Deepfake Detection

ÖZET

Deepfake uygulamaları son yıllarda çok fazla dikkat çeken konulardan biridir. Sosyal ağların artan kullanımı neticesinde cihazların kameraları ile oluşturulan fotoğraf ve videoların sahtecilik düzeyinde değiştirilmesi olayı suç düzeyine ulaştırmıştır. Deepfake tekniği ile oluşturulan ve sosyal ağlarda dağıtımı yapılan birçok sahte görüntü ve video sadece kişilerin özel hayatı ile birlikte toplum huzurunu da tehdit etmektedir. İnsan yüzü, insanlar arası etkileşimde ve biyometrik doğrulama sistemlerinde önemli bir role sahiptir. Bu çalışmada, Deepfake tespit modelinin oluşturulmasında bir sınıflandırma yaklaşımıyla ilerlenmiştir. Öznitelik çıkarıcı olarak EfficientNet model ailesi kullanılmıştır.

1. Deepfake Nedir?

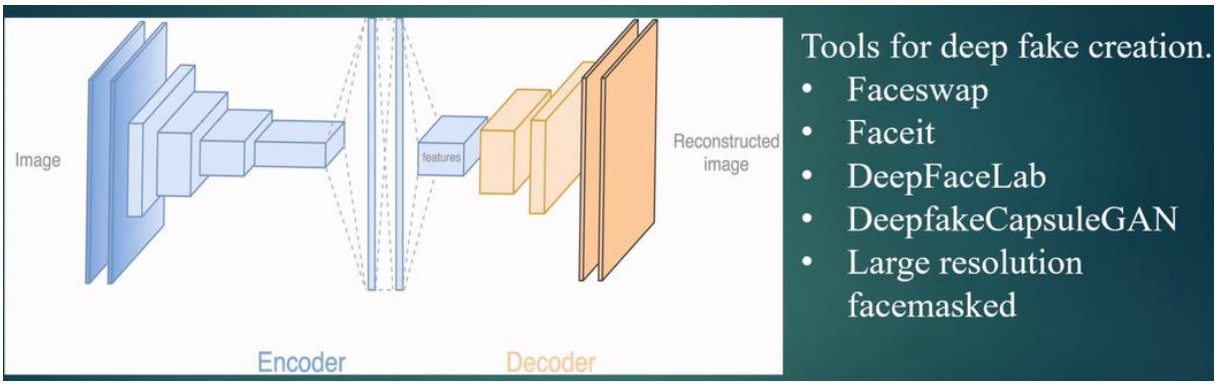
Deepfake, mevcut bir görüntü veya videoda yer alan bir kişinin, yapay sinir ağları kullanarak bir başka kişinin görüntüsü ile değiştirildiği bir medya türüdür. Sıklıkla, otomatik kodlayıcılar ve üretken çekişmeli ağlar (GAN'lar) olarak bilinen makine öğrenme tekniklerini kullanarak mevcut medyanın kaynak medya üzerinde birleştirilmesi ve üst üste konması ile üretilirler. Deepfake, "deep learning" (derin öğrenme) ve "fake" (sahte) kelimelerden türetilmiş bir birleşik kelimedir. [1] Deepfake videoları, bazı durumlarda gerçek kişilerin sesleri de kullanılarak oluşturulabilir.

Deepfake ile yapılan bu tür sahteciliklerin tespit edilmesi, son zamanlarda bir araştırma alanı olarak dikkatleri üzerine çekmiştir. Konvolüsyonel sinir ağları, deepfake sahtekarlıklarını tespit etmede ve bunlardan kaçınmada özellikle etkili olan derin öğrenme tekniklerinin temeli olarak karşımıza çıkmaktadır. Günümüzde internet ve sosyal medya platformları, hızla yayılan ve insanların hayatlarını olumsuz yönde etkileyen sahte videoların yayılmasına olanak tanıyor. Deepfake olarak adlandırılan bu sahte videolar, yapay zeka teknolojileri sayesinde oluşturuluyor ve gerçekçi görünüyorlar. Bu nedenle, deepfake tespiti üzerine yapılan çalışmalar son derece önemli hale geldi.

