**T. C.**

**İSTANBUL GELİŞİM ÜNİVERSİTESİ**

Yönetim Bilişim Sistemleri

**Yüz Manipülasyonlarının Tespiti: EfficientNetV2S Tabanlı Bir Model Önerisi**

Lisans Tezi

**Ertuğrul Köse**

Danışman

Dr.Öğr.Üyesi HADI POURMOUSA

**İstanbul – 2025**

**TEZ TANITIM FORMU**

|  |  |  |
| --- | --- | --- |
| **Yazar Adı Soyadı** | **:** | Ertuğrul Köse |
| **Tezin Dili** | **:** | Türkçe |
| **Tezin Adı** | **:** | Yüz Manipülasyonlarının Tespiti: EfficientNetV2S Tabanlı Bir Model Önerisi |
| **Enstitü** | **:** | İstanbul Gelişim Üniversitesi |
| **Anabilim Dalı** | **:** | Yönetim Bilişim Sistemleri |
| **Tezin Türü** | **:** | Lisans |
| **Tezin Tarihi** | **:** | 01.06.2025 |
| **Sayfa Sayısı** | **:** | 57 |
| **Tez Danışmanları** | **:** | Dr.Öğr.Üyesi HADI POURMOUSA |
| **Dizin Terimleri** | **:** |  |
| **Türkçe Özet** | **:** |  |
| **Dağıtım Listesi** | **:** | 1. İstanbul Gelişim Üniversitesi Lisansüstü Eğitim Enstitüsüne 2. YÖK Ulusal Tez Merkezine |

*İmzası*

*Adı SOYADI*

**T. C.**

**İSTANBUL GELİŞİM ÜNİVERSİTESİ**

Yönetim Bilişim Sistemleri

**Yüz Manipülasyonlarının Tespiti: EfficientNetV2S Tabanlı Bir Model Önerisi**

Lisans Tezi

**Ertuğrul Köse**

Danışman

Dr.Öğr.Üyesi HADI POURMOUSA

**İstanbul – 2025**

**BEYAN**

Bu tezin hazırlanmasında bilimsel ahlak kurallarına uyulduğu, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğu, kullanılan verilerde herhangi tahrifat yapılmadığını, tezin herhangi bir kısmının bu üniversite veya başka bir üniversitedeki başka bir tez olarak sunulmadığını beyan ederim.

Ertuğrul Köse

TARİH

|  |  |  |  |
| --- | --- | --- | --- |
| **İSTANBUL GELİŞİM ÜNİVERSİTESİ**  **LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜNE** | | | |
| ...............................................’ın ................................................................................. ........................................................................... adlı tez çalışması, jürimiz tarafından ......................................... anabilim dalı, ......................................... bilim dalında YÜKSEK LİSANS / DOKTORA tezi olarak kabul edilmiştir. | | | |
|  | Başkan | *İmza* |  |
| *Prof. Dr. Adı SOYADI* |
|  | Üye | *İmza* |  |
| *Prof. Dr. Adı SOYADI*  (Danışman) |
|  | Üye | *İmza* |  |
| *Prof. Dr. Adı SOYADI* |
|  | Üye | *İmza* |  |
| *Doç. Dr. Adı SOYADI* |
|  | Üye | *İmza* |  |
| *Dr. Adı SOYADI*  (Var ise İkinci Danışman) |
| ONAY  Yukarıdaki imzaların, adı geçen öğretim üyelerine ait olduğunu onaylarım.  ... / ... / 20..  *İmzası*  *Prof. Dr. Ad SOYADI*  Enstitü Müdürü | | | |

ÖZET

Son yıllarda yapay zekâ teknikleri ile üretilen manipüle edilmiş video ve görseller, özellikle “deepfake” olarak adlandırılan, içerikler, bireylerin mahremiyetini tehdit etmekte ve kamuoyunu yanıltmak amacıyla kötüye kullanılabilmektedir. Bu sebeple gerçek ve manipüle edilmiş içeriklerin ayrımını yapabilecek otomatik sistemlerin geliştirilmesi büyük önem arz etmektedir. Bu tez çalışmasında, EfficientNetV2S mimarisi kullanılarak deepfake görüntülerin tespit edilmesine yönelik bir model geliştirilmiştir.

Çalışma kapsamında, Celeb-DF ve DFD adlı yüksek kaliteli ve güncel veri setleri kullanılmış, doğrulama ve test veri setleri oluşturulmuştur. Ön işleme sürecinde MTCNN kullanılarak yüz çıkarımı yapılmış; ardından görüntüler EfficientNetV2S modeline uygun boyutlandırılmış ve normalize edilmiştir.

Model eğitiminin ilk aşamasında %81 doğruluk ve %82 AUC değerine ulaşılmıştır. Daha sonra model üzerinde fine-tuning (ince ayar) uygulanarak eğitim sonunda model doğruluk oranı %89’a, AUC skoru ise %96’ya çıkarılmıştır. Elde edilen sonuçlar, EfficientNetV2S tabanlı yaklaşımların deepfake tespiti gibi karmaşık sınıflandırma problemlerinde başarılı olduğunu ortaya koymuştur.

**Anahtar Kelimeler:** Derin Sahtecilik, EfficientNetV2S, Aktarım Öğrenmesi, Derin Öğrenme

SUMMARY

In recent years, manipulated videos and images generated using artificial intelligence techniques—particularly those known as deepfakes—have posed a serious threat to individuals' privacy and have been maliciously exploited to mislead the public. Therefore, the development of automated systems capable of distinguishing between real and manipulated content has become critically important.

In this thesis study, a model for detecting deepfake images was developed using the EfficientNetV2S architecture. High-quality and up-to-date datasets, Celeb-DF and DFD, were utilized to construct validation and test sets. During the preprocessing stage, faces were extracted using MTCNN, and the images were resized and normalized to be compatible with the EfficientNetV2S model.

In the initial training phase, the model achieved 81% accuracy and an AUC score of 82%. After applying fine-tuning techniques, the model's accuracy increased to 89%, and the AUC score reached 96%. These results demonstrate that EfficientNetV2S-based approaches are effective for complex classification tasks such as deepfake detection.

**Keywords:** DeepFake, EfficientNetV2S, Transfer Learning, Deep Learning

İÇİNDEKİLER

[ÖZET i](#_Toc198845173)

[SUMMARY ii](#_Toc198845174)

[İÇİNDEKİLER iii](#_Toc198845175)

[TABLOLAR LİSTESİ vi](#_Toc198845176)

[ŞEKİLLER LİSTESİ viii](#_Toc198845177)

[ÖNSÖZ ix](#_Toc198845178)

[Birinci bölüm 10](#_Toc198845179)

[Giriş 10](#_Toc198845180)

[1.1. Yapay Sinir Ağları Kavramı 10](#_Toc198845181)

[1.2. Yapay Sinir Ağının Bileşenleri 11](#_Toc198845182)

[1.1.1. Nöronlar 11](#_Toc198845183)

[1.1.2. Giriş Katmanı 12](#_Toc198845184)

[1.1.3. Gizli Katman 12](#_Toc198845185)

[1.1.4. Çıkış Katmanı 12](#_Toc198845186)

[1.1.5. Bağlantı Ağırlıkları 12](#_Toc198845187)

[1.1.6. Aktivasyon fonksiyonları 13](#_Toc198845188)

[1.2. Yapay Sinir Ağı Mimarisi 14](#_Toc198845189)

[1.2.1. Tek Katmanlı Ağ 14](#_Toc198845190)

[1.2.2. Çok Katmanlı Ağ 15](#_Toc198845191)

[1.2.3. Tekrarlayan Sinir Ağları (RNN) 16](#_Toc198845192)

[1.2.4. Uzun Kısa Süreli Bellek (LSTM) 17](#_Toc198845193)

[1.2.5. Evrişimsel Sinir Ağları 18](#_Toc198845194)

[1.2.5.1. Evrişim katmanı 18](#_Toc198845195)

[1.2.5.2. Havuzlama katmanı 19](#_Toc198845196)

[1.2.5.3. Tam bağlantı katmanı 20](#_Toc198845197)

[1.2.5.4. Çekişmeli üretken ağlar (GAN) 21](#_Toc198845198)

[1.3. Deepfake Nedir? 22](#_Toc198845199)

[İkinci bölüm 24](#_Toc198845200)

[LİTERATÜR TARAMASI 24](#_Toc198845201)

[2.1. Görüntü üretimi üzerine yapılan çalışmalar 24](#_Toc198845202)

[2.2. Deepfake tespiti için oluşturulmuş veri setleri 25](#_Toc198845203)

[2.2.1. Celeb-DF 25](#_Toc198845204)

[2.2.2. FaceForensics++ 25](#_Toc198845205)

[2.2.3. DFDC (Deepfake Detection Challenge) 26](#_Toc198845206)

[2.2.4. DFD (Deep Fakes Dataset) 26](#_Toc198845207)

[2.2.5. UADFV 26](#_Toc198845208)

[2.2.6. DeepfakeTIMIT 27](#_Toc198845209)

[2.3. Deepfake Tespiti Üzerine Yapılan Çalışmalar 28](#_Toc198845210)

[Üçüncü bölüm 40](#_Toc198845211)

[YöntEm (metodoloji) 40](#_Toc198845212)

[3.1. Veri Seti 40](#_Toc198845213)

[3.1.1. 3.1.1 Veri örnekleme ve dengeleme süreci 41](#_Toc198845214)

[3.2. Veri Ön İşleme Aşamaları 42](#_Toc198845215)

[3.2.1. Video karelerine ayrıştırma ve yüz tespiti 42](#_Toc198845216)

[3.2.2. Veri etiketleme ve DataFrame oluşturma 42](#_Toc198845217)

[3.2.3. Sınıf dengesizliği giderme 43](#_Toc198845218)

[3.2.4. Görüntü ölçekleme ve ön işleme 43](#_Toc198845219)

[3.2.5. Görüntü akışlarının (generator) oluşturulması 43](#_Toc198845220)

[3.3. Model Mimarisi ve Eğitimi 43](#_Toc198845221)

[3.3.1. Modelin temeli EfficentNetV2S ile transfer learning 43](#_Toc198845222)

[3.3.2. Yeni katmanların eklenmesi ve özelleştirilmesi 44](#_Toc198845223)

[3.3.3. Modelin derlenmesi ve eğitim aşaması 45](#_Toc198845224)

[3.4. Fine-Tuning (İnce Ayar) 46](#_Toc198845225)

[Dördüncü bölüm 47](#_Toc198845226)

[Bulgular 47](#_Toc198845227)

[4.1. Ön Eğitim Aşaması Sonuçları 47](#_Toc198845228)

[4.2. Fine-Tuning (İnce Ayar) Aşaması Sonuçları 49](#_Toc198845229)

[4.3. Literatürdeki Diğer Çalışmalarla Karşılaştırma 51](#_Toc198845230)

[Beşinci bölüm 52](#_Toc198845231)

[Sonuçlar 52](#_Toc198845232)

[Altıncı bölüm 53](#_Toc198845233)

[Öneriler 53](#_Toc198845234)

[Yedinci bölüm 54](#_Toc198845235)

[Kaynakça 54](#_Toc198845236)

TABLOLAR LİSTESİ

[Tablo 1: Yapay Sinir Ağı ve Beyin Sinir Ağının Karşılıklı Tanımlamaları. 11](#_Toc198845237)

[Tablo 2:Aktivasyon Fonksiyonları 15](#_Toc198845238)

[Tablo 3 : Deepfake Tespitinde Kullanılan Popüler Veri Setleri 29](#_Toc198845239)

[Tablo 4 : Deepfake Tespit Yöntemlerinin Özeti 37](#_Toc198845240)

[Tablo 5: Model Katmanları 46](#_Toc198845241)

[Tablo 6 : Model Derlenme ve Eğtim Bileşenleri 46](#_Toc198845242)

[Tablo 7 : Fine-Tuning Süreci Özet 48](#_Toc198845243)

[Tablo 8: Modelin Sonuç Özeti 52](#_Toc198845244)

[Tablo 9 : Literatürdeki Diğer Çalışmalar ve Mevcut Çalışmanın Karşılaştırılması 53](#_Toc198845245)

ŞEKİLLER LİSTESİ

[Şekil 1: Yapay Sinir Ağına Ait Bir Nöron/Işlem Birimi. 12](#_Toc198845252)

[Şekil 2: Yapay Sinir Ağı Bileşenleri 14](#_Toc198845253)

[Şekil 3: Tek Katmanlı Ağ 16](#_Toc198845254)

[Şekil 4: Çok Katmanlı Yapay Sinir Ağı Model 17](#_Toc198845255)

[Şekil 5: (Solda) RNN Için Devre Şeması Gösterilmiştir. (Sağda) Aynı RNN Için Açılmış (Unfolded) Hesaplama Şeması Gösterilmiştir. 17](#_Toc198845256)

[Şekil 6 : LSTM Yapısı Hücre Durumu **c t**,Gizli Durum **h t**, Girdi **x t** ve Çıktı **o t** 18](#_Toc198845257)

[Şekil 7 : Klasik Bir CNN Mimarisi 19](#_Toc198845258)

[Şekil 8 : İki Boyutlu Bir Tensör Üzerindeki Evrişim İşlemine Bir Örnek. 20](#_Toc198845259)

[Şekil 9 : 2x2’lik Maksimum Havuzlama ve Ortalama Havuzlama Yöntemi 21](#_Toc198845260)

[Şekil 10: Tam Bağlantı Katmanı Şeması 22](#_Toc198845261)

[Şekil 11: Çekişmeli Üretken Ağ Yapısı. 23](#_Toc198845262)

[Şekil 12 : (Sol) Google Arama Motorunun ‘'Deepfake'' Anahtar Kelimesini Içeren Web Sayfa Sayısı Sonuçları. (Sağ) Google Arama Motorunun “Deepfake” Ile Ilgili Videoları Içeren Web Sayfa Sayısı Sonuçları. 24](#_Toc198845263)

[Şekil 13 : Ön Eğitim Aşaması Sırasında Her Bir Epoch Sonunda Elde Edilen Accuracy, Loss ve AUC Değerleri 50](#_Toc198845264)

[Şekil 14 : Fine-Tuning Aşaması Sırasında Her Bir Epoch Sonunda Elde Edilen Accuracy, Loss ve AUC Değerleri 52](#_Toc198845265)

ÖNSÖZ

Bu tez çalışması, günümüzde dijital ortamda artan manipülasyon teknikleri ve özellikle deepfake içeriklerin oluşturduğu tehditleri ele alarak, yapay zekâ tabanlı bir çözüm önerisi sunmayı amaçlamaktadır. Akademik bilgi birikimimi bu alanda geliştirerek, toplumsal fayda sağlayabilecek bir çalışmaya dönüştürmek benim için son derece değerli olmuştur.

