Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 13(4), 2353-2363, 2023
Journal of the Institute of Science and Technology, 13(4), 2353-2363, 2023

ISSN: 2146-0574, eISSN: 2536-4618

Bilgisayar Mühendisliği / *Computer Engineering* **DOI: 10.21597/jist.1281262**

Araştırma Makalesi / *Research Article*

Geliş tarihi / *Received*: 11.04.2023 **Kabul tarihi / *Accepted*: 16.08.2023**

Atıf İçin: Taşyürek, M. ve Gül, E. (2023). Nesne Tespitinde En Uygun Modelin Seçimi İçin Görüntüler Üzerinde Evrişimli Sinir Ağları ile Çekişmeli Saldırı Tespiti. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 13(4), 2353-2363.

To Cite: Tasyurek, M. & Gul, E. (2023). Adversarial Attack Detection with Convolutional Neural Networks on Images for Selection of the Most Suitable Model in Object Detection. *Journal of the Institute of Science and Technology*, 13(4), 2353-2363.

Nesne Tespitinde En Uygun Modelin Seçimi İçin Görüntüler Üzerinde Evrişimli Sinir Ağları ile Çekişmeli Saldırı Tespiti

Murat TAŞYÜREK[1*], Ertuğrul GÜL[2]

Öne Çıkanlar:
- Çekişmeli saldırı tespiti
- Nesne tespiti için evrişimli sinir ağı seçimi
- YOLO v5 ve Faster R-CNN modellerinin transfer öğrenmeli ve transfer öğrenmesiz eğitimi

Anahtar Kelimeler:
- Nesne tespiti
- Çekişmeli saldırı
- Faster R-CNN
- YOLO v5
- ESA

ÖZET:

Görüntülerdeki nesnelerin yüksek doğrulukta tespit edilmesi gerçek zamanlı uygulamalar başta olmak üzere birçok uygulama alanı için önemli bir konudur. Evrişimli sinir ağları ise son yıllarda nesne tespiti uygulamalarında kullanılan ve yüksek doğrulukta başarılar elde edilebilen derin öğrenme tabanlı yöntemlerdir. Klasik Evrişimli sinir ağları orijinal görüntülerdeki nesneleri yüksek doğruluk tespit edebilmesine rağmen ağların FGSM, PGD ve APGD gibi çekişmeli saldırıların uygulandığı görüntülerde başarımları yetersiz kalabilmektedir. Bu problemin üstesinden gelmek için saldırılı görüntülerde nesne tespiti için farklı modeller ve ön işlemler geliştirilmektedir. Ancak saldırılı ve saldırısız durumlar için modellerin başarımları değişebilmektedir. Bu yüzden saldırının olup olmadığının tespit edilmesi ve duruma göre en başarılı modelin seçilmesi gerekmektedir. Bahsedilen problemi çözmek için bu çalışmada görüntülerde çekişmeli saldırı olup olmadığının evrişimli sinir ağları kullanarak tespit edilmesi gerçekleştirilmektedir. Çalışma kapsamında YOLO v5 ve Faster R-CNN modelleri transfer öğrenmeli ve transfer öğrenmesiz olarak çekişmeli saldırı tespiti görevi için eğitilmiştir. Deneysel sonuçlar transfer öğrenmeli Faster R-CNN modelinin 0.971 f1 skoru ile dört model arasından en başarılı sonucu elde ettiğini göstermektedir.