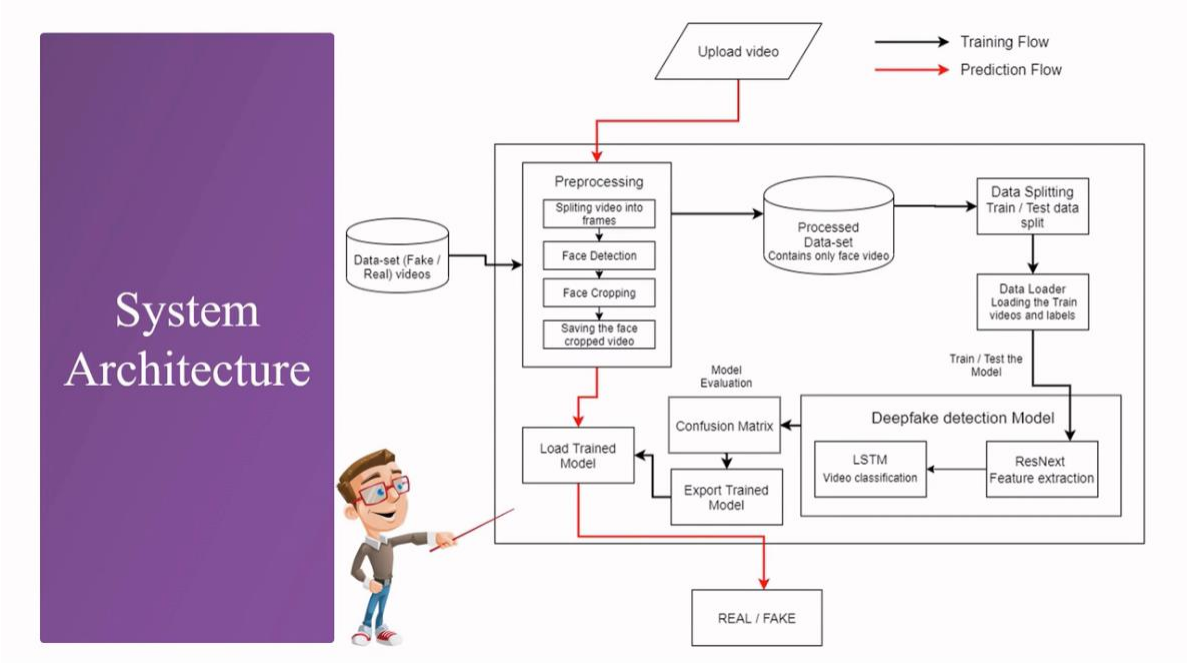
Deepfake videoları, hem eğlence amaçlı hem de kötü amaçlı kullanılabilirler. Örneğin, bir kişi, kendi yüzünü bir ünlü yüzü ile değiştirerek komik bir video oluşturabilir. Ancak deepfake videoları, kötü niyetli kişiler tarafından da kullanılabilir. Örneğin, bir siyasi liderin yüzü, yapay olarak oluşturulan bir yüz ile değiştirilerek bir propaganda videosu oluşturulabilir.

2. Deepfake Nasıl Oluşturulur?

Deepfake oluşturma, genellikle bir kodlayıcı-çözücü ağı (encoder-decoder network) kullanılarak gerçekleştirilir. İlk adımda, hedef kişinin yüz verileri toplanır ve ön işleme adımında işlenir. Kodlayıcı aşamasında, hedef yüzü düşük boyutlu bir latent vektör olarak temsil eder. Çözücü aşamasında, latent vektör kullanılarak orijinal yüz görüntüsü yeniden oluşturulur. Deepfake görüntüsü düzenlenir ve hedef videolara veya görüntülere entegre edilir. Bu teknoloji, etik ve yasal sorunları da beraberinde getirir.