Tez sürecim boyunca bilgi, sabır ve desteğini esirgemeyen değerli danışmanım Dr.Öğr.Üyesi HADI POURMOUSA’ya en içten teşekkürlerimi sunarım. Çalışmamın her aşamasında rehberliği, bilimsel bakış açısı ve yapıcı yönlendirmeleriyle bana yol göstermiştir.

Ayrıca, eğitim hayatım boyunca bana koşulsuz destek veren sevgili anneme, babama ve yakın çevreme teşekkür ederim. Onların manevi desteği, bu süreci daha güçlü ve kararlı bir şekilde tamamlamamı sağlamıştır.



Giriş

* 1. Yapay Sinir Ağları Kavramı

Beyin, tıp dilinde “cerebrum” olarak adlandırılır. Bu organ, iki yarı küreden (sağ ve sol) zarla örtülü bir şekilde kafatasında bulunur. Beyindeki iki yarı kürenin farklı işlevleri vardır. Sağ yarı küre, görsel ve işitsel algıda güçlüdür; bilgileri parça parça alıp işlemek yerine bir bütün olarak alan ve işleyen kısımdır. Sol yarı küre ise sözel bilgilerin işlenmesinde güçlüdür; bilgileri parça parça işleyen ve odak noktasının sürdürülmesinde önemli rol oynayan kısımdır. Bu iki bölümün birbirleriyle iletişim kurması önem arz eder. Bu iletişimi ise, yaklaşık 250 milyon sinir lifinden oluşan “corpus callosum” adı verilen yapı sağlamaktadır (Ersoy ve Karal, 2012).

“Yapay sinir ağları, insan beyninin sinir hücresinin çalışma şeklini taklit ederek, sistemlere öğrenme, genelleme ve hatırlama özelliklerinin kazandırılması” (Saraç, 2004) şeklinde ifade edilir. Dr. Robert Hecht-Nielsen’e göre, yapay sinir ağı harici girdilere dinamik olarak yanıt vererek bilgiyi işleyen, parçalar halinde basit unsurlardan oluşan bir bilgi işlem sistemidir. Bu tanıma yakın bir başka tanım da Teuvo Kohonen’e aittir. Teuvo Kohonen’e göre, yapay sinir ağları, gerçek dünyadaki nesnelerle biyolojik sinir sistemine benzer bir şekilde etkileşime giren, paralel olarak bağlanmış birçok basit öğeden oluşan hiyerarşik organizasyonlardır (Bayır, 2016). Yapay sinir ağı ve beyin sinir ağının karşılıklı tanımları Tablo 1’de gösterilmiştir.

**A screenshot of a computer

AI-generated content may be incorrect.**

Tablo 1: Yapay Sinir Ağı ve Beyin Sinir Ağının Karşılıklı Tanımlamaları.

***Kaynak****: Öztürk ve Şahin, (2018).*

1.2. Yapay Sinir Ağının Bileşenleri

* + 1. A white circle with black letters and a white circle with black letters

       AI-generated content may be incorrect. Nöronlar

Şekil 1: Yapay Sinir Ağına Ait Bir Nöron/Işlem Birimi.

***Kaynak:*** *URL1*

Yapay sinir ağına ait nöron modeli, biyolojik nöronlardan ilham alınarak elde edilmiş bir işlem modelidir. Görseli Şekil 2’de verilmiştir. Nöron, kendisine gelen sinyalleri önceden belirlenmiş ağırlıklarla çarpar, bu sonuca yanlılık (bias) ekleyerek aktivasyon fonksiyonuna gönderir. Aktivasyon fonksiyonu ise kendisine gelen değeri işleyerek anlamlı bir çıktı oluşturur (Goodfellow ve diğ., 2016). Kısaca, nöron, kendisine verilen girdileri toplayan ve bu toplamın eşik değerini aşması durumunda çıktı üreten bir işlem birimidir (Ersoy ve Karal, 2012).

Şekil 1’de bir nörona N adet giriş (X1, X2, …, Xn) uygulanmaktadır. Her bir giriş, ilgili ağırlık (w1, w2, …, wn) ile çarpılır ve bu çarpımlar toplanarak toplam giriş değeri hesaplanır. Bu toplam, aktivasyon fonksiyonuna uygulanarak nöronun çıktısı (y) elde edilir. Şekil 1’de bir nörona N adet giriş (X1, X2, …, Xn) uygulanmaktadır. Her bir giriş, ilgili ağırlık (w1, w2, …, wn) ile çarpılır ve bu çarpımlar toplanarak toplam giriş değeri hesaplanır. Bu toplam, aktivasyon fonksiyonuna uygulanarak nöronun çıktısı (y) elde edilir.

* + 1. Giriş Katmanı

Dış ortamdan gelen bilgiler ham bir şekilde bu katmanda tutulur. Bazı araştırmacılar burada herhangi bir işlem yapılmadığı için bu katmanın ağların katman sayısına dahil olmadığını savunurlar (Bayır, 2016). Giriş katmanındaki her bir nöron, veri setindeki her bir özelliğin temsilidir (Koç ve diğ., 2018). Bu katmandan gelen bilgiler, ağırlıklarla çarpılarak gizli katmana iletilir (Öztemel, 2003).

* + 1. Gizli Katman

Bu katmanlar, girdi katmanlarından gelen bilginin işlenerek çıktıya dönüştürüldüğü katmanlardır. Gizli katman sayısı, modelin çözeceği probleme göre değişkenlik gösterir (Öztemel, 2003). Bu katmanlar, ağın öğrenme kapasitesini arttıran katmanlardır. Daha fazla katman, ağın öğrenme kapasitesini arttırır; ancak bunun da maliyet ve aşırı öğrenme (overfitting) dezavantajı vardır (Ersoy ve Karal, 2012).

* + 1. Çıkış Katmanı

Çıkış katmanı, yapay sinir ağlarının son katmanıdır; bu katman, belirlenen amaca göre nihai sonuçları üretir (Walczak, 2018). Üretilen çıktılar dış ortama ya da başka bir ağ modeline iletilir. Çıktı katmanı da girdi katmanı gibi, bir modelde yalnızca bir tane bulunur (Bayır, 2006). Çıkış katmanındaki nöronlar, probleme bağlı olarak uygun bir aktivasyon fonksiyonu kullanır (Walczak, 2018).

* + 1. Bağlantı Ağırlıkları

Bağlantı ağırlıkları nöronlar arasındaki ilişkinin etkisini belirleyen önemli parametrelerdir. Ağırlıklar eğitim sırasında konfigüre edilerek girdiler ile çıktılar arasındaki ilişkiyi modeller. Bu bağlantı ağırlıkları aslında modelin öğrendiği bilgileri depolayarak benzer girdilere doğru çıktılar üretilmesini sağlar (Arı ve Berberler, 2017). Biyolojik sinir hücresindeki sinapslara denk gelir (Öztürk ve Şahin, 2018). Aslında ağın eğitilmesi doğru ağırlıkların bulunması demektir. Bağlantı ağırlıkları en başta rastgele belirlenir; train veri setindeki verilerin hepsi için doğru çıktılar üretilene kadar bu ağırlıklar değişmeye devam eder. Ağ üzerindeki bu ağırlıkların her birinin ne olduğu ve neden olduğu bilinmediği için “kara kutu” yakıştırması yapılmıştır (Öztemel, 2003). Ağırlıkların değişmesi çıktıların değişmesine yol açar. Bu ağırlıkların pozitif olması bir nöronun diğerini uyarıcı (eksitatör) etkiyle etkilemesi; negatif olması ise baskılayıcı (inhibitör) etkiyle etkilemesi anlamına gelir (Okumuş, 2025).

![A diagram of a network

AI-generated content may be incorrect.]()

Şekil 2: Yapay Sinir Ağı Bileşenleri

***Kaynak:*** *URL6*

* + 1. Aktivasyon fonksiyonları

Aktivasyon fonksiyonlarının bir diğer adı transfer fonksiyonlarıdır. Aktivasyon fonksiyonları genellikle doğrusal değildir; bu nedenle kompleks ve farklı problemlere çözüm uygulanmasını sağlar (Bayır, 2006). Aktivasyon fonksiyonları, ağırlıklar kadar önemlidir; bu nedenle çözülen probleme göre uygun aktivasyon fonksiyonu seçilmesi gerekir (Öztürk ve Şahin, 2018). Aktivasyon fonksiyonunu önemli kılan sebeplerden biri, nöronların değer sınırlarını bilmemesi ve bir hücrede (+∞, -∞) değer bulunabilmesidir. Bu sebeple nöronun bağlantıları, nöronun aktifliğine aktivasyon fonksiyonu ile karar verir (Ser ve Bati, 2019). Altı adet aktivasyon fonksiyonu Tablo 2’de gösterilmiştir.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Tablo 2:Aktivasyon Fonksiyonları

***Kaynak:*** *URL2*

* 1. Yapay Sinir Ağı Mimarisi
     1. Tek Katmanlı Ağ

Tek katmanlı ağ, 1958 yılında F. Rosenblatt tarafından ortaya atılmıştır (Bayır, 2006). Yapay sinir ağları mimarileri arasındaki en basit yapı, tek katmanlı ağ diyebiliriz. Bu yapıda gizli katman yer almaz; sadece girdi katmanı ve çıktı katmanı vardır. Bu katmandaki çıktı fonksiyonu doğrusaldır; bu nedenle sadece doğrusal olarak ayrılabilir problemleri çözebilir (Minsky ve Papert, 2017). Çıktı, 1 veya –1 değerini alır; 1 birinci sınıfı, –1 diğer sınıfı temsil eder (Öztürk ve Şahin, 2018). Şekil 2’de tek katmanlı ağın görseli verilmiştir.

metin, diyagram, ekran görüntüsü, daire içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 3: Tek Katmanlı Ağ

**Kaynak:** Öztemel, (2003).

* + 1. Çok Katmanlı Ağ

Çok katmanlı ağ, doğrusal olmayan fonksiyonlara sahip olup yalnızca girdi ve çıktı katmanından oluşmaz, aynı zamanda gizli katmanları da içerir. Bu yapının ortaya çıkmasının nedeni tek katmanlı yapının yetersiz kalmasıdır (Öztürk ve Şahin, 2018). Çok katmanlı ağ mimarisi, tek katmanlı yapının XOR problemi karşısında yetersiz kalması nedeniyle bazı araştırmacılar tarafından geliştirilmiştir (Bayır, 2006). Bu mimaride çok katmanlı soyutlamalar yapılarak doğrusal olmayan modeller elde edilir. Nöronlar arasındaki ağırlıklar, geri yayılım (backpropagation) ve gradyan iniş (gradient descent) yöntemleriyle güncellenir; bu sayede model örüntüleri daha doğru öğrenir (LeCun ve diğ., 2015).

diyagram, çizgi, origami içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 4: Çok Katmanlı Yapay Sinir Ağı Model

***Kaynak:*** *Ersoy ve Karal, (2012).*

* + 1. Tekrarlayan Sinir Ağları (RNN)

Tekrarlayan sinir ağları (Recurrent Neural Networks), farklı uzunluklardaki dizi girdisini işleyen, ileri beslemeli sinir ağının uzantısıdır. Tekrarlayan sinir ağları, her zaman adımından önceki duruma dayalı olarak gizli durumunu günceller; bu sebeple bir hafızaya sahiptir (Chung ve diğ., 2014). Gizli durum, bellek görevi görerek geçmiş bilgileri saklar (Das ve diğ., 2023). Tekrarlayan sinir ağları, sıralı (ordinal) verileri işlemek için sunulmuştur. Bir aktivasyon fonksiyonu ile zaman adımı t, ağ girişi x\_t ve bir önceki adımdan gelen gizli durum h\_{t-1}, birleştirilerek elde edilir; bu formül şu şekilde ifade edilir:

ℎ*t* = 𝜎(𝑊*hh***⋅** ℎ *t-1*+ 𝑊*xh*⋅ 𝑥*t* + 𝑏*h* )

* + - * σ: Aktivasyon fonksiyonu
      * W*hh*: Önceki gizli duruma ait ağırlık matrisi
      * W*xh*​: Girişe ait ağırlık matrisi
      * b*h*​: Bias terimi

Bu işleyişe “tekrarlayan süreç” denir (Lipton ve diğ.,2015). Şekil 4’te RNN yapısı gösterilmiştir.

A diagram of a diagram

AI-generated content may be incorrect.

Şekil 5: (Solda) RNN Için Devre Şeması Gösterilmiştir. (Sağda) Aynı RNN Için Açılmış (Unfolded) Hesaplama Şeması Gösterilmiştir.

***Kaynak:*** *Das ve diğ., (2023)*

* + 1. Uzun Kısa Süreli Bellek (LSTM)

Uzun süreli kısa bellek RNN’lerde yaşanan gradyan kaybolmasına çözüm olarak 1997’de Sepp Hochreiter ve Jürgen Schmidhuber tarafından sunulmuştur. RNN’ler her zaman bir önceki duruma göre ilişkileri tutarken uzun süreli bilgileri saklamakta eksik kalır ve bu sebeple gradyan patlaması veya gradyan kaybolması ortaya çıkar. LSTM, geleneksel RNN’lerin aksine kendi içinde bir döngüye sahiptir; bu sayede gradyanlar daha uzun süre saklanabilir. LSTM mimarisi, giriş kapısı, unutma kapısı ve çıkış kapısı olmak üzere üç ana kapıdan oluşmaktadır.