Adversarial Attack Detection with Convolutional Neural Networks on Images for Selection of the Most Suitable Model in Object Detection

Highlights:
- Adversarial attack detection
- Convolutional neural network selection for object detection
- Training of YOLO v5 and Faster R-CNN models with and without transfer learning

Keywords:
- Object detection
- Adversarial attack
- Faster R-CNN
- YOLO v5
- CNN

ABSTRACT:

Object detection on images with high accuracy is an essential issue for many application areas, especially real-time applications. Convolutional neural networks, on the other hand, are deep learning-based methods that have been used in object detection applications in recent years and have achieved high accuracy. However, although classical convolutional neural networks can detect objects on original images with high accuracy, their performance may be insufficient on images where adversarial attacks such as FGSM, PGD, and APGD are applied. To overcome this problem, different models and pre-processes are developed for object detection on attacked images. However, the performance of the models may vary for attacked and non-attacked cases. Therefore, it is necessary to determine whether the attack occurred and select the most successful model according to the case. To solve the problem mentioned above, detect whether there is an adversarial attack on the images using convolutional neural networks has been performed in this study. Within the scope of the study, YOLO v5 and Faster R-CNN models were trained for the adversarial attack detection task with and without transfer learning. Experimental results show that the Faster R-CNN model with transfer learning achieved the most successful result among the four models with an f1 score of 0.971.

[1]Murat TAŞYÜREK (**Orcid ID:** 0000-0001-5623-8577), Kayseri Üniversitesi, Mühendislik Mimarlık ve Tasarım Fakültesi, Bilgisayar Mühendisliği Bölümü, Kayseri, Türkiye
[2]Ertuğrul GÜL (**Orcid ID:** 0000-0002-5591-3435), Kayseri Üniversitesi, Mühendislik Mimarlık ve Tasarım Fakültesi, Yazılım Mühendisliği Bölümü, Kayseri, Türkiye

***Sorumlu Yazar/*Corresponding Author**: Murat TAŞYÜREK, e-mail: murattasyurek@kayseri.edu.tr

GİRİŞ

Nesne tespiti, görüntülerdeki araç, insan, kedi, yüz ve tabela gibi bilinen bir veya birden fazla sınıftaki nesne örneklerinin tespit edilmesidir. Görüntülerde genellikle az sayıda nesne bulunur, ancak bu nesnelerin çok sayıda olası konumu ve nesnelerin görüntü üzerindeki çok sayıda olası ölçeği bulunmaktadır (Amit ve ark., 2020). Son yıllarda nesne tespiti uygulamalarının tıp (Shelatkar ve ark., 2022; Terzi ve Terzi, 2022), askeri (Du ve ark., 2022; Liu ve ark., 2022) ve endüstri (Wang ve ark., 2021; Guo ve ark., 2022) gibi birçok alanda kullanımı büyük oranda artmıştır. Bu yüzden özellikle gerçek zamanlı uygulamalarda nesnelerin doğru olarak tespit edilmesi büyük önem taşımaktadır.

Evrişimli sinir ağları nesne tespiti için kullanılan yöntemlerin başında gelmektedir (Gu ve ark., 2022). Evrişimli sinir ağları, bilgisayarla görü ve görüntü işleme gibi uygulama alanlarında yaygın olarak kullanılan derin öğrenme yöntemleridir. Evrişimli sinir ağları nesne tespitinin yanı sıra sınıflandırma ve segmentasyon uygulamaları içinde sıkça kullanılmaktadır (Längkvist ve ark., 2016; Balamurugan ve Gnanamanoharan, 2023).

