veli Üniversitesi Hukuk Fakültesi Dergisi*, 25(4), 655-703.
- Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2020, December). CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. *IEEE International Workshop on Information Forensics and Security*, 1-6.
- Bau, D., Zhu, J. Y., Strobel, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv Preprint*.
- Beckmann, A., Hilsmann, A., & Eisert, P. (2023, June). Fooling state-of-the-art deepfake detection with high-quality deepfakes. *In Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, 175-180.
- Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv Preprint*.

- Brooks, T., Hellsten, J., Aittala, M., Wang, T. C., Aila, T., Lehtinen, J. & Karras, T. (2022). Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 31769-31781.
- Canesche, M., Bragança, L., Neto, O. P. V., Nacif, J. A., & Ferreira, R. (2021, May). Google Colab CAD4U: Hands-on cloud laboratories for digital design. *IEEE International Symposium on Circuits and Systems*, 1-5.
- Carrara, F., Amato, G., Brombin, L., Falchi, F., & Gennaro, C. (2021, January). Combining gans and autoencoders for efficient anomaly detection. *25th International Conference on Pattern Recognition*, 3939-3946.
- Chadha, A., Kumar, V., Kashyap, S. and Gupta, M. (2021). Deepfake: An overview. *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, 557-566.
- Chai, L., Bau, D., Lim, S. N., & Isola, P. (2020). What makes fake images detectable? understanding properties that generalize. *Computer Vision Proceedings*, 103-120.
- Chaudhari, P., Agrawal, H., & Kotecha, K. (2020). Data augmentation using MG-GAN for improved cancer classification on gene expression data. *Soft Computing*, 11381-11391.
- Chhabra, S., Thakral, K., Mittal, S., Vatsa, M., & Singh, R. (2023). Low quality deepfake detection via unseen artifacts. *IEEE Transactions on Artificial Intelligence*, 5(4), 1573-1585.
- Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789-8797.
- Damer, N., Raja, K., Süßmilch, M., Venkatesh, S., Boutros, F., Fang, M., & Kuijper, A. (2021). Regenmorph: Visibly realistic GAN generated face morphing attacks by attack re-generation. *In Advances in Visual Computing: International Symposium Proceedings*, 251-264.
- Ding, Y., Thakur, N., & Li, B. (2021). Does a GAN leave distinct model-specific fingerprints?. *arXiv Preprint*.
- Do, N. T., Na, I. S., & Kim, S. H. (2018). Forensics face detection from GANs using convolutional neural network. *International Symposium on Information Technology Convergence*, 376-379.
- Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv Preprint*.
- Donahue, J., Krähenbühl, P., Darrell, T. (2016). Adversarial feature learning. *arXiv Preprint*.

Dong, C., Kumar, A., Liu, E. (2022). Think twice before detecting GAN-generated fake images from their spectral domain imprints. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7865-7874.

İnternet: Warzel, C. Faceapp shows we care about privacy but don't understand it. URL-1: <https://www.nytimes.com/2019/07/18/opinion/faceapp-privacy.html>, Son Erişim Tarihi: 19.05.2025.

İnternet: Ry, C. FaceApp was a test. We didn't pass. URL-2: <https://www.cnet.com/home/smart-home/faceapp-privacy-concerns/>, Son Erişim Tarihi: 19.05.2025.

İnternet: Awesome Face Forgery Generation and Detection. URL-3: <https://github.com/clpeng/Awesome-Face-Forgery-Generation-and-Detection>, Son Erişim Tarihi: 19.05.2025.

İnternet: Hindipur, A., The gan zoo/gans.tsv. URL-4: <https://github.com/hindupuravinash/the-gan-zoo/blob/master/gans.tsv>. Son Erişim Tarihi: 19.05.2025.

İnternet: Hindipur, A., The gan zoo. URL-5: <https://github.com/hindupuravinash/the-gan-zoo>. Son Erişim Tarihi: 19.05.2025.

İnternet: Dufour N., Google Research and Gully A., Jigsaw. URL-6: <https://research.google/blog/contributing-data-to-deepfake-detection-research/>. Son Erişim Tarihi: 19.05.2025.

İnternet: CycleGAN veri seti. URL-7: <https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>. Son Erişim Tarihi: 19.05.2025.

İnternet: StarGAN veri seti. URL-8: <https://github.com/yunjey/stargan>. Son Erişim Tarihi: 19.05.2025.