* Giriş kapısı hücreye girmesi gereken girdileri seçer ve ne kadar süre tutulacağına karar verir.
* Unutma kapısı hücreden hangi bilginin silineceğine karar verir. Aktivasyon fonksiyonu olarak sigmoid kullanılır çıkan değer 1 yakınsa bilgiyi korur 0 yakınsa bilgiyi siler.
* Çıkış kapısı hücreden hangi bilginin sonraki katmana iletileceğine karar verir.

Bu kapılar sayesinde LSTM uzun dönem ilişkileri saklar (Das ve diğ., 2023). Şekil 5’te LSTM yapısı gösterilmiştir.

A diagram of a flowchart

AI-generated content may be incorrect.

Şekil 6 : LSTM Yapısı Hücre Durumu **c t**,Gizli Durum **h t**, Girdi **x t** ve Çıktı **o t**

***Kaynak:*** *Das ve diğ, (2023).*

* + 1. Evrişimsel Sinir Ağları

Evrişimsel sinir ağları, görüntülerden yerel özellikleri otomatik bir şekilde çıkarabilen, çok katmanlı ve ileri beslemeli bir yapay sinir ağı mimarisidir (Akın ve Şahin, 2024). Evrişimsel sinir ağlarının çok katmanlı bir yapıya sahip olduğunu söyledik; ancak çok katmanlı bir yapıya evrişimsel sinir ağı diyebilmek için bu katmanlardan en az birinde genel matris çarpımı yerine evrişim (convolution) adı verilen özel bir matematiksel işlemin kullanılmış olması gerekmektedir (Büyükavcılar, 2023). Evrişimsel sinir ağları, kompleks görüntü örüntülerini etkili bir şekilde öğrenmesini sağlayan konvolüsyon, düzleştirme, havuzlama ve tam bağlantı bileşenlerine sahiptir. Bu bileşenlerin tasarımı ise elde edilmek istenen görüntünün niteliğine göre farklılık gösterir (Akın ve Şahin, 2024). Şekil 6’da klasik bir evrişimsel sinir ağı yapısı gösterilmiştir.

metin, meyve içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 7 : Klasik Bir CNN Mimarisi

***Kaynak:*** *Fırıldak ve Talu, (2019).*

* + - 1. Evrişim katmanı

Yapay sinir ağlarında bulunmayan bu katman, evrişimsel sinir ağlarının temel yapı taşıdır (Arı ve Hanbay, 2019). Evrişim katmanı, girilen görüntülere filtreler (kernel) uygular; bu uygulamayla görüntüdeki yerel özellikleri (kenarlar, köşeler, dokular vb.) otomatik bir şekilde çıkarır. Uygulanan bu filtreler, görüntülerin belirli bölgelerindeki önemli bilgileri vurgular ve bu sayede özellik haritaları (feature maps) üretir. Elde edilmek istenen özellik haritalarının boyutu ve çözünürlüğü, adım (stride), çekirdek boyutu ve sıfır doldurma (padding) hiper parametreleri ile kontrol edilir (Krizhevsky ve diğ., 2012). Şekil 7’de evrişim işlemi örneği gösterilmiştir.

metin, yazı tipi, diyagram, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 8 : İki Boyutlu Bir Tensör Üzerindeki Evrişim İşlemine Bir Örnek.

***Kaynak:*** *Al-Dulaimi ve Kurnaz, (2024).*

* + - 1. Havuzlama katmanı

Havuzlama katmanı, işlemleri hızlandırmak için konvolüsyon katmanı sonrası elde edilen görüntü matrisinin boyutunu küçültür. Ortalama havuzlama ve maksimum havuzlama olmak üzere iki yöntem vardır. Maksimum havuzlama yönteminde seçilen matris içindeki en yüksek değer alınırken, ortalama havuzlamada matrisin ortalaması hesaplanır (Akın ve Şahin, 2024). Havuzlama katmanı, aşırı öğrenme (overfitting) riskini düşürerek daha genelleyici bir yapı elde edilmesini sağlar (Krizhevsky ve diğ., 2012). Bazı mimarilerde havuzlama katmanı olmayabilir; bunun sebebi havuzlama katmanının isteğe bağlı olmasıdır (İnik ve Ülker, 2017). Şekil 8’de maksimum havuzlama ve ortalama havuzlama yöntemleri gösterilmiştir.

metin, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 9 : 2x2’lik Maksimum Havuzlama ve Ortalama Havuzlama Yöntemi

***Kaynak:*** *URL 3*

* + - 1. Tam bağlantı katmanı

Eğitim sırasında, tam bağlantılı katmandaki her nöron girdilerin ağırlıklı bir toplamını hesaplar; bu toplam, ağın çıktısını üretmek için belirli bir aktivasyon fonksiyonuna eklenir. Bu süreç, geri yayılım algoritması ile optimize edilir ve modelin genel performansını doğrudan etkiler (Goodfellow ve diğ., 2016).

Bu katman, özellikle görüntü tanıma ve nesne sınıflandırma gibi uygulamalarda derin öğrenme mimarisinin son aşaması olarak önemli bir rol oynar. Örneğin, konvolüsyonel sinir ağlarında çıkarılan özellikler tam bağlantılı katmanda işlenerek hangi nesneye ya da sınıfa ait oldukları belirlenir. Bu aşamada elde edilen sütundaki en yüksek değerlerin ortalaması doğruluk oranını, en düşük değerlerin ortalaması ise hata oranını yansıtarak modelin başarısının ölçülmesine yardımcı olur (LeCun ve diğ., 2015). Şekil 9’da tam bağlantı katmanı yapısı gösterilmiştir.

A diagram of a machine learning

AI-generated content may be incorrect.

Şekil 10: Tam Bağlantı Katmanı Şeması

***Kaynak:*** *URL 5*

* + - 1. Çekişmeli üretken ağlar (GAN)

Çekişmeli üretken ağlar, 2014 yılında Goodfellow ve arkadaşları tarafından ortaya atılmış, ayrımcı (Discriminator) ve üretici (Generator) ağlarından oluşan bir mimaridir. Üretici ağ, girdilerden yeni veriler oluşturarak ayrımcıyı kandırmaya çalışırken; ayrımcı ağ, bu verilerin gerçek mi sahte mi olduğunu ayırt etmeye çalışır. Bu iki ağın birbirleriyle oynadığı oyun sayesinde, gerçeğe yakın görüntüler elde edilir (Goodfellow ve diğ., 2014).

Çekişmeli üretken ağlar, görüntü üretimi, veri artırma, ses sentezi gibi alanlarda uygulanmış ve ileri nesil GAN'ların (DCGAN, CycleGAN, Conditional GAN) temelini oluşturmuştur. Şekil 9’da GAN yapısı gösterilmiştir.

A diagram of a computer generated model

AI-generated content may be incorrect.

Şekil 11: Çekişmeli Üretken Ağ Yapısı.

***Kaynak:*** *URL 4*

* 1. Deepfake Nedir?

Son zamanlarda makine öğrenmesi, yapay zekâ ve derin öğrenme alanlarındaki gelişmeler, hayatımızın birçok kısmına etki etmeye başlamıştır. Bu alanlardaki büyük ilerlemeler, görsel manipülasyonları daha gerçekçi hale getirmiştir. Bu gelişmelere paralel olarak, deepfake teknolojisi sahte görüntüler ve videolar oluşturma konusunda önemli ilerlemeler kaydetmiştir (Heo ve diğ., 2021).

Deepfake terimi, derin öğrenme (deep learning) ve sahte (fake) kelimelerinden türetilmiştir. Yapılan işlem, hedef bir kişinin yüz bölgesinin, kaynak bir kişinin yüz bölgesine aktarılmasıdır (Nguyen ve diğ., 2022). Deepfake teknolojisinin temelleri, 2014 yılında ortaya atılan özel bir sinir ağı mimarisi olan Generative Adversarial Networks (GAN) ile atılmıştır. "Deepfake" terimi ise 2017 yılında, Reddit’te "deepfake" adlı bir hesabın ünlülerin pornografik görüntülerini oluşturmasıyla ortaya çıkmış ve popülerlik kazanmıştır (Berk, 2020).

Deepfake teknolojisi özellikle siyasetçileri ve ünlüleri çok fazla etkilemiştir. Bunun başlıca sebebi, bu kişilerin internette çok sayıda görüntüsünün bulunması ve bu görüntüler üzerinden propaganda yapılmak istenmesidir. Deepfake teknolojisiyle birçok alan olumsuz etkilenebilir: dezenformasyon haberler üretilerek finansal piyasalar zarara uğratılabilir, askeri stratejistlerin büyük yanılgılara düşmesine neden olunabilir, uydu görüntüleri üzerinde oynanarak var olmayan objeler ve güzergâhlar görüntülere eklenebilir. Bu sebeplerden dolayı, deepfake teknolojisi sadece bireyleri değil, toplumları ve ülkeleri de büyük ölçüde etkileyebilmektedir (Nguyen ve diğ., 2022).

Bununla birlikte, deepfake teknolojisi yalnızca olumsuz işlerde kullanılmamakta; film, dizi, sosyal medya ve sanal dünyada olumlu katkılar da sunmaktadır. Film ve dizi yapımlarında, aktörlerin yüzlerinin dijital olarak değiştirilmesi, efektlerin daha gerçekçi hale getirilmesi ve hayatını kaybetmiş ünlü aktörlerin yeniden canlandırılması gibi uygulamalarda kullanılmaktadır. Sanal dünyada ise, kullanıcıların daha gerçekçi avatarlarının oluşturulmasına katkı sağlamaktadır (Dinçer ve diğ., 2024).

öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, diyagram, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 12 : (Sol) Google Arama Motorunun ‘'Deepfake'' Anahtar Kelimesini Içeren Web Sayfa Sayısı Sonuçları. (Sağ) Google Arama Motorunun “Deepfake” Ile Ilgili Videoları Içeren Web Sayfa Sayısı Sonuçları.

***Kaynak:*** *Rana ve diğ., (2022).*

Deepfake'in hayatımıza hızlı bir şekilde girip, teknolojiyi teknik olarak bilmeyen birinin bile basit bir donanımla gerçekçi deepfake videoları otomatik olarak oluşturmasının mümkün olduğu anlaşıldığında, otomatik tespit yöntemleri oluşturma ihtiyacı ortaya çıkmaktadır (Dolhansky ve diğ., 2020). Deepfake'in hızla yükselmesi, örnek Şekil 10'daki tabloya bakılarak, uzmanları ve akademisyenleri harekete geçirmiştir (Rana ve diğ., 2022).



LİTERATÜR TARAMASI

Son yıllarda, deepfake teknolojisinin farklı sektörlerde yoğun bir şekilde kullanılmaya başlanması, beraberinde önemli güvenlik ve etik sorunlarını da gündeme getirmiştir. Özellikle medya, eğlence, siyaset ve sosyal medya gibi alanlarda deepfake içeriklerin yaygınlaşması, bu tür sahte içeriklerin tespiti için yapılan akademik ve endüstriyel çalışmalarda dikkate değer bir artışa yol açmıştır. Deepfake içeriklerin potansiyel zararları, yalnızca bireysel mahremiyet ihlalleriyle sınırlı kalmayıp; kamu düzeni, siyasal istikrar ve toplumsal güven açısından da ciddi tehditler oluşturmaktadır.

Bu bağlamda, bu bölümde deepfake tespitine yönelik temel yaklaşımlar, deepfake tespitine özel olarak oluşturulmuş veri setleri ve deepfake içeriklerin üretiminde kullanılan ileri düzey mimariler ayrıntılı bir şekilde ele alınmaktadır. Öncelikle, deepfake tespit yöntemlerinin genel prensipleri açıklanacak; ardından, tespit çalışmalarında yaygın olarak kullanılan benchmark veri setleri tanıtılacak ve son olarak, görüntü üretiminde kullanılan önemli derin öğrenme mimarilerine değinilecektir.

* 1. Görüntü üretimi üzerine yapılan çalışmalar

Radford, Metz ve Chintala’nın çalışması, Deep Convolutional Generative Adversarial Networks (DCGAN) mimarisini tanıtarak, gözetimsiz temsil öğrenmesi alanında önemli bir adım atmaktadır. Bu çalışmada, denetimli öğrenme ve denetimsiz öğrenme için CNN’lerin başarısı arasındaki boşluğu doldurmak amaçlanmıştır. Çalışmada, geleneksel GAN mimarisinin derin evrişimsel sinir ağlarıyla entegrasyonu sağlanarak daha stabil ve etkili bir eğitim süreci elde edilmiştir. Önerilen mimaride strided convolution, ReLU aktivasyonu ve batch normalization kullanılması, modelin üreteç (generator) ve ayırt edici (discriminator) ağlarının performansını artırmıştır. Böylece, üretilen görüntülerde anlamlı ve hiyerarşik özellikler öğrenilmiş ve bu özelliklerin daha sonra farklı görevlerde kullanılabileceği gösterilmiştir (Radford ve diğ., 2015).

Zhang ve arkadaşları tarafından yapılan çalışmada, dikkat mekanizmasına dayalı ve uzun menzilli bağımlılıkları modelleyebilen, adına Self-Attention Generative Adversarial Network (SAGAN) verdikleri bir GAN mimarisi önerilmiştir. SAGAN, geleneksel GAN’lardan farklı olarak, tüm özellik konumlarından alınan ipuçları ile daha geniş bağlamı yakalayabilmektedir. Bu sayede, uzak bölgeler arasındaki tutarlılıklar sağlanarak görüntülerdeki detaylar oluşturulmaktadır. Ayrıca, bu modelin koşullandırılmasının GAN performansını etkilediği gözlemlenmiş ve buna dayanarak üreticiye spektral normalizasyon uygulanmıştır. Önerilen SAGAN, ImageNet veri seti üzerinde mevcut en iyi inception skorunu 15,72 birim artırırken, Fréchet Inception Distance değerini ise 27,62'den 18,65'e düşürerek önemli performans artışları göstermiştir (Zhang ve diğ., 2019).

* 1. Deepfake tespiti için oluşturulmuş veri setleri

Deepfake görüntü tespiti için sınırlı sayıda veri seti bulunmaktadır. Bu veri setleri, gerçek ve sahte görüntüler barındırmaktadır. Veri setlerinde bulunan görüntü sayısı, görüntülerin oluşturulma şekli, ışık açısı, görüntü kalitesi vb. özellikler, oluşturulacak modelin performansı açısından önem arz etmektedir (Kütükçü ve Polat, 2024). Bu kısımda, deepfake tespiti için kullanılan altı farklı popüler veri setinden bahsedilmektedir.