Son yıllarda internet teknolojilerinin hızla gelişmesi ile dijital görüntülerin kullanımı ve dağıtımı hızla artmış, görüntülere erişim kolaylaşmıştır. Bu yüzden görüntülerin paylaşılması veya dağıtılması sırasında çekişmeli saldırı gibi istenmeyen birçok manipülasyon ile karşılaşılabilmektedir. Ancak, klasik evrişimli sinir ağları orijinal görüntüler üzerinde başarılı sonuçlar elde ederken saldırıya uğrayan görüntüler üzerinde nesne tespitini yeterli doğrulukta gerçekleştiremeye bilmektedir. Çekişmeli saldırı, belirli görevler için eğitilmiş modellerin yanlış tahmin yapmasına neden olabilmektedir. Literatürde modellerin tahminlerinde hata yapması için birçok çekişmeli saldırı önerilmiştir. Fast gradient signed method (FGSM) (Goodfellow ve ark., 2014), Goodfellow ve arkadaşları tarafından, girdi görüntüsünün kayıp gradyanını kullanarak çekişmeli örneğinin oluşturulması için önerilmiştir. Kurakin ve arkadaşları, FGSM'yi küçük bir adım boyutuyla uygulayarak FGSM'nin yinemeli bir sürümü olan Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin ve ark., 2016)'yi önermiştir. Momentum Iterative Fast Gradient Sign Method (MI-FGSM) (Dong ve ark., 2018), I-FGSM'ye momentum terimi eklenerek Dong ve arkadaşları tarafından geliştirilmiştir (Zhang ve ark., 2020). Projected Gradient Descent (PGD) (Madry ve ark., 2017) saldırısı ise Madry ve diğerleri tarafından önerilen FGSM saldırısının güçlü yinelemeli bir versiyonudur. PGD saldırısının adım boyutu ve objektif fonksiyon sorunlarını ele alan ve PGD saldırısının uzantısı olan Auto Projected Gradient Descent (APGD) (Croce ve Hein, 2020) ise, Govindarajulu ve arkadaşları tarafından önerilmiştir.

Saldırı durumuna göre ağların başarımlarının değişmesinden dolayı, nesne tespiti uygulamalarında görüntüde saldırı olup olmadığının tespit edilerek kullanılacak ağın belirlenmesi önemli bir hale gelmiştir. Yapılan son çalışmada, çekişmeli saldırı uygulanmış görüntülerde nesne tespiti için ayrık dalgacık dönüşümü ve gri tonlama tabanlı derin öğrenme yöntemleri önerilmiştir (Tasyurek ve Gul, 2023). Gerçekleştirilen deneylerde, klasik derin öğrenme yöntemleri saldırısız görüntülerde daha iyi sonuçlar üretirken ayrık dalgacık dönüşümü ve gri tonlama tabanlı derin öğrenme yöntemleri çekişmeli saldırı uygulanmış görüntülerde daha iyi sonuçlar üretmiştir. Yapılan çalışmadaki deney sonuçları dikkate alındığında, saldırı durumuna göre derin öğrenme modelinin seçilmesi daha yüksek doğrulukta nesne tespitinin gerçekleştirilebileceğini göstermektedir.

Bu çalışmada, görüntünün türüne göre nesne tespitinde kullanılacak ağların seçimi için evrişimli sinir ağları ile görüntüler üzerinde çekişmeli saldırı uygulanıp uygulanmadığının tespiti gerçekleştirilmiştir. Evrişimli sinir ağları olarak son yıllarda en çok kullanılan derin öğrenme ağlarından olan YOLO v5 (Jocher ve ark., 2020) ve Faster R-CNN (Ren ve ark., 2015) modelleri kullanılmıştır. Faster R-CNN modeli bölge tabanlı modellerin başında gelmekteyken, YOLO ise düşük hesaplama

maliyeti ve yüksek doğruluk oranına sahiptir. Bu yüzden bu iki model çekişmeli saldırı tespitinde kullanılmak üzere evrişimli sinir ağı olarak seçilmiştir. Çalışma kapsamında öncelikle doğal sahne görüntülerinden oluşan veri setindeki görüntülerin belirli bir kısmına FGSM, PGD ve APGD çekişmeli saldırıları ayrı ayrı uygulanmıştır. Böylelikle içerisinde orijinal ve çeşitli çekişmeli görüntülerden oluşan ve çekişmeli saldırı tespiti için kullanılacak veri seti oluşturulmuştur. Veri setinin oluşturulmasından sonra, YOLO v5 ve Faster R-CNN evrişimli sinir ağları çalışma kapsamında transfer öğrenmeli ve transfer öğrenmez olarak ayrı ayrı eğitilmiştir. Eğitilmiş ağlar daha sonra veri kümesindeki test veri seti ile test edilmiştir. Deneyler sonucunda en yüksek başarı oranı 0.971 f1 skoru ile transfer öğrenmeli Faster R-CNN ağında elde edilirken en düşük başarı oranı 0.960 f1 skoru ile transfer öğrenmesiz YOLO v5 ağında elde edilmiştir. Deneysel sonuçlar görüntülerde çekişmeli saldırı olup olmadığının tespitinin yüksek oranda doğruluk ile elde edilebileceğini göstermektedir. Bu yüzden, nesne tespiti için kullanılacak ağ seçiminde bir ön işlem olarak görüntülerin çekişmeli saldırılı olup olmadığının sınıflandırılmasının gerçekleştirilebileceği görülmüştür. Görüntüde saldırı tespit edilirse saldırılı görüntülerde daha yüksek doğruluk ile nesne tespiti gerçekleştiren ve bu görev için eğitilmiş evrişimli sinir ağlar seçilerek daha yüksek başarılar elde edilebilir.