Resim 1 Deepfake Tools



Resim 2 Deepfake System Mimarisi

3. Deepfake Tespitinin Yöntemleri

Deepfake tespiti için birçok yöntem kullanılabilir. Bunlar arasında, yapay zeka teknolojileri, fotoğraf ve video analizi ve diğer teknikler yer alır. Aşağıda, deepfake tespiti için kullanılan bazı yöntemler açıklanmıştır:

3.1. Yapay Zeka Teknolojileri

Deepfake videoları, yapay zeka teknolojileri kullanılarak oluşturulduğu için, yapay zeka teknolojileri de deepfake tespiti için kullanılabilir. Örneğin, bir yapay zeka modeli, gerçek ve deepfake videolarını karşılaştırarak deepfake videolarını tespit edebilir.

Yapay zeka modelleri, deepfake videolarını tespit etmek için farklı özellikler kullanır. Örneğin, bazı modeller, yüz hatlarının hareketlerini analiz ederek deepfake videolarını tespit edebilir. Diğer modeller ise, deepfake videolarının oluşturulduğu yapay yüzlerin özelliklerini analiz eder ve gerçek yüzlerden farklılıkları belirleyerek deepfake videolarını tespit eder.

Yapay zeka teknolojileri, deepfake videolarını tespit etmek için çok etkili olabilir. Ancak, yapay zeka modelleri de yanlış sonuçlar üretebilir. Örneğin, bazı modeller, gerçek videoları deepfake olarak tanıyabilir. Bu nedenle, deepfake tespiti için yapay zeka teknolojileri, diğer yöntemlerle birlikte kullanılmalıdır.

3.2. Fotoğraf ve Video Analizi

Deepfake videolarının oluşturulduğu yüzler, genellikle gerçek yüzlerin fotoğraflarından veya videolarından alınır. Bu nedenle, deepfake tespiti için fotoğraf ve video analizi de kullanılabilir.

Fotoğraf ve video analizi, deepfake videolarının oluşturulduğu yüzlerin özelliklerini belirleyebilir. Örneğin, deepfake videoları oluşturmak için kullanılan yüzlerin gözlerinin pozisyonları, burun şekilleri ve ağız hareketleri, gerçek yüzlerden farklılık gösterir. Bu nedenle, deepfake videolarının oluşturulduğu yüzlerin özellikleri analiz edilerek deepfake videoları tespit edilebilir.

3.3. Ses Analizi

Deepfake videoları, bazı durumlarda gerçek kişilerin sesleri de kullanılarak oluşturulabilir. Bu nedenle, deepfake tespiti için ses analizi de kullanılabilir.

Ses analizi, deepfake videolarının seslerini analiz ederek deepfake videolarını tespit edebilir. Örneğin, deepfake videoları oluşturmak için kullanılan sesler, gerçek kişilerin seslerinden farklılık gösterir. Bu nedenle, deepfake videolarının sesleri analiz edilerek deepfake videoları tespit edilebilir.

4. Deepfake Tespitinde Zorluklar

Deepfake tespiti, birçok zorlukla karşılaşabilir. Örneğin, deepfake videoları, gerçek videolara oldukça benzer görünebilirler. Ayrıca, deepfake videoları, gerçek videolardan daha gerçekçi görünebilirler.

Deepfake tespiti için kullanılan teknolojilerin de bazı sınırlamaları vardır. Örneğin, yapay zeka modelleri, bazı durumlarda yanlış sonuçlar üretebilir. Ayrıca, deepfake videoları, sürekli olarak gelişen yapay zeka teknolojileri tarafından daha gerçekçi hale getirilebilirler.