* + 1. Celeb-DF

Celeb-DF, Li ve arkadaşları tarafından oluşturulmuş zorlu ve büyük ölçekli bir veri setidir. Mevcut deepfake veri setlerine göre daha gerçekçi ve yüksek kaliteli görüntüler içermektedir. Li ve arkadaşları, YouTube’dan topladıkları ünlü kişilere ait gerçek videolara uyguladıkları zorlu bir teknikle, manipülasyon izlerinin az olduğu görüntüler elde etmiştir. Elde edilen görüntülerin derlenmesiyle oluşturulan veri seti, 590 gerçek ve 5639 deepfake video içermektedir (Li ve diğ., 2020).

* + 1. FaceForensics++

FaceForensics++, Rössler ve arkadaşları tarafından tanıtılan, zengin ve çeşitli kaynaklar sunan bir veri setidir. Bu veri seti, gerçek yüz videoları ile dört farklı manipülasyon tekniği (DeepFakes, Face2Face, FaceSwap ve NeuralTextures) kullanılarak oluşturulmuş görüntülerden oluşmaktadır. Buradaki her bir teknik için farklı kalitede versiyonlar sunulmakta, böylece değerlendirme tekniklerine çeşitlilik katılmaktadır. FaceForensics++ veri seti, 1000 gerçek ve 4000 sahte videodan oluşmaktadır (Rössler ve diğ., 2019).

* + 1. DFDC (Deepfake Detection Challenge)

Kaggle yarışmalarında kullanılmak üzere ve mevcut deepfake veri setlerinin çeşitlilik ve ölçek açısından yetersiz kalması sebebiyle Dolhansky ve arkadaşları tarafından yayınlanmış bir veri setidir. Gerçek ve deepfake videolar barındıran bu veri seti, çeşitli yaş ve etnik gruplardan oluşan 3.426 aktörle hazırlanmış olup, 100.000’den fazla görüntü içermektedir. Veri setinin boyutu yaklaşık 470 GB’tır (Dolhansky ve diğ., 2020).

* + 1. DFD (Deep Fakes Dataset)

Google Research ve Jigsaw tarafından, deepfake tespit araştırmalarına katkıda bulunmak amacıyla oluşturulmuş bir veri setidir. DFD veri seti, farklı çekim koşulları ve çeşitli sahnelerle çekilmiş gerçek ve sahte videolardan oluşmaktadır. Bu veri setini oluşturmak amacıyla, farklı yaş ve etnik gruptan oluşan 28 aktör kullanılmış olup, 364 gerçek ve 3.068 sahte videodan oluşmaktadır (Google AI Blog, 2019).

* + 1. UADFV

New York Eyaleti Üniversitesi Albany tarafından, sınırlı sayıda kimlikle oluşturulmuş erken dönem video veri setlerinden biridir. UADFV veri setinde 49 gerçek ve 49 sahte video olmak üzere toplam 98 video bulunmaktadır. Her bir video yaklaşık 11 saniye olmakla birlikte, toplamda 32.752 frame içermektedir. UADFV, deepfake tespiti çalışmalarında özellikle yüz manipülasyonlarının yarattığı bozuklukları inceleyen yöntemlerin eğitilmesinde önemli bir referans veri kaynağı sunarken, sınırlı sayıda kimlik içermesi nedeniyle kimliklere aşırı uyum sağlanmasını önlemektedir (Yang ve diğ., 2019).

* + 1. DeepfakeTIMIT

Idiap Araştırma Enstitüsü tarafından oluşturulmuş, gerçek ve sahte videolar içeren bir veri setidir. Sahte videolar, 16 benzer görünümlü kişi çiftine uygulanan açık kaynaklı GAN tabanlı yüz değiştirme yöntemi kullanılarak üretilmiştir. Her çift için 128x128 piksel çözünürlüğünde yüksek kaliteli (HQ) ve 64x64 piksel çözünürlüğünde düşük kaliteli (LQ) olmak üzere iki farklı model eğitilmiştir (Korshunov ve Marcel, 2018).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Veri Seti** | **Gerçek Video Sayısı** | **Sahte Video Sayısı** | **Toplam** | **Yöntem** |
| Celeb-DF | 590 | 5.639 | 6.229 | Deepfake |
| FaceForensics++ | 1.000 | 4.000 | 5.000 | (DeepFakes, Face2Face, FaceSwap ve NeuralTexture |
| DFDC (Deepfake Detection Challenge) | **~23.000** | **~101.000** | ~123.000 | Belirtilmemiş |
| DFD (Deep Fakes Dataset) | 364 | 3.068 | 3.432 | FaceSwap |
| UADFV | 49 | 49 | 98 | FakeApp |
| DeepfakeTIMIT | 320 | 640 | 960 | Dynamic Time Warping (DTW) |

Tablo 3 : Deepfake Tespitinde Kullanılan Popüler Veri Setleri

* 1. Deepfake Tespiti Üzerine Yapılan Çalışmalar

Coccomini ve arkadaşları çalışmalarında hibrit bir yaklaşımı önermektedir. Bu hibrit yaklaşım, deepfake tespitinde global ve yerel özelliklerin etkin bir biçimde çıkarılabilmesi amacıyla EfficientNet ve Vision Transformers (ViT) mimarilerini birleştirmiştir. Çalışmada EfficientNet’in yüksek performanslı evrişimsel sinir ağı yapısı kullanılarak videolardan alınan karelerin ayrıntılı uzamsal özellikleri elde edilmiş, daha sonra bu özellikler Vision Transformers bileşeni aracılığıyla, dikkat mekanizmaları yardımıyla uzun menzilli ilişkiler ve küresel bağlam bilgisi harmanlanmıştır. Önerilen mimari, deepfake videolardaki tutarsızlıkları ve yerel anomalileri yakalamayı amaçlamaktadır. Bu çalışmada elde edilen en iyi model, DeepFake Detection Challenge (DFDC) veri seti üzerinde 0.951 AUC ve %88.0 F1 skoru elde etmiştir (Coccomini ve diğ., 2022).

Afchar ve arkadaşları, deepfake tespiti için hafif ve kompakt bir derin öğrenme mimarisi önermişlerdir. Bu mimariye MesoNet adını vermişlerdir. MesoNet, geleneksel derin ağlardan farklı olarak az katman sayısı ile düşük hesap maliyetli deepfake tespiti için uygun bir çözüm sunmaktadır. Afchar ve arkadaşlarının çalışmasında, medyaların mezoskopik özelliklerine (orta ölçekli uzamsal desenler ve anomaliler) odaklanılarak yüzdeki manipülasyonların belirlenmesi amaçlanmıştır. Bu doğrultuda iki model üretilmiştir: MesoInception-4 ve Meso-4, her iki model de deepfake tespit etmek için optimize edilmiştir. Model, FaceForensics++ veri seti ile değerlendirilip, Face2Face yöntemi ile üretilen deepfake videolarda %95, deepfake tekniği ile oluşturulan videolarda ise %98 doğruluk oranı elde edilmiştir (Afchar ve diğ., 2018).

Wodajo ve Atnafu tarafından yapılan çalışmada, Convolutional Vision Transformer (CVT) tabanlı bir deepfake video tespit yöntemi önerilmektedir. Bu yaklaşım, CNN’lerin yerel özellik çıkarma yeteneği ile Transformer tabanlı modellerin uzun menzilli bağıntıları öğrenme kapasitesini birleştirmektedir. Bu sayede videolarda görülen manipülasyon izlerinin tespiti sağlanmıştır. Çalışmada, DeepFake Detection Challenge (DFDC) veri seti kullanılarak %91.5 doğruluk, 0.91 AUC ve 0.32 kayıp değeri elde edilmiştir (Wodajo ve Atnafu, 2021).

Nguyen ve arkadaşları, kapsül ağların gücünden yararlanarak "Capsule Forensics" adlı bir yöntem geliştirmişlerdir. Bu yöntem, geleneksel CNN’lerin uzamsal ilişkilerdeki sınırlı kapasitesine karşı, kapsül ağlarının çalışma mekanizmasındaki dinamik yönlendirme sayesinde hiyerarşik ve mekânsal ilişkileri daha etkili modelleyebilmesini ele alır. Bu ağın eğitiminde FaceForensics++ veri seti kullanılmış olup, DeepFake tespiti için %92.17, FaceSwap tespiti için %92.79, Face2Face tespiti için %90.36 doğruluk oranı elde edilmiştir (Nguyen ve diğ., 2019).

Li ve Lyu tarafından yapılan çalışmada, deepfake tespiti için yüz deformasyonu artefaktlarının analizi ele alınmıştır. Deepfake video üretiminde, kaynak yüz hedef yüze yerleştirilirken kaynak ve hedef yüzlerin boyut farkı ve perspektiflere sahip olması nedeniyle bazı uyumsuzluklar ortaya çıkmaktadır. Bu çalışmada hedef, bu oluşan geometrik ve renk uyumsuzluklarını tespit ederek deepfake videoları ayırt edebilmektir. Evrişimsel sinir ağı tabanlı bu model, düşük bulanık alanları ve düşük çözünürlüklü bölgeleri tespit ederek deepfake görüntülerini tespit etmiştir. Araştırmacılar bu çalışma için kendi veri setini oluşturmuş, oluşturulan bu veri setinde karmaşık deepfake üretim yöntemi yerine basit işleme yöntemi (yüz hizalama, Gauss bulanıklığı, affine dönüşüm simülasyonu) kullanılmıştır. Model yüksek doğruluk oranları vermiştir, ancak yöntemin daha karmaşık deepfake üretim tekniklerine karşı geliştirilmesi gerektiği vurgulanmıştır (Li ve Lyu, 2018).

Li, Chang ve Lyu tarafından yapılan çalışmada, AI ile üretilen deepfake videoların tespiti için göz kırpma davranışlarına odaklanılmıştır. Çalışmadaki varsayım, deepfake üretim modellerinin eğitim verilerinde göz kırpma örneklerinin az olması nedeniyle, üretilen videoların da insanların doğal göz kırpma davranışını tam olarak sağlayamamasıdır. Bu çalışmada yüz bölgelerinden göz alanları çıkarılmış ve göz kırpma hareketleri analiz edilerek gerçek videolar ve sahte videolar arasındaki farklar belirlenmeye çalışılmıştır. Çalışmada Evrişimsel Sinir Ağı (CNN) ve Tekrarlayan Sinir Ağı (RNN) kombinasyonuyla oluşturulan Uzun Süreli Tekrarlayan Evrişimli Sinir Ağı (LRCN) kullanılarak 0.99 AUC değeri elde edilmiştir (Li ve diğ., 2018).

Zhang, Ni ve Nie tarafından yapılan çalışmada, deepfake tespitinde karşılaşılan Alan Kayma problemine odaklanan Alan Kayma Modelleme (DSM) çerçevesi önermişlerdir. Bu çalışmada, deepfake örneklerine karşı genelleme yeteneğini artırmak için eğitim ve test verileri arasında oluşan dağılım farklılıkları modellemiştir. Önerilen model, FaceForensics++, Celeb-DF ve DeepfakeTIMIT veri setleriyle eğitilmiş ve kayıp değerini minimize edebilmek için özel bir kayıp fonksiyonu olan Maksimum Ortalama Uyumsuzluğu (MMD) kullanılmıştır. FaceForensics++ üzerinde 0.98 AUC, Celeb-DF üzerinde ise 0.92 AUC değeri elde edilmiştir (Zhang ve diğ., 2025).

Güera ve arkadaşları, video üzerindeki manipülasyonları tespit etmek için piksel seviyesinde çalışılmasına gerek duyulmadan, videonun akış özelliklerini kullanarak deepfake tespiti yöntemi önermektedir. Bu yaklaşım, video sıkıştırma sürecinde ortaya çıkan manipülasyon izlerini ve istatistiksel özellikleri yakalamayı amaçlamaktadır ve piksel bilgisine dayalı yöntemlerden farklı bir süreç izlemektedir. Bu çalışmada Media Forensic Challenge (MFC) veri seti kullanılmıştır. Sonuçlar, az sayıda etiketli veride bile destek vektör makinesi (SVM) ve basit bir rastgele orman ensemble’ı kullanılarak oluşturulan modellerde yüksek skorlar elde edilmiştir. Örneğin, destek vektör makinesi modeli ile 0.91 F1 ve 0.951 AUC değeri elde edilmiştir (Güera ve diğ., 2019).

Gu ve arkadaşları tarafından yapılan çalışmada, deepfake tespit problemini, sahte videolarda ortaya çıkan zamansal ve mekânsal tutarsızlıkları modelleyerek çözmeyi hedefleyen bir yöntem önermektedir. Önerilen yöntem “Mekânsal-Zamansal Tutarsızlık Öğrenimi” (STIL) olarak formüle edilmiştir. Bu formül üç bileşenden oluşmaktadır. Mekânsal Tutarsızlık Modülü (SIM) her bir karede manipülasyon kaynaklı artefaktları tespit eder. Zamansal Tutarsızlık Modülü (TIM) kareleri hem dikey hem de yatay yönlerden inceleyerek kareler arasındaki farkların tespiti ile zamansal tutarsızlıkları ortaya çıkarır. Bilgi Tamamlayıcı Modül (ISM) SIM ve TIM’den elde edilen sonuçları birleştirerek daha kapsamlı ve güçlü mekânsal-zamansal temsil oluşturur. Çalışmada, **FaceForensics++**, **Celeb-DF**, **DFDC** ve **WildDeepfake** veri setleri kullanılmış olup her bir veri seti için yüksek değerler elde edilmiştir. Örneğin, **FaceForensics++** veri setinde kare bazlı değerlendirmede 0,99 AUC değeri elde edilirken, video bazlı değerlendirmede 0,98-0,99 aralığında AUC değeri elde edilmiştir (Gu ve diğ., 2021).