Makalenin organizasyonu şu şekildedir: Materyal ve metot bölümünde kullanılan derin öğrenme modelleri, çekişmeli saldırılar, saldırı tespiti için oluşturulan veri seti, metot ve başarım metrikleri sunulmuştur. Bulgular ve tartışma bölümünde deneysel sonuçlara ve son bölümde ise genel değerlendirme ve olası gelecek çalışmalara değinilmiştir.

MATERYAL VE METOT

Makalenin bu bölümünde çalışmada kullanılan derin öğrenme modelleri, saldırı uygulanmış görüntüleri üretmek için kullanılan çekişmeli saldırılar, derin öğrenme modellerini eğitmek ve test etmek için kullanılan veri seti, ağların başarımlarını ölçmek için kullanılan metrikler ve önerilen yaklaşımdan bahsedilmektedir.

You Only Once Look (YOLO)

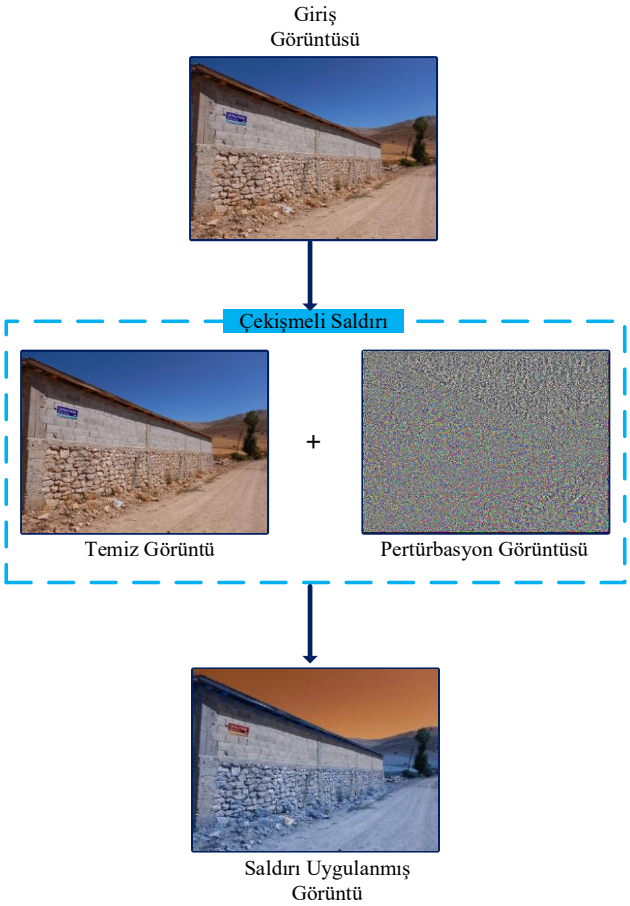
You Only Once Look (YOLO), Redmon ve diğerleri tarafından 2016 yılında hızlı nesne tespiti için geliştirilen evrişimli sinir ağıdır (Redmon ve ark., 2016). Küçük model boyutu, düşük hesaplama maliyeti ve yüksek doğruluk oranı YOLO'nun en önemli avantajlarıdır (Jiang ve ark., 2022). YOLO'nun V2 (Redmon ve Farhadi, 2017), V3 (Redmon ve Farhadi, 2018), V4 (Bochkovskiy ve ark., 2020), V5 gibi birçok versiyonu bulunmaktadır. Her yeni sürüm, önceki modelin hızını ve doğruluğunu artırmak için geliştirilmiştir. Günümüzde en çok tercih edilen YOLO versiyonlarından biri V5 versiyonudur.