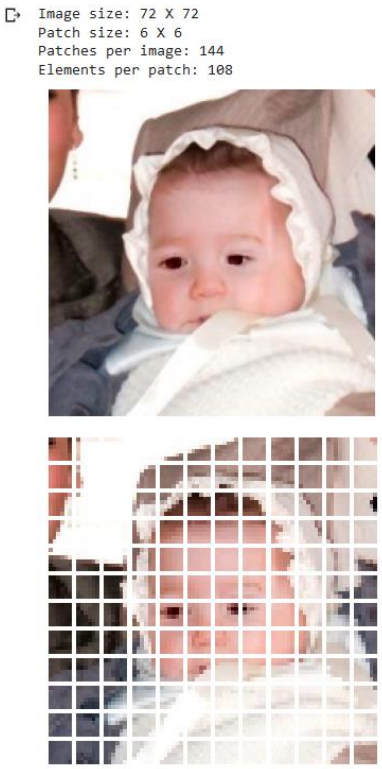
5. Deepfake tespiti için kullanılan teknolojilerin bir diğer zorluğu, deepfake videolarının hızlı bir şekilde oluşturulabilmesidir. Deepfake videoları, birkaç saat içinde oluşturulabilirler. Bu nedenle, deepfake videolarının hızlı bir şekilde tespit edilmesi gerekmektedir.

Deepfake tespiti için kullanılan teknolojilerin diğer bir sınırlaması da deepfake videolarının oluşturulduğu platformların sürekli olarak değişmesidir. Deepfake videoları, herhangi bir video düzenleme yazılımı veya uygulama ile oluşturulabilirler. Bu nedenle, deepfake tespiti için kullanılan teknolojilerin, her türlü platformda kullanılabilmesi gerekmektedir.

5. Vision Transformer: Deepfake Tespiti

Vision Transformer (ViT), video veya görüntülerdeki görsel ve uzamsal-zamansal ipuçlarını incelemek için derin öğrenme algoritmalarının, özellikle Transformers'ın gücünden yararlanan son teknoloji bir deepfake algılama modelidir. Başlangıçta doğal dil işleme görevleri için tasarlanan orijinal Transformer modelinin bir uzantısı olarak geliştirilen ViT, görsel analiz için uyarlanmıştır ve manipüle edilmiş medyayı belirlemede dikkate değer bir performans sergilemiştir.

ViT, video karelerini veya görüntüleri daha küçük parçalara bölerek ve bunları öz dikkat mekanizmaları kullanarak kodlayarak çalışır. Öz dikkat, modelin farklı yamalar arasındaki ilişkileri ve bağımlılıkları analiz etmesine izin vererek hem yerel hem de küresel özellikleri etkili bir şekilde yakalamasını sağlar. Bu kodlanmış yamalar daha sonra giriş verilerinden ayırt edici özellikleri çıkarmayı öğrenen çok sayıda dönüştürücü blok katmanı aracılığıyla işlenir.



Resim 3 Kod Çıktısı

ViT’nin dikkate değer bir yönü, uzay-zamansal bilgileri yakalama yeteneğidir. Model, zamansal dikkat mekanizmalarını dahil ederek, ardışık çerçeveler arasındaki zamansal bağımlılıkları dikkate alarak görsel bağlam anlayışını geliştirebilir. Bu zamansal farkındalık, derin sahte içeriğe özgü anormallikleri ve tutarsızlıkları tespit etmede çok önemlidir.

6. Eğitim ve Veri Kümesi:

ViT'yi eğitmek, hem gerçek hem de deepfake videolar veya görüntülerden oluşan büyük bir veri kümesi gerektirir. Bu veri seti, modelin sağlamlığını ve genelleme yeteneklerini sağlamak için çok çeşitli bireyleri, aydınlatma koşullarını, arka planları ve yüz ifadelerini kapsamalıdır. Eğitim sırasında model, tipik olarak denetimli öğrenme tekniklerine dayanan bir nesnel işlevi optimize ederek gerçek ve manipüle edilmiş medya arasında ayrım yapmayı öğrenir.

ViT, derin sahte algılama alanında çeşitli avantajlar sunar. İlk olarak, dikkat mekanizmaları, gerçek ve manipüle edilmiş içerik arasında ayrım yapma yeteneğini geliştirerek, ilgili bölgelere ve uzay-zamansal ilişkilere odaklanmasını sağlar. Ek olarak, ViT önceki yaklaşımlara kıyasla üstün performans sergileyerek derin sahtekarlıkların yayılmasıyla mücadelede etkili bir araç olma potansiyelini göstermiştir.