Bonettini ve arkadaşları tarafından yapılan çalışmada, yüzdeki manipülasyonları tespit etmek amacıyla farklı evrişimsel sinir ağlarının çıktılarını ensemble yöntemi ile birleştirilmesini önermişlerdir. Bu yöntem sayesinde, tek bir modelin yetersiz kaldığı yerlerde daha sağlam ve genelleyici bir tespit sistemi elde edilmiştir. Bu çalışmada **ImageNet** veri setinden %83 doğruluk oranı elde edilmiştir (Bonettini ve diğ., 2020).

De Lima ve arkadaşları tarafından yapılan çalışmada, deepfake tespiti için video tabanlı bir yöntem önerilmiştir. Bu yöntemde zamansal ve mekânsal bilgileri birlikte işleyen bir spatiotemporal konvolüsyonel sinir ağı (STCN) yaklaşımı ele alınmıştır. Çalışmadaki amaç, ardışık kareler arasındaki değişimleri tespit ederek sahte ve gerçek videolar arasındaki ince tutarsızlıkları ortaya çıkarmaktır. Deneysel değerlendirmeler bu yaklaşımın yalnızca mekânsal bilgiyi kullanan modellere göre daha yüksek doğruluk yeteneği sağladığını göstermektedir (De Lima ve diğ., 2020).

Coccomini ve arkadaşları tarafından yapılan bu çalışmada, deepfake tespiti için **Multi-Identity Size-Invariant Video Deepfake Detection** (MINTIME) adlı yeni bir yöntem önerilmiştir. Bu yöntem video içerisindeki yüzleri etkili bir şekilde işleyebilecek şekilde tasarlanmıştır. **MINTIME** tek yüz manipülasyonlarına ve aynı video içinde birden fazla yüz bulunması durumuna karşı dayanıklı bir tespit gerçekleştirir. Model, farklı ölçek ve kimliklerdeki yüzlerin varlığı durumunda ortaya çıkabilecek karışıklıkları minimize edecek şekilde, ölçek bağımsızlığı özelliklerini entegre etmektedir. Bu yöntem çapraz veri seti ve çapraz sahtecilik bağlamlarında önceki yöntemlere göre yüksek genelleme kabiliyeti sunmuştur (Coccomini ve diğ., 2024).

Jain ve arkadaşları tarafından yapılan bu çalışmada, veriye dayalı etiketlenmiş örnekler olmadan, sadece sentetik görüntüler kullanarak deepfake tespit metodolojisi önerilmiştir. Yöntem, sadece gerçek dünya verilerine bağımlı kalmadan, sentetik olarak oluşturulan görüntüler aracılığıyla öğrenmeyi hedefler. Bu yöntemde ayrıca gerçek verilerle de ince ayar yapıldığında genelleme yeteneği artmıştır. Deneysel değerlendirmede benchmark veri setleri üzerinde %93 doğruluk, 0,95 AUC ve 0,90 F1 skoru elde edilmiştir (Jain ve diğ., 2022).

Guarnera ve arkadaşları tarafından yapılan çalışmada, deepfake tespitinde ortaya çıkan evrişimsel izlerin analiz edilmesine dayalı bir yöntem önermişlerdir. Bu yöntem, deepfake üretim süreçlerinin, videolar üzerinde belirgin bir şekilde forensik izler (parmak izi) bıraktığı gözlemi üzerine kurulmuştur. Bu yöntemde **Expectation Maximization** (EM) algoritması kullanılarak izler üzerinden hesaplaması kolay ve hafif bir tespit sistemi geliştirilmiştir. Bu modelin eğitilmesi için **CelebA** ve 5 farklı **GAN (DWCT, STARGAN, ATTGAN, STYLEGAN, STYLEGAN2)** ile üretilmiş toplam 6 veri seti kullanılmıştır. Model, bu 6 farklı mimarideki kendine özgü izleri başarılı bir şekilde tespit etmiştir. Örneğin, **CelebA** vs **STYLEGAN2** karşılaştırmasında lineer **SVM** ile 4x4 boyutunda çekirdek kullanılarak %99,81 doğruluk elde edilmiştir (Guarnera ve diğ., 2020).

Zhao ve arkadaşları tarafından yapılan bu çalışmada, deepfake tespitinde mevcut yöntemlerinin çoğunun derin ağlarla çıkarılan global özellikler üzerinden gerçek-sahte ayrımını yaptığını ancak gerçek ve sahte görüntü arasındaki farklar genellikle ince ve lokal olduğu için mevcut bu çalışmaların optimal olmadığı düşünülmektedir ve buna uygun bir yöntem önerilmektedir. Bu yöntem üç temel bileşenden oluşmaktadır:

1. Farklı yerel bölgelere odaklanmak için çoklu mekânsal dikkat başlıkları
2. Dikkat haritaları tarafından yönlendirilen düşük seviyeli dokusal özellik ve yüksek seviyeli anlamsal özellikleri bir araya getirme.
3. Sığ özelliklerdeki ince yapaylıkları yakınlaştırmak için dokusal özellik geliştirme bloğu.

Ayrıca, bu ağın öğrenme zorluğunu gidermek için yeni bir bölgesel bağımsızlık kaybı ve dikkat güdümlü veri arttırma stratejisi sunulmuştur. Bu modelde Celeb-DF, FaceForensics++ ve DFDC veri setleri kullanılmıştır. FaceForensics++ veri seti üzerine uygulanan EfficientNet-B4 ağı ile %99,29 doğruluk, DFDC veri seti üzerinde ise 0.1679 loss değeri elde edilmiştir (Zhao ve diğ., 2021).

Wang ve arkadaşları tarafından yapılan çalışmada, deepfake tespiti için farklı uzamsal seviyelerde ortaya çıkan ince manipülasyon artifaktlarını transformer ağlar kullanarak yakalamayı hedeflemişlerdir. Bu çalışmada farklı boyutlardaki yamalar üzerinde çalışan Multi-modal Multi-scale Transformer (M2TR) ağı tanıtılmaktadır. Bu ağ, dikkatlice tasarlanmış bir çapraz modalite füzyon bloğu sayesinde RGB bilgilerine tamamlayıcı olarak frekans alanındaki sahtecilik artifaktlarını da öğrenir. Ayrıca çalışmada FaceForensics++ veri setinin gerçek görüntüleri üzerinden oluşturulan 4.000 sahte video’yu barındıran yüksek kaliteli SR-DF veri seti de sunulmuştur. Bu modelde SR-DF, FaceForensics++, Celeb-DF ve ForgeryNet veri setleri kullanılmış olup, Celeb-DF üzerinde 0,95, FaceForensics++ üzerinde 0,99, ForgeryNet üzerinde 0,82 ve SR-DF üzerinde 0.86 AUC değeri elde edilmiştir (Wang ve diğ., 2022).

Raza ve Malik tarafından yapılan çalışmada, deepfake tespiti için mevcut yöntemlerin tek bir modaliteye dayandığını ancak üretilen gerçekçi deepfake videoların ses ve görüntü üzerinden manipülasyon yapılarak üretildiği için ses ve görüntüleri senkronize ederek deepfake tespiti yapma yöntemi önermişlerdir. Önerilen bu yöntem “Multimodaltrace” adlı bir çerçevedir. Bu çerçevede, görüntü ve seslerden öğrenilen özellikler IntrAmodality Mixer Layer (IAML) ile bağımsız olarak karıştırılmakta ve IntErModality Mixer Layers (IEML) ile işlenerek çok etiketli sınıflandırma başlığına aktarılmaktadır. Bu çalışmada FakeAVCeleb, World Leaders ve Presidential Deepfake tespit veri setleri kullanılmış olup FakeAVCeleb veri setinde %92,9’luk doğruluk, World Leaders ve Presidential Deepfake veri kümeleri üzerinde çapraz değerlendirme yöntemi ile sırasıyla %83,61 ve %70 doğruluk değeri elde edilmiştir (Raza ve Malik, 2023).

Bayar ve Stamm tarafından yapılan çalışmada deepfake tespiti için derin öğrenmeye dayalı bir yöntem sunulmaktadır. Bayar ve Stamm, deepfake üretimi sırasında farklı manipülasyon türlerinin uygulanabildiği ve her bir manipülasyon türünün ayrı ayrı test edilmesinin zor olduğunu, bundan dolayı birçok farklı manipülasyon türünü tespit edebilen algoritma geliştirilmesi gerektiğini savunmuşlardır. Bu doğrultuda, eğitim verilerinden doğruca manipülasyon türlerini öğrenen yeni bir evrişimsel ağ mimarisi önerilmiştir. Bu modelde, ekstra ön işleme veya seçilmiş özelliklere ihtiyaç duyulmadan %99,10 doğruluk değeri elde edilmiştir (Bayar ve Stamm, 2016).

Xuan ve arkadaşları tarafından yapılan çalışmada, deepfake tespitinde genellemeyi artırmaya yönelik yöntemler incelenmiştir. Önerilen yöntemde, sahte ve gerçek görüntülere benzer şekilde uygulanmış ön işleme adımları kullanılarak, seviyesi düşük gürültülü ipuçlarının yok edilmesi sağlanmaktadır. Bu sayede evrişimsel sinir ağı tabanlı model, görüntü içeriğinden ziyade kalıcı ve daha özsel özellikleri öğrenerek deepfake tespiti yapabiliyor. Bu çalışmada yaygın deepfake veri setleri kullanılarak yüksek performans değerleri elde edilmiştir (Xuan ve diğ., 2019).

Qi ve arkadaşları tarafından yapılan çalışmada, sahte görüntü tespiti yapmak için yüz bölgesindeki görsel nabız ritimlerini analiz eden DeepRythem yöntemini sunmuşlardır. Qi ve arkadaşları, gerçek yüz videolarında, kan akışına bağlı olarak ciltte meydana gelen periyodik değişimlerin fotopletismografi (PPG) ile izlenebilir; deepfake videolarda ise bu doğal nabız ritimleri bozulmaktadır varsayımına dayanarak bu yöntemi geliştirmişlerdir. DeepRythem yöntemi, çift yönlü uzaysal-zamansal dikkat ve hareket büyütme tabanlı uzaysal-zamansal temsil (MMSTR) mekanizmalarını kullanarak farklı sahtecilik türlerine uyum sağlar. Bu çalışmada FaceForensics++ ve DFDC-preview veri setleri kullanılmış olup FaceForensics++ veri seti üzerinde %96 doğruluk elde edilmiştir (Qi ve diğ., 2020).

Barni ve arkadaşları tarafından yapılan çalışmada, deepfake tespiti yapmak için GAN tarafından üretilen görüntülerin doğasında bulunan farklılıkların analiz edilmesini savunmuşlardır. Bu doğrultuda sahte görüntü tespiti yapmak için cross-band co-occurrence (çapraz bant eşleşmeleri) analizi yöntemini kullanmışlardır. Bu yöntem, görüntüdeki farklı frekans bantları arasındaki ilişkileri inceleyerek sahte ve gerçek görüntü arasındaki farkları belirlemeyi amaçlamaktadır. Yöntem, evrişimsel sinir ağı modeli ile birleştirilerek GAN tarafından üretilen görüntülerin yüksek doğrulukla tespit edilmesini sağlamaktadır. Bu çalışmada ProGAN ve StyleGAN modelleri ile üretilmiş görseller kullanılmış olup yüksek doğruluk oranları elde edilmiştir (Barni ve diğ., 2020).

Ismail ve arkadaşları tarafından yapılan çalışmada, deepfake tespiti için derin öğrenme tabanlı bir metodoloji önerilmiştir. Bu yöntemde ilk olarak YOLO yüz tespit edici kullanılarak video framelerinden yüz bölgeleri çıkarılmakta; ardından çıkarılan bu yüz bölgelerinden InceptionResNetV2 ağı özellik çıkarımı yapılmaktadır. Çıkarılan bu özellikler, evrişimsel sinir ağının üst seviyesinde XGBoost sınıflandırıcısı tarafından beslenmektedir. Yapılan bu çalışmada CelebDF-FaceForencics++ birleşik veri seti kullanılmış olup 0,90 AUC, %90,73 doğruluk, %85,39 recall, %87,36 precision değeri elde edilmiştir (Ismail ve diğ., 2021).

Güera ve Delp tarafından yapılan çalışmada, deepfake tespiti için tekrarlayan sinir ağlarının (RNN) kullanıldığı bir yöntem önerilmiştir. Bu yöntemde, ilk önce videolardan frameler elde edilir, daha sonra bu frameler üzerinden evrişimsel sinir ağı (CNN) ile özellikler çıkarılır. Çıkarılan bu özellikler tekrarlayan sinir ağı (RNN) modeline beslenir. Bu sayede deepfake videolarda ortaya çıkan çerçeveler ile gerçek videolarda doğal olarak görülen zamansal tutarsızlıklar arasındaki farklar tespit edilir. Bu çalışmada çeşitli sitelerden toplanan 300 deepfake video ve HOHA veri setinden rastgele seçilen 300 gerçek video kullanılmış olup %97’lik bir doğruluk oranı elde edilmiştir (Güera ve Delp, 2018).

Caldelli ve arkadaşları tarafından yapılan çalışmada, eğitim sırasında görülmemiş deepfake manipülasyonlarının tespitine odaklanılmıştır. Yapılan çalışmada, optical flow tabanlı bir evrişimsel sinir ağı (CNN) mimarisi yöntemi kullanılmıştır. Bu yöntemde, deepfake videolarda oluşan zamansal tutarsızlıklar ve anomolilerini yakalamak için video kareleri arasındaki hareket (optical flow) hesaplaması yapılmaktadır. Bu sayede eğitim verilerinde yer almayan manipülasyon türleri bile etkili şekilde tespit edilebilmektedir. Yapılan çalışmada elde edilen deneysel sonuçlar, yöntemin genelleme yeteneğinin yüksek olduğunu ve farklı deepfake manipülasyonlarına karşı sağlam performans gösterdiğini ortaya koymaktadır (Caldelli ve diğ., 2021).