Faster R-CNN

Region Based CNN (R-CNN) modeli, birden fazla nesne içeren görüntülerde CNN lokalizasyon probleminin üstesinden gelmek için Girshick ve arkadaşları tarafından geliştirilmiştir. Ancak, R-CNN'nin eğitim ve tahmin süresi, bölgeye dayalı yaklaşımı nedeniyle çok yüksektir. Bu yüzden, Girshick, R-CNN'nin hesaplama maliyeti problemini çözmek Fast R-CNN (Girshick, 2015) modelini önermiştir. Yine de Fast R-CNN modelinin hesaplama maliyeti R-CNN modeline göre düşük olmasına rağmen, birçok uygulama için yeterli değildir. Bu nedenle, Fast R-CNN'nin arama bölgesi yaklaşımının neden olduğu darboğazın üstesinden gelmek için daha az tespit süresine sahip olan Faster R-CNN modeli, Ren ve ark. tarafından geliştirilmiştir (Ren ve ark., 2015). Faster R-CNN modeli son yıllarda tercih edilen bölge tabanlı modellerin başında gelmektedir.

Çekişmeli Saldırı

Çekişmeli saldırılar, modellerin tahminlerinde hata yapmasına neden olmak için kasıtlı olarak uygulanan ve çekişmeli örnekler oluşturmak için kullanılan yöntemlerdir. Çekişme saldırısı ile bozulan test örnekleri, belirli görevler için eğitilmiş modellerin yanlış tahmin yapmasını sağlayabilmektedir. Çekişme saldırılarında, saldırı uygulanmış görüntü pertürbasyonun test örneğine eklenmesi ile oluşturulmaktadır. Çekişmeli saldırı uygulanmış görüntü üretiminin örneği, Şekil 1'de gösterilmektedir. Bu çalışmada FGSM, PGD ve APGD olmak üzere üç farklı çekişmeli saldırı çeşidi kullanılmıştır. Goodfellow ve arkadaşları tarafından önerilen Fast gradient signed method (FGSM) saldırısı, en çok kullanılan çekişmeli saldırılardan biridir (Goodfellow ve ark., 2014). FGSM saldırısı, girdi görüntüsünün kayıp gradyanını kullanarak bir görüntünün rakip bir örneğini oluşturmak için görüntü üzerine uygulanmaktadır. Madry ve diğerleri tarafından önerilen Projected Gradient Descent (PGD) (Madry ve ark., 2017) saldırısı ise temel olarak FGSM saldırısının yinelemeli bir çeşididir (Ayas ve ark., 2022). Giriş, orijinalden başlayarak yinelemeli olarak güncellenir (Liu ve ark., 2019). PGD daha etkili bir saldırı olmasına rağmen, hesaplama açısından FGSM saldırısından daha maliyetlidir. Auto Projected Gradient Descent (APGD) (Croce ve Hein, 2020), PGD saldırısının optimal olmayan adım boyutu ve objektif fonksiyon sorunlarını ele alan bir uzantısıdır (Govindarajulu ve ark., 2023). PGD saldırısının standart formülasyonundaki zayıflıkların üstesinden gelmek için önerilmiştir (Croce ve Hein, 2020).