Ayrıca, ViT 'nin mimarisi, onu gerçek zamanlı veya büyük ölçekli uygulamalar için uygun hale getirerek, verimli bir şekilde ölçeklenmesini sağlar. Fotoğrafları ve videoları kare kare işleme yeteneği, uzun veya karmaşık videoların bile etkili bir şekilde analiz edilebilmesini sağlar.

6.1 Literatür Taraması

Yukarıda bahsedildiği gibi, otomatik kodlayıcılar ve üretken rakip ağlar, bir görüntünün veya videonun fotogerçekçi içeriğinin tamamını veya bir kısmını sentezlemek için kullanılan derin öğrenmeye dayalı üretici modellerin son iki örneğidir. Bu nedenle, önerdiğimiz metodolojiye ve uygulamaya geçmeden önce, şu anda deepfake görüntüler oluşturmak için uygulanan temel teknolojileri anlamak önemlidir.

6.2 Generative Adversarial Networks (GAN)

Üretken Çekişmeli Ağlar (GAN'lar), makine öğrenimi ve bilgisayar görüşünde öne çıkan ve etkili bir araştırma alanı olarak ortaya çıkmıştır. Ian Goodfellow ve meslektaşları tarafından 2014 yılında tanıtılan GAN'lar, iki sinir ağından oluşan bir çerçevedir: bir üretici ve bir ayrımcı. Üretici, gerçek verilere benzeyen sentetik örnekler üretmeyi amaçlarken, ayrımcı gerçek ve sahte örnekleri birbirinden ayırmayı amaçlar. İki ağ, oyun benzeri bir çekişme sürecine giriyor ve üretici son derece ikna edici çıktılar üretene kadar yeteneklerini sürekli olarak geliştiriyor. Başlangıcından bu yana GAN'lar, görüntü sentezi, stil aktarımı, görüntüden görüntüye çeviri ve video oluşturma dahil olmak üzere çeşitli uygulamalarda önemli gelişmeler gördü.

6.3 Autoencoders

Otomatik kodlayıcılar, denetimsiz öğrenme alanında temel bir kavramdır ve makine öğrenimi topluluğunda büyük ilgi görmüştür. Sinir ağı mimarileri olarak işlev gören otomatik kodlayıcılar, girdi verilerinin daha düşük boyutlu gizli bir alana kodlanması ve ardından orijinal biçimine geri kodunun çözülmesi yoluyla verimli temsillerini öğrenmek için tasarlanmıştır. Otomatik kodlayıcının kodlayıcı kısmı, girdi verilerini sıkıştırarak en göze çarpan özellikleri çıkarırken, kod çözücü kısmı, verileri sıkıştırılmış gösterimden yeniden oluşturur. Bu süreç, otomatik kodlayıcıların veriler içindeki temel kalıpları ve yapıları yakalamasına ve öğrenmesine olanak tanıyarak boyut azaltma, anormallik algılama ve veri oluşturma gibi görevlerde onları etkili kılar. Ayrıca, otomatik kodlayıcılar, yeniden yapılandırılmış çıktıyı orijinal girdiye çok yakın olmaya zorlayarak, verilerden istenmeyen gürültüyü veya yapay oluşumları gidermek ve gürültüyü gidermek için güçlü araçlar görevi görebilir. Otomatik kodlayıcıların çok yönlülüğü ve etkinliği, onları bilgisayar görüşü, doğal dil işleme ve öneri sistemleri dahil olmak üzere çeşitli alanlarda değerli bir araç haline getirdi.