Ciftci, Demir ve Yin tarafından yapılan çalışmada, ana hedef deepfake’in kaynağını tespit etmektir. Bu sebeple deepfake tespiti için insan biyolojik sinyallerinden yararlanan yenilikçi bir yöntem sunmuşlardır. Bu yöntemde yüz bölgesindeki küçük renk değişimlerinden türeyen fotopletismografi (PPG) sinyalleri kullanılmıştır. Elde edilen bu sinyaller analiz edilerek hem deepfake tespiti yapılmış hem de farklı GAN tabanlı yöntemler birbirinden ayırt edilmiştir. Bu çalışmada CelebDF, Deepfakes, Face2Face, FaceSwap, NeuralTex görüntüleri kullanılmış olup toplamda %93,69 doğruluk oranı elde edilmiştir (Ciftci, Demir ve Yin, 2020).

Schwarcz ve Chellappa tarafından yapılan çalışmada, deepfake tespiti için deepfake manipülasyonlarında ortaya çıkan artifaktaları parçalara dayalı (parts-based) dedektörler kullanarak tespit yapmayı önermektedir. Schwarcz ve Chellappa, mevcut üretim yöntemlerinin genellikle yüzün belirli bölgelerinde yerel tutarsızlıklar ve anomoliler ürettiğini gözlemlemişler. Önerilen bu yöntemde, yüz bölgelerine odaklanan ayrı ayrı dedektörler geliştirilmekte ve elde edilen lokal özellikler birleştirilerek, sahtecilik izlerini daha net bir şekilde ortaya çıkaran temsil oluşturulmaktadır. Yapılan bu çalışmada FaceForensics++, Celeb-DF ve DFDC veri setleri kullanılmış olup FaceForensics++ veri seti üzerinde 0,98, Celeb-DF veri seti üzerinde 0,94 AUC değeri elde edilmiştir (Schwarcz ve Chellappa, 2021).

Luo ve arkadaşları tarafından yapılan çalışmada, deepfake tespitinde genellemeyi artırmak için özelliklerin frekans sayısına göre kullanımını önermektedir. Luo ve arkadaşları yaptığı araştırmalarda mevcut tespit yöntemlerinin eğitildiği veri setlerinin özelliklerine aşırı uyum sağladığını gözlemlemiş ve bu nedenle ayrık kosinüs dönüşümü (DCT) tabanlı bir ön işleme adımı ekleyerek yüksek frekanslı özellikleri belirginleştirmişlerdir. Yapılan bu çalışmada FaceForensics++, Celeb-DF ve DFDC veri setleri kullanılmış olup FaceForensics++ ile eğitilen model Celeb-DF üzerinde %92,3 AUC değeri elde etmiştir (Luo ve diğ., 2021).

He ve arkadaşları tarafından yapılan çalışmada, deepfake tespiti için GazeForensics adını verdikleri yeni bir yöntem önermişlerdir. GazeForensics yöntemi, insan bakış yönünü (gaze direction) ve bölgesindeki mekânsal ilişkiler üzerinden tutarsızlıkları analiz ederek deepfake görüntüleri ortaya çıkarmayı hedeflemektedir. Yapılan çalışmada FaceForensics++, Celeb-DF ve DFDC veri seti kullanılmış olup, FaceForensics++ veri seti üzerinde %98,1 AUC ve %95,4 doğruluk değeri elde edilmiştir (He ve diğ., 2023).

Tablo 4 : Deepfake Tespit Yöntemlerinin Özeti

|  |  |  |
| --- | --- | --- |
| **Çalışma Model** | **En Yüksek Performans** | **Veri Seti** |
| **EfficientNet + Vision Transformers** (Coccomini ve diğ., 2022) | 0.951 AUC,  %88.0 F1 | DeepFake Detection Challenge (DFDC) |
| **MesoNet** (Afchar ve diğ., 2018) | %98 Doğruluk (DeepFake),  %95 Doğruluk (Face2Face) | FaceForensics++ |
| **Convolutional Vision Transformer (CVT)** (Wodajo ve Atnafu, 2021) | %91.5 Doğruluk,  0.91 AUC | DFDC |

*Tablo 4 :Deepfake Tespit Yöntemlerinin Özeti (Devam)*

|  |  |  |
| --- | --- | --- |
| **Capsule Forensics** (Nguyen ve diğ., 2019) | %92.79 Doğruluk  (FaceSwap) | FaceForensics++ |
| **Yüz Deformasyonu Analizi** (Li ve Lyu, 2018) | 0.97 AUC (UADFV) | UADFV (Özel Oluşturulmuş) |
| **LRCN (Göz Kırpma Analizi)** (Li ve diğ., 2018) | 0.99 AUC | CEW + Özel Veri Seti |
| **DSM (Domain Shift Modeling)** (Zhang ve diğ., 2025) | 0.98 AUC (FaceForensics++), 0.92 AUC (Celeb-DF) | FaceForensics++, Celeb-DF, DeepfakeTIMIT |
| **STIL (Mekânsal-Zamansal Tutarsızlık)** (Gu ve diğ., 2021) | 0.99 AUC  (FaceForensics++) | FaceForensics++, Celeb-DF, DFDC, WildDeepfake |
| **MultimodalTrace** (Raza ve Malik, 2023) | %92.9 Doğruluk  (FakeAVCeleb) | FakeAVCeleb, World Leaders, Presidential Deepfake |
| **DeepRhythm (PPG Sinyali Analizi)** (Qi ve diğ., 2020) | %96 Doğruluk | FaceForensics++, DFDC-preview |
| **Cross-band Co-occurrence** (Barni ve diğ., 2020) | Yüksek Doğruluk (GAN Görüntüleri) | ProGAN, StyleGAN Üretimi |
| **YOLO + InceptionResNetV2 + XGBoost** (Ismail ve diğ., 2021) | 0.90 AUC,  %90.73 Doğruluk | CelebDF-FaceForensics++ Birleşik |

*Tablo 4: Deepfake Tespit Yöntemlerinin Özeti (Devam)*

|  |  |  |
| --- | --- | --- |
| **RNN + CNN (Zamansal Analiz)** (Güera ve Delp, 2018) | %97 Doğruluk | Özel Derleme (300 Gerçek + 300 Sahte) |
| **Optical Flow Tabanlı CNN** (Caldelli ve diğ., 2021) | Yüksek Genelleme | FaceForensics++, Celeb-DF |
| **Fotopletismografi (PPG)** (Ciftci ve diğ., 2020) | %93.69 Doğruluk | CelebDF, Deepfakes, Face2Face, FaceSwap |
| **Parts-Based Detectors** (Schwarcz ve Chellappa, 2021) | 0.98 AUC (FaceForensics++), 0.94 AUC (Celeb-DF) | FaceForensics++, Celeb-DF, DFDC |
| **DCT + Yüksek Frekans Analizi** (Luo ve diğ., 2021) | %92.3 AUC  (Çapraz Dataset) | FaceForensics++, Celeb-DF, DFDC |
| **GazeForensics** (He ve diğ., 2024) | 0.98 AUC,  %95.4 Doğruluk | FaceForensics++, Celeb-DF, DFDC |
| **MINTIME** (Coccomini ve diğ., 2024) | Yüksek Genelleme (Çapraz Dataset) | FaceForensics++, Celeb-DF, SR-DF |
| **Evrişimsel İz Analizi** (Guarnera ve diğ., 2020) | %99.81 Doğruluk  (CelebA vs StyleGAN2) | CelebA, GAN Üretimleri |
| **Ensemble CNN** (Bonettini ve diğ., 2020) | %83 Doğruluk | ImageNet |
| **Spatiotemporal CNN (STCN)** (De Lima ve diğ., 2020) | Yüksek Doğruluk (Zamansal Analiz) | FaceForensics++, DFDC |

*Tablo 4: Deepfake Tespit Yöntemlerinin Özeti (Devam)*

|  |  |  |
| --- | --- | --- |
| **Sentetik Veri Tabanlı Model** (Jain ve diğ., 2022) | %93 Doğruluk, 0.95 AUC | FaceForensics++, Celeb-DF |
| **Multi-modal Multi-scale Transformer (M2TR)** (Wang ve diğ., 2022) | 0.99 AUC | FaceForensics++, SR-DF, Celeb-DF |
| **Manipülasyon Türü Tabanlı Model** (Bayar ve Stamm, 2016) | %99.10 Doğruluk | Özel Derleme (Çoklu Manipülasyonlar) |
| **Ön İşleme Odaklı Genelleme** (Xuan ve diğ., 2019) | Yüksek Doğruluk (Çapraz Dataset) | FaceForensics++, Celeb-DF |
| **Bölgesel Bağımsızlık Kaybı** (Zhao ve diğ., 2021) | %99.29 Doğruluk (FaceForensics++) | FaceForensics++, DFDC |



YöntEm (metodoloji)

* 1. Veri Seti

Bu tez kapsamında yapılan deepfake tespiti için Celeb-DF ve DeepFakeDetection Dataset olmak üzere iki farklı veri seti tercih edilmiştir. Bu iki veri seti de farklı senaryolar ve sahnelerden oluşmaktadır. Bu veri setlerinin kullanılma amacı, birçok farklı çeşit manipülasyondan oluşmaları nedeniyle modelin genelleme kabiliyetinin artırılması hedeflenmiştir.

Celeb-DF veri seti, 890 gerçek görüntü ve 5639 sahte görüntüden oluşmaktadır. Sahte videolar, gerçek videolara uygulanan face swap algoritmaları ile oluşturulmuştur. Videolar, yüksek çözünürlükte ve çok düşük seviyede artefakt içerecek şekilde tasarlandığından dolayı deepfake tespiti için zorlu bir kaynak sunmaktadır. Celeb-DF dizin yapısı:

Celeb-DF/

│

├── Celeb-real/ ← Gerçek videolar (590)

│ ├── id0\_0000.mp4

│ ├── id0\_0001.mp4

│ └── ...

│

├── YouTube-real/ ← Gerçek videolar (YouTube kaynaklı) (300)

│ ├── 00000.mp4

│ ├── 00001.mp4

│ └── ...

│

└── Celeb-synthesis/ ← Deepfake (sahte) videolar (5639)

├── id0\_ id0\_0000.mp4

├── id0\_ id0\_0001.mp4

└── ...

DFD (DeepFakeDetection Dataset) veri seti, Jigsaw ve Google Research tarafından oluşturulmuştur. DFD veri setinde 364 gerçek ve 3068 sahte video bulunmaktadır. Sahte videolar, gerçek videolara farklı deepfake algoritmaları uygulanarak elde edilmiştir. Bu veri seti, deepfake araştırmalarında güvenilirlik ve etik kaygılar gözetilerek oluşturulmuş ilk veri setlerinden biridir. DFD dizin yapısı: DFD/

│

├── DFD\_original\_sequences / ← Gerçek videolar (364)

│ ├── 01\_exit\_phone\_room.mp4

│ ├── 01\_hugging\_happy \_real.mp4

│ └── ...

│

└── DFD\_manipulated\_sequences/ ← Sahte videolar (3068)

├── 01\_02\_exit\_phone\_room\_YVGY8LOK.mp4

├── 01\_02\_hugging\_happy\_ \_YVGY8LOK.mp4

└── ...

* + 1. 3.1.1 Veri örnekleme ve dengeleme süreci

Bu çalışmada, Celeb-DF ve DFD veri setlerindeki tüm gerçek videolar ile DFD veri setinden 1000 adet, Celeb-DF veri setinden ise 4000 adet rastgele sahte video örneği seçilmiştir. Toplamda 1364 gerçek ve 5000 sahte video toplanmıştır. Bu verilerin %70’i eğitim, %15’i test ve %15’i doğrulama için ayrılmıştır.

* 1. Veri Ön İşleme Aşamaları

Bu çalışmada Celeb-DF ve DFD veri setinden alınan videolar doğrudan model eğitimi için uygun yapıda olmadığından, çeşitli ön işleme adımları uygulanarak modele uygun hale getirilmiştir. Bu adımlar aşağıda detaylı bir şekilde açıklanmıştır.

* + 1. Video karelerine ayrıştırma ve yüz tespiti

Her videodan her 10 karede bir çerçeve (frame) alınarak, gereksiz veri azaltılmış ve daha temsil edici örneklere sahip olunmuştur. Kare çıkarma işlemi sadece çerçeve ayrıştırmasıyla sınırlı kalmamış, aynı zamanda bu kareler üzerinden yüz tespiti yapılmıştır.

Bu süreçte yüz tespiti için Multi-Task Cascaded Convolutional Neural Network (MTCNN) modeli kullanılmıştır. Bu model ile %92 güven seviyesinin üstünde ve yalnızca 64x64 piksel boyutlarında tespit edilen yüzler işlenmiştir. İşlem süresini azaltmak için Python dilinde*concurrent.futures.ThreadPoolExecutor*yapısı kullanılarak paralel işlem yapılmıştır. Her frame işlenmeden önce RGB formatına dönüştürülmüş ve ardından yüz bölgeleri genişletilerek kırpılmıştır. Elde edilen yüz görselleri, *video\_adı\_frameNo\_faceNo.jpg* formatında isimlendirilmiş ve ilgili train, valve test dizinlerine kaydedilmiştir.

* + 1. Veri etiketleme ve DataFrame oluşturma

Modelin eğitimi, testi ve doğrulaması için görüntülerin sınıflarına (gerçek ve sahte) göre etiketlenerek uygun bir veri yapısına dönüştürülmesi gerekmektedir. Bu süreçte, çıkarılan yüz görüntüleri eğitim (train), doğrulama (validation) ve test alt kümelerine bölünmüş ve her biri ayrı üç farklı sınıf dizininde saklanmıştır. Bu dizinlerden görüntüler okunarak veri çerçeveleri (DataFrame) oluşturulmuştur. Her görüntü dosyasının yol bilgisi (*file\_path*) ile, ait olduğu sınıf etiketi (*label*) atanmış; gerçek görüntülere 0, sahte görüntülere ise 1 etiketi verilmiştir.

* + 1. Sınıf dengesizliği giderme

Gerçek ve sahte görüntü verileri arasındaki sayısal dengesizlik, modelin eğitim sürecinde aşırı öğrenme (overfitting) problemine yol açabileceğinden, bu problemi önlemek amacıyla sklearn kütüphanesinden  fonksiyonu kullanılarak sınıfların ağırlıkları hesaplanmıştır. Hesaplanan bu ağırlıklar, modelin eğitimi sırasında *class\_weight* parametresi olarak kullanılmıştır. Bu yöntem sayesinde modelin azınlık sınıfa (gerçek) daha fazla önem vermesi sağlanmış ve sahte videolara karşı aşırı uyum (overfitting) sağlaması engellenmiştir.