İnternet: generated. photos StyleGAN veri seti. URL-9: <https://generated.photos/>. Son Erişim Tarihi: 19.05.2025.

İnternet: CycleGAN veri seti. URL-10: <https://www.kaggle.com/datasets/utkarshsaxenadn/imagetoimage-translation-cyclegan/data>. Son Erişim Tarihi: 19.05.2025.

İnternet: Real-ESRGAN veri seti. URL-11: [https://github.com/xinntao/Real-ESRGAN/blob/master/docs/model\\_zoo.md](https://github.com/xinntao/Real-ESRGAN/blob/master/docs/model_zoo.md) Son Erişim Tarihi: 19.05.2025.

İnternet: iFakeFaceDB veri seti. URL-12: <https://github.com/socialabubi/iFakeFaceDB>. Son Erişim Tarihi: 19.05.2025.

İnternet: The DeepFake Detection Challenge (DFDC) veri seti. URL-13: <https://www.kaggle.com/c/deepfake-detection-challenge/>. Son Erişim Tarihi: 19.05.2025.

İnternet: Diverce Fake Face Dataset - DFFD. URL-14: <http://cvlab.cse.msu.edu/dffd-dataset.html>. Son Erişim Tarihi: 19.05.2025.

İnternet: 100K-Faces veri seti. URL-15: <https://github.com/NVlabs/stylegan>. Son Erişim Tarihi: 19.05.2025.

İnternet: CelebA-HQ veri seti. URL-16: [https://github.com/tkarras/progressive\\_growing\\_of\\_gans](https://github.com/tkarras/progressive_growing_of_gans). Son Erişim Tarihi: 19.05.2025.

İnternet: BigDatasetGAN veri seti. URL-17: [https://drive.google.com/drive/u/2/folders/1NC0VVSZrVZsd1Z\\_IRSdadyfCfTXMAPsfl](https://drive.google.com/drive/u/2/folders/1NC0VVSZrVZsd1Z_IRSdadyfCfTXMAPsfl). Son Erişim Tarihi: 19.05.2025.

İnternet: DCGAN veri seti. URL-18: [https://figshare.com/articles/dataset/Dataset\\_Outlier\\_GAN\\_/14754309](https://figshare.com/articles/dataset/Dataset_Outlier_GAN_/14754309). Son Erişim Tarihi: 19.05.2025.

İnternet: LSUN veri seti. URL-19: <https://github.com/fyu/lsun>. Son Erişim Tarihi: 19.05.2025.

İnternet: BigGAN-Deep veri seti. URL-20: <https://github.com/ajbrock/BigGAN-PyTorch>. Son Erişim Tarihi: 19.05.2025.

İnternet: StackGAN tarafından üretilen veri seti. URL-21: <https://github.com/hanzhanggit/StackGAN>. Son Erişim Tarihi: 19.05.2025.

Dundar, A., Gao, J., Tao, A., & Catanzaro, B. (2023). Progressive learning of 3d reconstruction network from 2d gan data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2), 793-804.

Dzanic, T., Shah, K., & Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. *Advances in Neural Information Processing Systems*, 3022-3032.

Elaraby, N., Barakat, S., & Rezk, A. (2022). A conditional GAN-based approach for enhancing transfer learning performance in few-shot HCR tasks. *Scientific Reports*, 12(1), 16271.

Erdoğan, M. A. (2023). *Deepfake teknolojisinin hukuka aykırı kullanımından doğan hukuki ve cezai sorumluluk*. Yayımlanmamış Yüksek Lisans Tezi. Altınbaş Üniversitesi Kamu Hukuku Ana Bilim Dalı, İstanbul.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020, November). Leveraging frequency analysis for deep fake image recognition. *International Conference on Machine Learning*, 3247-3258.

Gargar, S. S. (2022). Explicit use of fourier spectrum in generative adversarial networks. *arXiv Preprint*.

- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 580-587.
- Giudice, O., Guarnera, L., & Battiato, S. (2021). Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, 7(8), 128.
- Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: datasets, algorithms, and challenges. *electronics*, 13(3), 585.
- Gong, W. (2017). *Wasserstein generative adversarial network*. Master Thesis. Department of Engineering University of Cambridge, England.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. and Bengio, Y. (2014). Generative adversarial networks. *arXiv Cornell University*, 63(11), 139-144.
- Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., & Verdoliva, L. (2021, July). Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. *In 2021 IEEE International Conference on Multimedia and Expo*, 1-6.
- Granot, N., Feinstein, B., Shocher, A., Bagon, S., & Irani, M. (2022). Drop the gan: In defense of patches nearest neighbors as single image generative models. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13460-13469.
- Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. (2020). Development of video-based emotion recognition using deep learning with Google Colab. *Telecommunication Computing Electronics and Control*, 18(5), 2463-2471.
- Guo, H., Hu, S., Wang, X., Chang, M. C., & Lyu, S. (2022, May). Eyes tell all: Irregular pupil shapes reveal gan-generated faces. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2904-2908.
- Hackney, D. N. (2023). Analysis of artifact formation and removal in GAN training doctoral dissertation, University of Cincinnati, Ohio.
- Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2), 15-20.
- Huang, Y., Juefei-Xu, F., Wang, R., Xie, X., Ma, L., Li, J. & Pu, G. (2020). Fakelocator: Robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. *arXiv Preprint*.
- Huang, Y., Juefei-Xu, F., Wang, R., Guo, Q., Ma, L., Xie, X. & Pu, G. (2020, October). Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. *In Proceedings of the 28th ACM International Conference on Multimedia*, 1217-1226.

- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125-1134.
- Jenni, S., & Favaro, P. (2018). Self-supervised feature learning by learning to spot artifacts. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2733-2742.
- Jeong, Y., Kim, D., Kim, P., Ro, Y., & Choi, J. (2021). Self-supervised gan detector. *arXiv Preprint*.
- Jeong, Y., Kim, D., Min, S., Joe, S., Gwon, Y., & Choi, J. (2022). Bihpf: Bilateral high-pass filters for robust deepfake detection. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 48-57.
- Jia, L., (2022). *Using three GAN-based models to provide modeling inspiration*. Master Thesis, Department of Computer Science, University of Reading, Berkshire.
- Johnson, M. K., & Farid, H. (2007). Exposing digital forgeries through specular highlights on the eye. *In Information Hiding: 9th International Workshop*, 311-325.
- 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.
- Kamar, D., Üre, N. K., & Ünal, G. (2022). GAN-based intrinsic exploration for sample efficient reinforcement learning. *arXiv Preprint*.
- Kanani, P., & Padole, M. (2019). Deep learning to detect skin cancer using google colab. *International Journal of Engineering and Advanced Technology Regular Issue*, 8(6), 2176-2183.
- Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv Preprint*.
- Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 852-863.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401-4410.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110-8119.
- Keleş, M. K., & Güvenoğlu, E. (2023). Çekişmeli üretici ağlar kullanılarak hasarlı mozaik görüntülerinin tamamlanması. *Journal of the Institute of Science and Technology*, 13(2), 736-746.

- Kench, S., & Cooper, S. J. (2021). Generating 3D structures from a 2D slice with GAN-based dimensionality expansion. *arXiv Preprint*.
- Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv Preprint*.
- Khodabakhsh, A., Ramachandra, R., & Busch, C. (2019, June). Subjective evaluation of media consumer vulnerability to fake audiovisual content. *Eleventh International Conference on Quality of Multimedia Experience*, 1-6.
- Khoo, B., Phan, R. C. W., & Lim, C. H. (2022). Deepfake attribution: On the source identification of artificially generated images. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3), 14-38.
- Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? assessment and detection. *arXiv Preprint*.
- Kumar, M., & Sharma, H. K. (2023). A GAN-based model of deepfake detection in social media. *Procedia Computer Science*, 218, 2153-2162.
- Le, B. M., & Woo, S. S. (2021). Exploring the asynchronous of the frequency spectra of gan-generated facial images. *arXiv Preprint*.
- Lee, S., Tariq, S., & Shin, Y. (2020). Hand-crafted facial manipulation (HFM) dataset. *Mendeley Data*, 1.
- Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z. & Palaniappan, K. (2020, October). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. *IEEE Applied Imagery Pattern Recognition Workshop*, 1-9.
- Li, H., Li, B., Tan, S., & Huang, J. (2020). Identification of deep network generated images using disparities in color components. *arXiv Preprint*.
- Li, M., Zhong, Q., Zhang, L. Y., Du, Y., Zhang, J., & Xiang, Y. (2020, December). Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach. *IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications*, 402-409.
- Li, W., He, P., Li, H., Wang, H., & Zhang, R. (2021). Detection of GAN-generated images by estimating artifact similarity. *IEEE Signal Processing Letters*, 862-866.
- Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint*.
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3207-3216.

- Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *In Proceedings of the IEEE International Conference on Computer Vision*, 3730-3738.
- Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019, March). Do gans leave artificial fingerprints?. *IEEE Conference on Multimedia Information Processing and Retrieval*, 506-511.
- Marra, F., Saltori, C., Boato, G., & Verdoliva, L. (2019, December). Incremental learning for the detection and classification of gan-generated images. *IEEE International Workshop on Information Forensics and Security*, 1-6.
- Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2020). Two-branch recurrent network for isolating deepfakes in videos. *Computer Vision*, 667-684.
- Matern, F., Riess, C., & Stamminger, M. (2019, January). Exploiting visual artifacts to expose deepfakes and face manipulations. *Winter Applications of Computer Vision Workshops*, 83-92.
- McCloskey, S., & Albright, M. (2018). Detecting gan-generated imagery using color cues. *arXiv Preprint*.
- McCloskey, S., & Albright, M. (2019, September). Detecting GAN-generated imagery using saturation cues. *International Conference on Image Processing*, 4584-4588.
- Mo, H., Chen, B., & Luo, W. (2018, June). Fake faces identification via convolutional neural network. *ACM workshop on Information Hiding and Multimedia Security*, 43-47.
- Mohebbi Moghaddam, M., Boroomand, B., Jalali, M., Zareian, A., Daeijavad, A., Manshaei, M. H., & Krunz, M. (2023). Games of GANs: Game-theoretical models for generative adversarial networks. *Artificial Intelligence Review*, 56(9), 9771-9807.
- Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., & Manjunath, B. S. (2019). Detecting GAN generated fake images using co-occurrence matrices. *arXiv Preprint*.
- Natsume, R., Yatagawa, T., & Morishima, S. (2018). Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv Preprint*.
- Nickabadi, A., Fard, M. S., Farid, N. M., & Mohammadbagheri, N. (2022). A comprehensive survey on semantic facial attribute editing using generative adversarial networks. *arXiv Preprint*.
- Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8), 1-3.
- Ntavelis, E., Romero, A., Kastanis, I., Van Gool, L., & Timofte, R. (2020). Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. *Computer Vision*, 394-411.

- Ortaç, G., Doğan, Z., Orman, Z., & Şamlı, R. (2020). Baby face generation with generative adversarial neural networks: a case study. *Acta Infologica*, 4(1), 1-9.
- Oshea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv Preprint*.
- Öztemel, E. (2006). *Yapay Sinir Ağları* (3. Baskı). İstanbul: Papatya Yayıncılık.
- Padmanabhan, S. ve Kishan, S. (2021). Deepfake creation and detection: A survey. *Proceedings of the Third International Conference on Inventive Research in Computing Applications*, 584-588.
- Pei, G., Zhang, J., Hu, M., Zhai, G., Wang, C., Zhang, Z. & Tao, D. (2024). Deepfake generation and detection: a benchmark and survey. *arXiv Preprint*.
- Peng C. L., Gao X., B. & Wang N, N. (2021). Deep visual identity forgery and detection (in Chinese). *Scientia Sinica Informationis*, 1451-1474.
- Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., & Moreno-Noguer, F. (2020). Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 698-713.
- Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020, August). Thinking in frequency: Face forgery detection by mining frequency-aware clues. *European Conference on Computer Vision*, 86-103.
- Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 25494-25513.
- Reesink, T.B.J. (2020) *Creating 3D faces from 2D images using GANs*, Master Thesis, Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, Netherlands.
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1-11.
- Sandotra, N. ve Arora, B. (2023). A comprehensive evaluation of feature-based AI techniques for deepfake detection. *Neural Computing and Applications*, 36(8), 3859-3887.
- Saxena, P., Gupta, R., Maheshwari, A., & Maheshwari, S. (2020). Semantic image completion and enhancement using gans. *High Performance Vision Intelligence: Recent Advances*, 151-170.
- Say, T., Alkan, M., & Kocak, A. (2025). Advancing GAN deepfake detection: Mixed datasets and comprehensive artifact analysis. *Applied Sciences*, 15(2), 923-947.