6.4 Variational Autoencoder (VAE)

Varyasyonel Otomatik Kodlayıcılar (VAE'ler), hem otomatik kodlayıcıların hem de olasılıksal modellemenin unsurlarını birleştiren önemli bir üretken modeller sınıfı olarak ortaya çıkmıştır. VAE'ler, veri dağılımındaki anlamlı ve sürekli değişimleri yakalayan girdi verilerinin gizli temsillerini öğrenmek için tasarlanmıştır. Geleneksel otomatik kodlayıcılardan farklı olarak, VAE'ler, girdi verilerini gizli uzaydaki bir dağılıma eşleyen bir kodlayıcı ağı dahil ederek olasılıksal bir değişiklik getirir. Bu dağılım, ortalama ve varyans parametreleriyle tipik olarak Gauss dağılımıdır. Eğitim sırasında VAE'ler, kodlanmış örneklerin genellikle bir Gauss birimi olan önceki bir dağılımı takip etmesini sağlamak için gizli alanı optimize etmeyi amaçlar. Bu düzenlileştirme, modeli öğrenilen gizli alandan örnek alarak ve bunları kod çözücü ağı aracılığıyla çözerek yeni örnekler oluşturmaya teşvik eder. VAE'ler, olasılık öğelerini birleştirerek, yeni veri örnekleri oluşturma, mevcut örnekler arasında enterpolasyon yapma ve gizli alan manipülasyonları gerçekleştirme gibi çeşitli uygulamaları etkinleştirir. VAE'lerin esnekliği ve anlamlılığı, onları görüntü oluşturma, metin sentezi ve anormallik tespiti gibi alanlarda değerli araçlar haline getirdi.

6.5 Rakip Otomatik Kodlayıcılar

Görüntüdeki sahte kısımlar, içsel istatistiksel bilgiler kullanılarak bulunabilir [3]. GAN tarafından oluşturulan sahte resimler, düşük boyutlu rastgele vektörden oluşturulur. Bu nedenle, pasif görüntü sahteciliği dedektörleri bunları tespit etmek için kullanılamaz. Özellikle, GAN'lar tarafından üretilen sahte görüntüler, kaynak fotoğraflarından değiştirilmeden kalır. Çok sayıda tanıma görevinde yaygın olarak kullanıldıklarından, GAN'lar tarafından üretilen sahte resimleri belirlemek için derin bir sinir ağı kullanabiliriz. Son zamanlarda, hileli resim tanımlama için denetimli öğrenmeye dayalı derin öğrenme yaklaşımları araştırılmaktadır. Başka bir deyişle, sahte görüntülerin belirlenmesi ikili bir sınıflandırma sorunu (yani, sahte veya gerçek görüntü) olarak ele alınmıştır.

7. Deney Kurulumu

Bu çalışmanın amacı, çeşitli yüz görüntüleri kullanarak derin sahte görüntüleri belirlemektir. İlk yaklaşımımız, derin sahte tanıma için trafo tabanlı bir mimari kullanır. Bazen ViT olarak adlandırılan Vision Transformer, görüntünün belirli bölümlerine Transformer benzeri bir tasarım uygulayan görüntüleri kategorize etmek için bir metodolojidir. Bir görüntü, her biri doğrusal olarak gömülmüş ve konum yerleştirmeleri eklenmiş sabit boyutlu yamalara bölündükten sonra, ortaya çıkan vektör dizisi, geleneksel bir Transformer kodlayıcıya sağlanır. Sınıflandırma yapmak için, diziye ek, öğrenilebilir bir "sınıflandırma belirteci" dahil etme geleneksel yaklaşımı kullanılır. İkinci yaklaşımımız, Şekil 2'de gösterildiği gibi, karşılaştırmalı bir çalışma için görüntü dönüştürücü yerine sınıflandırıcı olarak Efficient net B07 mimarisinin kullanıldığı ilk boru hattımızın bir varyasyonudur.