Şekil 1. Çekişmeli saldırı uygulanmış görüntü üretimi

Veri Seti

Bu çalışmada, nesne tespitinde kullanılacak olan görüntülerin orijinal ya da çekişmeli saldırıya maruz kalıp kalmadığını tespit etmek için bir veri seti oluşturulmuştur. Veri setini oluşturmak için Kayseri Büyükşehir Belediyesi'nden alınan 2.904 adet doğal sahne görüntüsü kullanılmıştır. Bu sahne

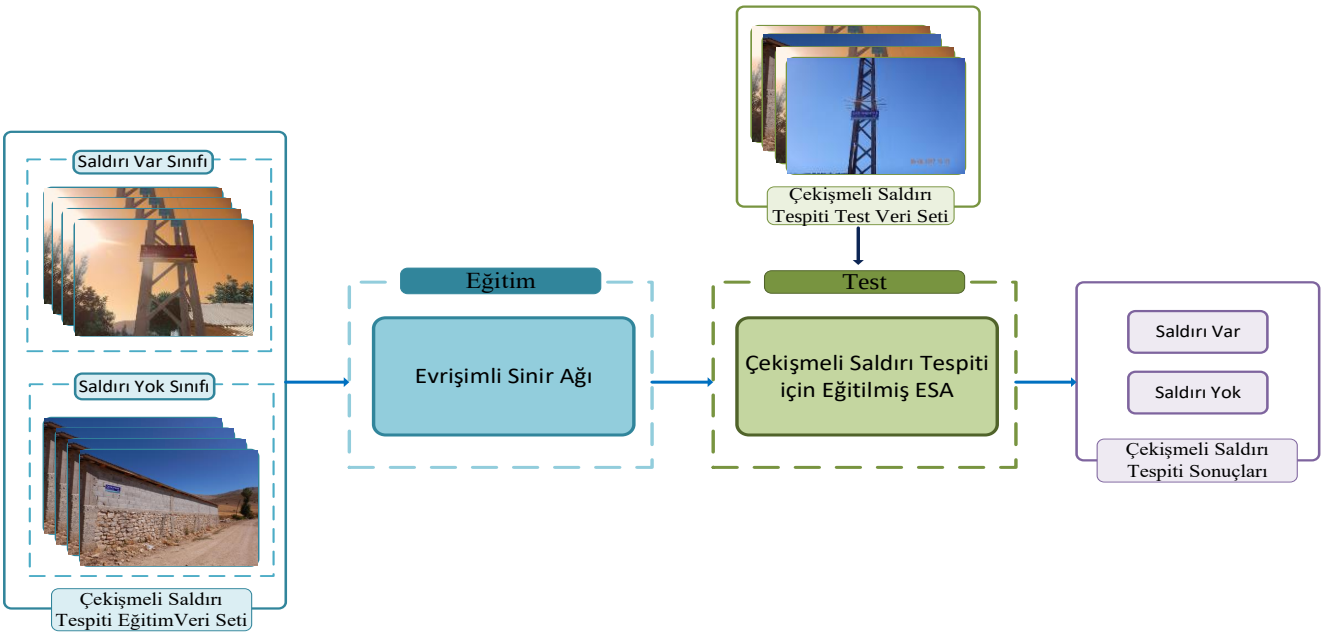
görüntülerinden 2.178 (%75) tanesine çekişmeli saldırı uygulanmıştır. Saldırılmış görüntülerin 726 (%33,33) tanesi FGSM, 726 (%33,33) tanesi PGD ve 726 (%33,33) tanesi APGD kullanılarak üretilmiştir. Veri setindeki orijinal görüntülerden 420 (%57,85) tanesi eğitim verisi olarak kullanılırken, doğrulama ve test verisi olarak sırasıyla 106 (%14,60) ve 200 (%27,54) adet orijinal görüntü kullanılmıştır. Saldırılmış görüntülerden ise 1260 (%57,85) tanesi eğitim verisi olarak kullanılırken, doğrulama ve test verisi olarak sırasıyla 318 (%14,60) ve 600 (%27,54) adet görüntü kullanılmıştır. Toplamda 1680 (%57,85) adet görüntü eğitimde, 424 (%14,60) adet doğrulamada ve 800 (%27,54) adet ise testte kullanılmıştır. Oluşturulan veri seti Çizelge 1'de sunulmuştur.