- Shaham, T. R., Dekel, T., & Michaeli, T. (2019). Singan: Learning a generative model from a single natural image. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4570-4580.
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306-132322.
- Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2016). Learning from simulated and unsupervised images through adversarial training. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2107-2116.
- Sorensen, M. M. Z. (2024). Deepfake face-Swap animations and affect: In human perception and digital information technologies. *Bristol University Press*, 195-212.
- Suri, S., Meshry, M., Davis, L. S., & Shrivastava, A. (2024). GRIT: GAN residuals for paired image-to-image translation. *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4965-4975.
- Tan, W., Wen, B., Chen, C., Zeng, Z., & Yang, X. (2021, September). Systematic analysis of circular artifacts for stylegan. *IEEE International Conference on Image Processing*, 3902-3906.
- Tejero, Á. J. G. (2020). *Detecting overfitting in GANs with a metric based on the Fourier spectrum*. Master Thesis. Universidad Nacional de Educacion, Ekvador.
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J. (2020). Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64, 131-148.
- Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8695-8704.
- Wang, W. (2022, September). Evolution of StyleGAN3. *International Conference on Electronics and Devices, Computational Science*, 5-13.
- Wang, X., Guo, H., Hu, S., Chang, M. C., & Lyu, S. Gan-generated faces detection: A survey and new perspectives. *arXiv Preprint*.
- Wang, Y. A mathematical introduction to generative adversarial nets (GAN). arXiv 2020. *arXiv Preprint*.
- Wesselkamp, V., Rieck, K., Arp, D., & Quiring, E. (2022, May). Misleading deep-fake detection with GAN fingerprints. *IEEE Security and Privacy Workshops*, 59-65.
- Wirth, S. (2023). InterFace experiments: FaceApp as everyday AI. *Interface Critique*, 159-169.

- Wu, X., Xie, Z., Gao, Y., & Xiao, Y. (2020, May). Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2952-2956.
- Xia, W., Zhang, Y., Yang, Y., Xue, J. H., Zhou, B., & Yang, M. H. (2022). Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 3121-3138.
- Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X., & Ding, N. (2022). GAN-based anomaly detection: A review. *Neurocomputing*, 497-535.
- Xie, L., Wang, X., Chen, X., Li, G., Shan, Y., Zhou, J., & Dong, C. (2023). Desra: detect and delete the artifacts of gan-based real-world super-resolution models. *arXiv Preprint*.
- Yang, H., & Weinan E., (2022). Generalization error of GAN from the discriminator's perspective. *Research in the Mathematical Sciences*, 9(1), 8-36.
- Yang, X., Li, Y., & Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 8261-8265.
- Yang, X., Li, Y., Qi, H., & Lyu, S. (2019, July). Exposing GAN-synthesized faces using landmark locations. *ACM Workshop on Information Hiding and Multimedia Security*, 113-118.
- Yılmaz, B., & Korn, R. (2023). Understanding the mathematical background of Generative Adversarial Networks (GANs). *Mathematical Modelling and Numerical Simulation with Applications*, 3(3), 234-255.
- Yu, N., Davis, L. S., & Fritz, M. (2019). Attributing fake images to gans: Learning and analyzing gan fingerprints. *International Conference on Computer Vision*, 7556-7566.
- Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *In Proceedings of the IEEE International Conference on Computer Vision*, 5907-5915.
- Zhang, M., Wang, H., He, P., Malik, A., & Liu, H. (2022). Exposing unseen GAN-generated image using unsupervised domain adaptation. *arXiv Preprint*.
- Zhang, N., Liu, X., Li, X., & Qi, G. J. (2023). Morphganformer: Transformer-based face morphing and de-morphing. *arXiv Preprint*.
- Zhang, X., Karaman, S., & Chang, S. F. (2019, December). Detecting and simulating artifacts in gan fake images. *International Workshop on Information Forensics and Security*, 1-6.

- Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). Learning rich features for image manipulation detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 1053-1061.
- Zhu, J.-Y., Park, T., Isola, P. ve Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *In Proceedings of the IEEE International Conference on Computer Vision*, 2223-2232.
- Zhu, Y., Wang, X., Salloum, R., Chen, H. S., & Kuo, C. C. J. (2022). Rggid: A robust and green gan-fake image detector. *APSIPA Transactions on Signal and Information Processing*, 11(2), 1-26.





*Gazili olmak ayrıcalıktır*