8. Değerlendirme Metrikleri

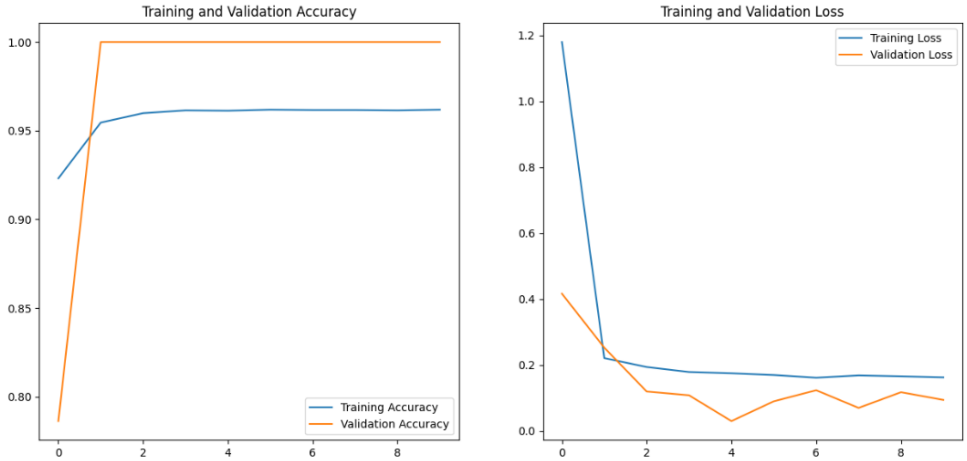
Çalışma, sınıflandırıcıların doğruluğunu ölçmek için çeşitli ölçütler çıktısı verecektir. Kesinlik (accuracy) ve geri çağırma (recall), sınıflar ciddi şekilde dengesiz olduğunda tahmin başarısı için yararlı ölçütleridir. Kesinlik ölçümleri (accuracy), bilgi almadaki ilgililikten kaynaklanırken, geri çağırma (recall), kaç tane alakalı sonucun döndürüldüğünü ölçer.

	precision	recall	f1-score	support
Healthy(class 0)	0.00	0.00	0.00	200
UnHealthy(class 1)	0.15	1.00	0.26	35
accuracy			0.15	235
macro avg	0.07	0.50	0.13	235
weighted avg	0.02	0.15	0.04	235

Accuracy: 14.893617021276595

Resim 2 Kod Çıktısı

Kesinlik-geri çağırma eğrisi, çeşitli eşikler için kesinlik ve geri çağırma arasındaki dengeyi gösterir.