* + 1. Görüntü ölçekleme ve ön işleme

Eğitim verileri üzerinde EfficientNetV2S ağı ile uyumlu olacak şekilde  fonksiyonu ile görüntüler normalize edilmiştir. Bunun yanı sıra, modelin genelleme yeteneğini artırmak amacıyla çeşitli veri arttırma işlemleri uygulanmıştır. Görüntüler rastgele döndürülmüş, kaydırılmış, kontrast ve parlaklık gibi renk tabanlı bozulmalar uygulanmıştır. Bu sayede modelin farklı aydınlatma ve pozisyon koşullarına karşı daha dayanıklı hale gelmesi sağlanmıştır.

* + 1. Görüntü akışlarının (generator) oluşturulması

Veri seti, *train\_df*, *val\_df*ve *test\_df* olarak ayrı ayrı veri çerçevelerine ayrılmıştır. Bu veri çerçevelerinden *ImageDataGenerator* kullanılarak eğitim, test ve doğrulama setleri için akışlar tanımlanmıştır. Bu akışlar oluşturulurken eğitim veri setine, 3.2.4 Görüntü Ölçekleme ve Ön İşleme başlığı altında belirtilen adımlar uygulanmış, doğrulama ve test veri setlerine ise yalnızca EfficientNetV2S için gerekli normalizasyon işlemi yapılmıştır. Bu sayede model, yalnızca eğitim aşamasında veri çeşitliliği ile beslenmiş ve test aşamasında gerçek veriye en yakın senaryo korunmuştur.

* 1. Model Mimarisi ve Eğitimi
     1. Modelin temeli EfficentNetV2S ile transfer learning

Bu çalışmada transfer learning yaklaşımı uygulanmıştır. Transfer learning modeli olarak daha önce ImageNet veri seti üzerinde eğitilmiş olan EfficientNetV2S ağı kullanılmıştır. Bu ağ EfficientNetV2 serisinin daha küçük ve hızlı varyantlarından biridir bunun yanın da sınırlı veri setlerinde bile yüksek doğruluk sağlayabilmektedir.

Model, parametresi ile son sınıflandırma katmanları olmadan yüklenmiştir. Model yüklendikten sonra ise modelin gövdesi dondurulmuş ve uygulamaya yönelik yeni katmanlar eklenerek model eğitilmiştir. Parametresi ile modelin çıktısında Global Average Pooling uygulanması sağlanmış, bu sayede vektörleştirilmiş özellik temsilleri elde edilmiştir.

* + 1. Yeni katmanların eklenmesi ve özelleştirilmesi

Modelin çıktısına uygulamaya yönelik sınıflandırıcı bloğu eklenerek gerçek ve deepfake görüntü ayrımı yapması hedeflenmiştir. Eklenen bloklar aşağıdaki bileşenleri içermektedir:

* **Dense (Yoğun) Katmanlar:** 512, 128 ve 64 adet nöron içeren ve Aktivasyon fonksiyonu ReLU olan katmanlar sayesinde modelin karmaşık desenleri öğrenmesi sağlanmıştır.
* **L2 Regularizasyonu**: Bu 3 adet Dense katmanına uygulanarak modelin genelleme yeteneği arttırılmıştır.
* **Batch Normalization**: Her Dense katmanından sonra bu katman eklenerek, katman çıktılarının normalleştirilmesi sağlanmıştır bu sayede öğrenme hızı artmıştır.
* **Dropout**: %40 ila %60 oranlarında dropout katmanları eklenerek overfitting (aşırı öğrenme) riski azaltılmıştır.
* **Çıkış Katmanı**: Tek nöronlu ve sigmoid aktivasyonlu bu katman sayesinde çıktı değeri [0,1] aralığında olacak şekilde binary sınıflandırma yapılmıştır.

|  |  |
| --- | --- |
| **Bileşen** | **Değerler** |
| Dense Katmanlar | 512, 128, 64 (ReLU) |
| L2 Regularizasyon | 0.001 |
| Batch Normalization | ✓ |
| Dropout | %60, %50, %40 |
| Çıkış Katmanı | 1 nöron (sigmoid) |

Tablo 5: Model Katmanları

* + 1. Modelin derlenmesi ve eğitim aşaması

Modelin eğitiminde istikrar sağlamak ve performansı artırmak amacıyla aşağıdaki ayarlamalar yapılmıştır:

* **Optimizer:** Optimizer olarak *Adam* optimizasyon algoritması kullanılmıştır.
* **Öğrenme Oranı Planlayıcısı: kullanılarak her belirli bir adımda öğrenme oranı %2 oranında merdiven tipi ile azaltılmıştır. Bu sayede model başta hızlı öğrenirken, zamanla daha stabil öğrenme yapmıştır.**
* **Kayıp (Loss) Fonksiyonu:** İkili bir sınıflandırma olduğundan dolayı kayıp(loss) fonksiyonu olarak kullanılmıştır.
* **Değerlendirme Metriği: Modelin başarımı** accuracy (doğruluk) ve ROC AUC metriği ile ölçülmüştür.
* **Callback’ler:** Modelin eğitimi sırasında overfitting’i engellemek amacı ile ve en iyi modeli kaydetmek için callback’leri kullanılmıştır.

|  |  |
| --- | --- |
| **Parametre** | **Ayar** |
| Optimizer | Adam |
| Öğrenme Oranı Planı | ExponentialDecay (%2 azalma) |
| Kayıp Fonksiyonu | binary\_crossentropy |
| Metrikler | accuracy, AUC |
| Callback’ler | EarlyStopping, ModelCheckpoint |

Tablo 6 : Model Derlenme ve Eğtim Bileşenleri

* 1. Fine-Tuning (İnce Ayar)

Transfer learning aşamasında, EfficientNetV2S modelinin gövdesindeki tüm katmanlar dondurulmuş ve yalnızca üst sınıflandırıcı bloğu eğitilmiştir. İnce ayar (fine-tuning) adımında ise, modelin gövdesinde yer alan bazı katmanlar tekrar eğitilebilir hâle getirilerek, ağın daha derin özelliklerinin mevcut veri setine daha iyi uyum sağlaması amaçlanmıştır;

1. **Veri Dengesinin Yeniden Sağlanması**
   * Pretrained model eğitimi sırasında sınıf dengesizliği, compute\_class\_weight() fonksiyonu kullanılarak ağırlıklandırma yöntemiyle giderilmiştir. Fine-tuning aşamasında ise aynı eğitim veri çerçevelerinden gerçek ve sahte örnek sayıları eşit olacak şekilde rastgele satırlar seçilerek (sample () ile) dengeli yeni DataFrame’ler oluşturulmuştur.
2. **Veri Artırma (Augmentation) ve Ön İşleme**
   * Fine-Tuning (İnce Ayar) aşamasına özel tanımlandı.
   * Görüntülere pretrained modeldeki kadar sert augmentasyonlar uygulamak yerine daha hafif augmentasyonlar uygulandı.
3. **Model Katmanlarının Dondurulması ve Açılması**
   * Başlangıçta, pretrained modelde bulunan 517 adet katman *trainable=False* yapılarak donduruldu. Daha sonra 350. katmandan sonra BatchNormalization katmanları hariç tüm katmanlar yeniden eğitilebilir hale getirildi.
4. **Yeniden Derleme (Re-Compilation)**
   * İnce ayar sırasında çok daha düşük bir başlangıç öğrenme oranı (3 × 10⁻⁵) kullanıldı.
   * ile her birkaç bin adımda öğrenme oranı %6 azaltılarak merdiven tipi düşüş uygulandı.
   * Optimizasyon için yine Adam, kayıp fonksiyonu olarak *binary\_crossentropy*, başarı metriği olarak doğruluk ve AUC kullanıldı.
5. **Ek Callback’ler ve Eğitim**
   * (3 sabır adımı) ve  ile en iyi modelin kaydedilmesi callback’leri devam ettirildi.
   * İnce ayar eğitimi 30 epoch boyunca, eğitim ve doğrulama akışları kullanılarak gerçekleştirildi.

|  |  |
| --- | --- |
| **Adım** | **Ayar/İşlem** |
| Veri Dengesinin Yeniden Sağlanması | ile gerçek/sahte örnek sayıları eşitlendi |
| Veri Artırma & Ön İşleme | İnce ayar için hafif augmentasyon (ImageDataGenerator) ve uygulandı |
| Katmanların Dondurulması & Açılması | Başlangıçta 517 katman donduruldu; 350. katmandan sonrası (BatchNorm hariç) açıldı |
| Yeniden Derleme | LR=3×10⁻⁵; (%6 azalma, merdiven tipi); Adam; binary\_crossentropy; accuracy, AUC |
| Callback’ler & Eğitim | ve ; 30 epoch boyunca eğitim yapıldı |

Tablo 7 : Fine-Tuning Süreci Özet



Bulgular

* 1. Ön Eğitim Aşaması Sonuçları

Transfer learning sürecinde, ImageNet üzerinde eğitilmiş EfficientNetV2S gövdesi dondurularak yalnızca üst sınıflandırıcı bloğu eğitilmiştir. Bu eğitim, 5 epoch sonunda aşırı ezberleme nedeniyle EarlyStopping callback’i tarafından durdurulmuş; test kümesinde 0,59 kayıp (loss), %81 doğruluk (accuracy) ve 0,82 (Area Under the ROC Curve) değeri elde edilmiştir.

Sınıf bazlı performans analizinde ise gözlemlerim şunlardır:

* **Gerçek (real) sınıfı**:
  + Support = 9 838 örnek
  + Precision = 0.68
  + Recall = 0.43
  + F₁ skoru = 0.53

Bu değerler, modelin gerçek videoları doğru sınıflandırma konusundahâlâ yetersiz kaldığını, sadece dörtte bir oranında gerçek videodan yakaladığını işaret etmektedir.

* **Sahte (fake) sınıfı**:
  + Support = 31 181 örnek
  + Precision = 0.84
  + Recall = 0.94
  + F₁ skoru = 0.89

Sahte videoların tespitinde yüksek recall ve precision, modelin manipüle edilmiş içeriklerde bıraktığı artefaktları tanıma konusunda güçlü olduğunu ortaya koymuştur.

Modelin sınıf bazlı performans değerlerine baktığımızda, dengesiz sınıf dağılımından etkilenerek gerçek görüntülerde iyi performans gösteremediği; ancak sahte görüntü sınıfında güçlü genellemeler yapabildiği gözlemlenmiştir.

A graph of training and validation

AI-generated content may be incorrect.

A graph of training and validation

AI-generated content may be incorrect.

Şekil 13 : Ön Eğitim Aşaması Sırasında Her Bir Epoch Sonunda Elde Edilen Accuracy, Loss ve AUC Değerleri

* 1. Fine-Tuning (İnce Ayar) Aşaması Sonuçları

Fine-tuning (ince ayar) adımlarıyla modelin derin katmanları probleme özgü verilerle güncellenmiştir. İnce ayar sonrası model, test kümesinde 0,35 kayıp (loss), %89 doğruluk (accuracy) ve 0,96 AUC (Area Under the ROC Curve) değeri elde etmiştir. Bu metriklerdeki yükseliş ve sonuç, modelin genel karar sınırlarını hem gerçek hem de sahte görüntülerde daha doğru şekilde ayarladığını göstermektedir.

Sınıf bazlı performans analizinde ise gözlemlerim şöyledir:

* **Gerçek sınıfı** (support: 9 838):
  + Precision = 0.91
  + Recall = 0.87
  + F₁ skoru = 0.89
* **Sahte sınıfı** (support: 9 838):
  + Precision = 0.88
  + Recall = 0.91
  + F₁ skoru = 0.89

Artan precision ve recall değerleri, özellikle gerçek görüntü sınıfındaki false negative (kaçırma) oranının ciddi oranda azaldığını göstermektedir. Fine-tuning (ince ayar) işlemi, recall ve precision odaklı metriklerde önceki aşamaya kıyasla %40’a varan mutlak artış sağlamış ve modelin daha dengeli bir sınıf performansı sergilemesine olanak tanımıştır.

A graph of training and validation

AI-generated content may be incorrect.

A graph with blue and orange lines

AI-generated content may be incorrect.