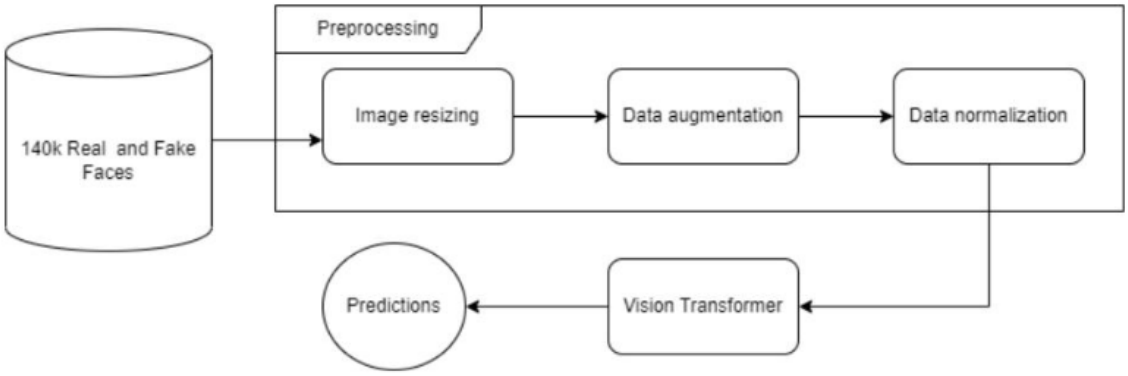


Resim 3 Kod Çıktısı

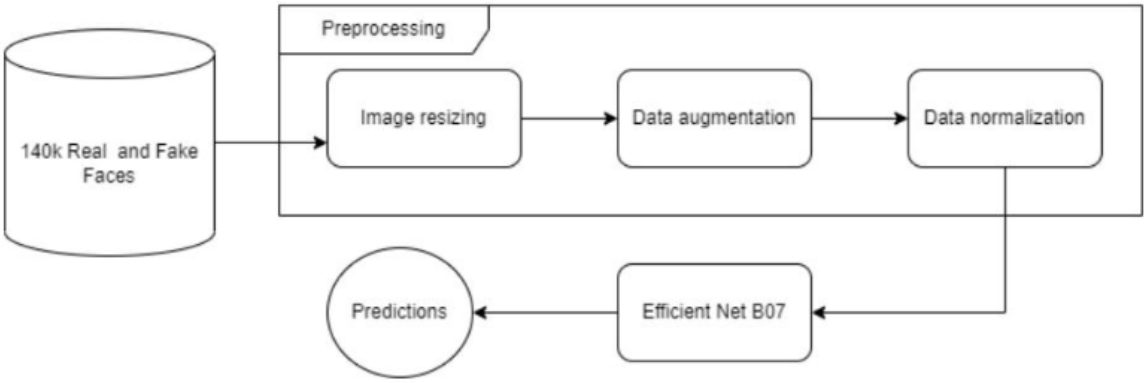
Yüksek doğruluk, düşük bir yanlış pozitif oranı gösterir ve yüksek geri çağırma, düşük bir yanlış negatif oranı gösterir; eğrinin altında önemli bir alan, iyi hatırlama ve kesinlik anlamına gelir. Yüksek puanlar, sınıflandırıcının doğru sonuçlar verdiğini ve çoğu olumlu sonucun olumlu olduğunu gösterir.

9. Metodoloji

Ön işleme, eğitim ve doğrulama, Şekil 1'de gösterildiği gibi yöntemin üç temel bileşenini oluşturur. Veri setini seçtikten sonra, görüntü dönüştürücü modelini eğitmek için eğitim setindeki görüntüleri alıyoruz. Görüntü ön işleme aşamasında, görüntü verileri normalleştirilir ve homojenleştirme amacıyla boyutlar ayarlanır. Bu araştırma, önceden eğitilmiş herhangi bir model kullanmadan geleneksel bir görüntü transformatörünü eğiterek deepfake tespi etmeyi ve karşılaştırmalı bir çalışma sağlamak için bunu bir Efficient Net B07 CNN mimarisiyle karşılaştırmak amaçlandı.



Şekil 1 Yöntem



Şekil 2 Efficient Net B07 ile revize edilmiş Hali

10. **Çözüm**

Karşılaştırmalı bir çalışma için derin sahte tanıma için iki farklı sınıflandırıcı sunulmuştur. İlk yaklaşımımız, sınıflandırıcı olarak bir görsel dönüştürücünün kullanılmasını sağladı. Buna karşılık, aynı boru hattının bir varyasyonu, derin sahte görüntü tanıma için Efficient net B07 mimarisini kullandı. Çalışma, çeşitli avantajlara ve iyi bir doğruluk puanına sahip iki derin öğrenme yaklaşımı sundu. Görüntü dönüştürücü modeli %80,7 doğruluk elde ederken, CNN tabanlı mimari %81,8 doğruluk elde etmiştir. Bu nedenle, derin sahte fotoğrafları tespit edecek sistemlerin geliştirilmesine yönelik yöntemler, derin sahte görüntülerin tespiti, sınıflandırma ve sahtecilik tespiti için standartlaştırılmış araçlar oluşturmak için de temel oluşturabilir. Bu çalışmada elde edilen sonuçlar yeterli sayılabilir, ancak daha fazla veri ve hiperparametre ayarı ile daha iyi sonuçlar elde edebiliriz

KAYNAKÇA

1. https://tr.wikipedia.org/wiki/Deepfake
2. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial Autoencoders. (2015).
3. Sitara, K., Mehtre, B.M.: Digital Video Tampering Detection: An overview of passive techniques. Digital Investigation. 18, 8–22 (2016).