Şekil 14 : Fine-Tuning Aşaması Sırasında Her Bir Epoch Sonunda Elde Edilen Accuracy, Loss ve AUC Değerleri

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Aşama | Loss | Accuracy | AUC | Sınıf | Precision | Recall | F1 Skoru | Support |
| Pretrained | 0.59 | 0.81 | 0.82 | Gerçek | 0.68 | 0.43 | 0.53 | 9 838 |
|  |  |  |  | Sahte | 0.84 | 0.94 | 0.89 | 31 181 |
| Fine-Tuned | 0.35 | 0.90 | 0.96 | Gerçek | 0.91 | 0.87 | 0.89 | 9 838 |
|  |  |  |  | Sahte | 0.88 | 0.91 | 0.89 | 9 838 |

Tablo 8: Modelin Sonuç Özeti

* 1. Literatürdeki Diğer Çalışmalarla Karşılaştırma

Tablo 8 incelendiğinde, literatürde %95–99 doğruluk seviyelerine ulaşan bazı gelişmiş modellerin, özel olarak tasarlanmış ağ mimarileri (örneğin, Capsule Networks, Transformer tabanlı yapılar), genellikle çok daha büyük veri setleri veya görüntü ve sesi birlikte kullanan multimodal yaklaşımlar kullandığı görülmektedir. Buna karşılık, mevcut tez çalışmasında yalnızca görüntü temelli ve EfficientNetV2S mimarisi kullanılarak eğitilen modelin, literatürdeki yöntemlerle karşılaştırılabilir düzeyde başarılı sonuçlar ürettiği söylenebilir. Geliştirilen modelin hem hesaplama maliyetinin düşük olması hem de yüksek doğruluklu sonuçlar üretmesi nedeniyle gerçek dünya uygulamaları için daha uygulanabilir bir çözüm olarak öne çıkmakta ve potansiyelini ortaya koymaktadır.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Çalışma | Yöntem / Mimarî | Veri Seti | Accuracy | AUC |
| Afchar vd. (2018) | MesoNet (Mezoskopik Özellikler) | FaceForensics++ | 0.846 | 0.91 |
| Li vd. (2018) | RNN + Göz Kırpma Tespiti | Özgün veri seti | 0.830 | 0.99 |
| Sabir vd. (2019) | Uzay-zamansal Özellikler + RNN | FaceForensics++ | 0.926 | 0.96 |
| Rossler vd. (2019) | XceptionNet | FaceForensics++ (HQ) | 0.9926 | 0.995 |
| Ciftci vd. (2020) | Biyolojik İşaret + Residual | FF++ + Celeb-DF | 0.895 | 0.94 |
| Coccomini vd. (2022) | EfficientNet + Vision Transformer | DFDC | 0.920 | 0.95 |
| Rosler vd. (2019) | Xception | FaceForensics++ | 0.84 | 0.876 |
| Nguyen vd. (2019) | Capsule Network + Dynamic Routing | FaceForensics++, DeepFakeTIMIT | 0.96 | - |
| Jain vd. (2022) | **Sentetik Veri Tabanlı Model** | FaceForensics++, Celeb-DF | 0.93 | 0.95 |
| Mevcut Çalışma | EfficientNetV2S + Fine-Tuning | Celeb-DF + DFD | **0.90** | **0.96** |

Tablo 9 : Literatürdeki Diğer Çalışmalar ve Mevcut Çalışmanın Karşılaştırılması

|  |
| --- |
| Sonuçlar  Yapılan tez çalışmasında, son zamanlarda giderek artan bir şekilde kullanılan ve dijital alanda bireylerin gizliliğini ve mahremiyetini tehdit eden, aynı zamanda kamuoyunu yanıltmak ve yönlendirmek amacıyla kötüye kullanılabilen deepfake içeriklerin tespitine yönelik bir derin öğrenme tabanlı yaklaşım geliştirilmiştir. Çalışmanın temel amacı gerçek ve manipüle edilmiş yüz görüntülerini ayırt edebilen sınıflandırma modeli tasarlamak ve literatüre kazandırmaktır.  Çalışmada kullanılan veri setleri, Celeb-DF ve Deepfake Detection Dataset (DFD) gibi yaygın olarak tercih edilen güncel ve yüksek kaliteli veri deepfake veri kümeleridir.  Eğitim sürecinde EfficientNetV2S ağı kullanılarak modelin hem sınıflandırma başarımı hem de parametre verimliliği açısından güçlü bir performans göstermesi sağlanmıştır. İlk olarak önceden eğitilmiş (pretrained) model üzerine sınıflandırıcı blok eklenmiş, ardından fine-tuning aşaması ile modelin daha derin katmanları yeniden eğitime açılmıştır.  Modelin başarımı değerlendirilirken sadece doğruluk değil; AUC, precision, recall ve F1 skoru gibi çok boyutlu metrikler de dikkate alınmıştır. Yapılan tüm bu değerlendirmeler neticesinde, modelin yüksek güvenilirliğe sahip sonuçlar ürettiği ve literatürdeki diğer güncel yöntemlerle rekabet edebilir bir düzeye ulaştığı gözlemlenmiştir.      Öneriler  Bu çalışmada sadece görüntü verileri kullanılarak deepfake tespiti üzerine yoğunlaşılmıştır. Ancak deepfake içeriklerin çeşitliliği ve üretim tekniklerinin gelişimi göz önünde bulundurulduğunda, ileride yapılacak çalışmalarda hem**modelin değerlendirme boyutları** hem **de veri kümelerinin kapsamı** açısından aşağıdaki öneriler dikkate alınabilir:   * **Veri Setlerinin Güncellenmesi ve Genişletilmesi:** Mevcut veri setleri, özellikle sahte içeriklerin üretildiği teknikler açısından zamanla güncelliğini yitirmektedir. Günümüzde deepfake üretim teknikleri daha sofistike hale gelmiştir. Bu nedenle, gelecekteki çalışmalarda**güncel tekniklerle üretilmiş ve çeşitli yaş, cinsiyet, etnik yapı ve ortam koşullarını içeren yeni veri setlerinin kullanılması**, modelin genelleme ve gerçek dünya verileri başarımını artıracaktır. * **Veri Kaynağına Erişim ve Üretim:** Açık kaynaklı ve güncel deepfake video içeriklerine erişimin sınırlı olması, modelin gerçek dünya koşullarında test edilmesini güçleştirmektedir. Bu bağlamda, **yeni veri setlerinin oluşturulması**, etik kurallar çerçevesinde kullanıcı izniyle video kaydı alınarak veya çeşitli sosyal medya platformlarından kamuya açık verilerin derlenmesiyle sağlanabilir. * **Gerçek Zamanlı Uygulama Olanakları:** Eğitim süreci tamamlanan modellerin düşük gecikme süresiyle çalışabilmesi için **mobil veya gömülü sistemlerle uyumlu hafif modellerin** geliştirilmesi, deepfake tespit sistemlerinin sosyal medya platformlarına veya video konferans uygulamalarına entegre edilebilmesi açısından önemli bir adımdır.   Bu önerilerin ileride gerçekleştirilecek çalışmalara önemli katkılar sunması beklenmektedir.      Kaynakça  Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE.  Akın, E., & Şahin, M. E. (2024). Derin öğrenme ve yapay sinir ağı modelleri üzerine bir inceleme. *EMO Bilimsel Dergi*, *14*(1), 27-38.  Al-Dulaimi, O. A. H. H., & Kurnaz, S. (2024, May). Deep fake Image Detection Based on Deep Learning Using a Hybrid CNN-LSTM with Machine Learning Architectures as Classifier. In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-7). IEEE. |
|  |
| Arı, A., & Berberler, M. E. (2017). Yapay sinir ağları ile tahmin ve sınıflandırma problemlerinin çözümü için arayüz tasarımı. *Acta Infologica*, *1*(2), 55-73.  Arı, A., & Hanbay, D. (2019). Bölgesel evrişimsel sinir ağları tabanlı MR görüntülerinde tümör tespiti. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, *34*(3), 1395-1408. |
| Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2020, December). CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE.  Bayar, B., & Stamm, M. C. (2016, June). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM workshop on information hiding and multimedia security* (pp. 5-10). |
| BAYIR, F. (2006). Yapay sinir ağları ve tahmin modellemesi üzerine bir uygulama. *Yayınlanmamış Yüksek Lisans Tezi, İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü*.  Berk, M. E. (2020). Dijital çağın yeni tehlikesi “deepfake”. *OPUS International Journal of Society Researches*, *16*(28), 1508-1523.  Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021, January). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012-5019). IEEE.  Büyükavcılar, M. (2023). Derin öğrenme ile resim ve videolar üzerinde derin sahte tespiti Yüksek lisans tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı.  Caldelli, R., Galteri, L., Amerini, I., & Del Bimbo, A. (2021). Optical Flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, *146*, 31-37.  Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.  Ciftci, U. A., Demir, I., & Yin, L. (2020, September). How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE international joint conference on biometrics (IJCB)* (pp. 1-10). IEEE.  Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219-229). Cham: Springer International Publishing.  Coccomini, D. A., Zilos, G. K., Amato, G., Caldelli, R., Falchi, F., Papadopoulos, S., & Gennaro, C. (2024). MINTIME: multi-identity size-invariant video deepfake detection. *IEEE Transactions on Information Forensics and Security*.  Das, S., Tariq, A., Santos, T., Kantareddy, S. S., & Banerjee, I. (2023). Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. *Machine learning for Brain disorders*, 117-138.  De Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. (2020). Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*. |
|  |
| Dinçer, S., Ulutaş, G., Ustubioglu, B., Tahaoglu, G., & Sklavos, N. (2024). Golden ratio based deep fake video detection system with fusion of capsule networks. *Computers and Electrical Engineering*, *117*, 109234.  Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.  Ersoy, E., & Karal, Ö. (2012). Yapay sinir ağları ve insan beyni. *İnsan ve Toplum Bilimleri Araştırmaları Dergisi*, *1*(2), 188-205.  Fırıldak, K., & Talu, M. F. (2019). Evrişimsel sinir ağlarında kullanılan transfer öğrenme yaklaşımlarının incelenmesi. *Computer Science*, *4*(2), 88-95.  Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.  Google AI Blog. (2019, 24 Eylül). Deepfake tespiti araştırmalarına veri katkısı. Google Research. <https://research.google/blog/contributing-data-to-deepfake-detection-research/>  Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., & Ma, L. (2021, October). Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3473-3481).  Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 666-667).  Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE. |
|  |

Güera, D., Baireddy, S., Bestagini, P., Tubaro, S., & Delp, E. J. (2019). We need no pixels: Video manipulation detection using stream descriptors. *arXiv preprint arXiv:1906.08743*.

He, Q., Peng, C., Liu, D., Wang, N., & Gao, X. (2024). GazeForensics: DeepFake detection via gaze-guided spatial inconsistency learning. *Neural Networks*, *180*, 106636.

Heo, Y. J., Choi, Y. J., Lee, Y. W., & Kim, B. G. (2021). Deepfake detection scheme based on vision transformer and distillation. *arXiv preprint arXiv:2104.01353*.

Ismail, A., Elpeltagy, M., S. Zaki, M., & Eldahshan, K. (2021). A new deep learning-based methodology for video deepfake detection using XGBoost. *Sensors*, *21*(16), 5413.

İnik, Ö., & Ülker, E. (2017). Derin öğrenme ve görüntü analizinde kullanılan derin öğrenme modelleri. *Gaziosmanpaşa Bilimsel Araştırma Dergisi*, *6*(3), 85-104.

Jain, A., Memon, N., & Togelius, J. (2022, October). A dataless faceswap detection approach using synthetic images. In *2022 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1-7). IEEE.

Koç, E., Çalışkan, S., Yazıcıoğlu, S. A., Demirci, U., & Kuş, Z. (2018). Yapay Sinir Ağları, Kelime Vektörleri ve Derin Öğrenme Uygulamaları.

Korshunov, P., & Marcel, S. (2018). Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kütükçü, Y. E., & Polat, H. (2024, October). BYOL Yaklasimini Kullanarak MTCNN ile Yakalanan Yüzlerde Deepfake Tespiti Deepfake Detection on Faces Captured with MTCNN, by Using the BYOL Approach. In *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

|  |
| --- |
| Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.  Li, Y., Chang, M. C., & Lyu, S. (2018, December). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International workshop on information forensics and security (WIFS) (pp. 1-7). Ieee.  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* (pp. 3207-3216).  Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2015). Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*. |
|  |

Luo, Y., Zhang, Y., Yan, J., & Liu, W. (2021). Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16317-16326).

Minsky, M., & Papert, S. A. (2017). *Perceptrons, reissue of the 1988 expanded edition with a new foreword by Léon Bottou: an introduction to computational geometry*. MIT press.

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., ... & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525.

Okumuş, H. (2025). Yapay sinir ağlarının dalgacık dönüşümlü aktivasyon fonksiyonları ile iyileştirilmesi (Yüksek lisans tezi, Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Ana Bilim Dalı).

Öztemel, E. (2003). Yapay sinir ağlari. *PapatyaYayincilik, Istanbul*.

Öztürk, K., & Şahin, M. E. (2018). Yapay sinir ağları ve yapay zekâ’ya genel bir bakış. *Takvim-i Vekayi*, *6*(2), 25-36.

Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., ... & Zhao, J. (2020, October). Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 4318-4327).

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE access*, *10*, 25494-25513.

Raza, M. A., & Malik, K. M. (2023). Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 993-1000).

Saraç, T. (2004). Yapay sinir ağları. *Seminer projesi*, 22-75.

Schwarcz, S., & Chellappa, R. (2021). Finding facial forgery artifacts with parts-based detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 933-942).

Ser, G., & Bati, C. T. (2019). Derin sinir ağları ile en iyi modelin belirlenmesi: mantar verileri üzerine Keras uygulaması. *Yuzuncu Yıl University Journal of Agricultural Sciences*, *29*(3), 406-417.

URL1. (18.03.2025). [https://www.mimarmustafakemal.com](https://www.mimarmustafakemal.com/post/yapay-si%CC%87ni%CC%87r-a%C4%9Fi-nedi%CC%87r-yapay-si%CC%87ni%CC%87r-a%C4%9Fi-mi%CC%87mari%CC%87si%CC%87-nasil-olu%C5%9Fturulurhttps:/www.mimarmustafakemal.com/post/yapay-si%CC%87ni%CC%87r-a%C4%9Fi-nedi%CC%87r-yapay-si%CC%87ni%CC%87r-a%C4%9Fi-mi%CC%87mari%CC%87si%CC%87-nasil-olu%C5%9Fturulur)

URL2. (19.03.2025). [https://ennurcitir.wordpress.com](https://ennurcitir.wordpress.com/2020/02/19/yapay-sinir-aglari/)

URL3. (19.03.2025). [https://www.teknoloji.org/](https://www.teknoloji.org/cnn-convolutional-neural-networks-nedir/)

URL4. (07.04.2025). [https://ilge.com.tr](https://ilge.com.tr/generative-adversarial-netWorks-gan--yapay-zekanin-yeni-siniri)

URL5. (16.05.2025). [https://medium.com/](https://medium.com/@gkdemrmelike/evrişimsel-sinir-ağları-convolutional-neural-network-18a05f54bdb6)

URL6. (16.05.2025). [https://bilimfili.com/](https://bilimfili.com/dunyayi-degistirmekte-olan-yapay-sinir-aglari-nedir)

Walczak, S. (2018). Artificial neural networks. In *Encyclopedia of Information Science and Technology, Fourth Edition* (pp. 120-131). IGI Global Scientific Publishing

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y. G., & Li, S. N. (2022, June). M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 615-623).

Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*.

Xuan, X., Peng, B., Wang, W., & Dong, J. (2019, October). On the generalization of GAN image forensics. In *Chinese conference on biometric recognition* (pp. 134-141). Cham: Springer International Publishing.

Yang, X., Li, Y., & Lyu, S. (2019, May). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 8261-8265). IEEE.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354-7363). PMLR.

Zhang, J., Ni, J., & Nie, F. (2025). DSM: Domain Shift Modeling for general deepfake detection. *Signal Processing*, *230*, 109822.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2185-2194).