Çizelge 1. Oluşturulan veri seti

Veri Seti	Eğitim Veri Seti (adet)	Doğrulama Veri Seti (adet)	Test Veri Seti (adet)
Orijinal görüntü sayısı	420	106	200
FGSM uygulanmış görüntü sayısı	420	106	200
PGD uygulanmış görüntü sayısı	420	106	200
APGD uygulanmış görüntü sayısı	420	106	200
Toplam görüntü sayısı	1680	424	800

Metot

Bu çalışmada, nesne tespiti gerçekleştirilecek görüntülerde çekişmeli saldırının olup olmadığının tespit edilmesi amaçlanmaktadır. Bu amaçla, son yıllarda nesne tespiti için sıklıkla kullanılan YOLO v5 ve Faster R-CNN modelleri transfer öğrenmeli ve transfer öğrenmesiz olarak eğitilmiştir. Eğitimlerde önceki bölümde bahsedilen çekişmeli saldırıların tespiti için oluşturulan veri seti kullanılmıştır. Transfer öğrenmeli eğitimde öncelikle farklı bir görev için eğitilmiş olan ön eğitimli YOLO v5 ve Faster R-CNN ağlarının sınıflandırıcı katmanı yerine problemimizdeki iki sınıfa özgü sınıflandırıcı eklenmiştir. Ardından, çekişmeli saldırı uygulanmış görüntüler ve orijinal görüntülerden oluşturulan eğitim veri seti ile ağlar yeniden eğitilmiştir. Son olarak ise transfer öğrenme ile eğitilen ağların başarımları, veri setindeki test verileri kullanılarak gözlemlenmiştir. Transfer öğrenmesiz eğitim için ise transfer öğrenmede olduğu gibi ön eğitimsiz ağların sınıflandırma katmanı probleme özgü olacak şekilde iki sınıf için ayarlanmıştır. Daha sonra, bu ağlar görüntülerde çekişmeli saldırı olup olmadığını tespit etmek için eğitim veri seti ile transfer öğrenmesiz olarak eğitilmiştir. Transfer öğrenmesiz eğitilen ağlar son olarak test veri seti ile test edilmiştir. Çalışmada kullanılan metodoloji Şekil 2'de gösterilmektedir.



Şekil 2. Çekişmeli saldırı tespiti metodolojisi

Performans kriterleri

Derin öğrenme modellerinin nesne tespit başarımını (sınıflandırma) değerlendirmek için doğruluk, duyarlılık, hassaslık ve f1 skor metrikleri yaygın olarak kullanılmaktadır (Ming ve ark., 2021). Bu metrikler true positive (TP), true negative (TN), false positive (FP) ve false negative (FN) değerleri kullanılarak hesaplanmaktadır. TP değeri derin öğrenme modeli tarafından tespit edilen nesne sayısını göstermektedir. Diğer bir ifade ile girdi görüntüde nesne var ve derin öğrenme modeli bu nesneyi tespit etmişse bu işlem TP olarak kabul edilmektedir. TN değeri ise nesne olmadığı ve derin öğrenme modelinin nesne tespit etmediği görüntü sayısını ifade etmektedir. Görüntüde nesne yok ise ve derin öğrenme modeli nesne tespit etmemişse bu durum TN olarak ifade edilmektedir. FP değeri ise nesne olmadığı halde derin öğrenme modellinin tespit ettiği nesne sayısını göstermektedir. Görüntüde nesne olmamasına rağmen derin öğrenme modeli nesne tespit etmektedir ve bu durum FP olarak ifade edilir. FN değeri ise nesne olduğu halde derin öğrenme modelinin tespit edemediği nesne sayısını göstermektedir. Görüntüde nesne olduğu halde derin öğren modeli bu nesneyi tespit edememektedir ve bu durum FN olarak ifade edilir.

Derin öğrenme metriklerinden doğruluk doğru olarak sınıflandırılan nesne sayısının toplam sınıflandırma sayısına oranını gösterir ve Eşitlik 1. ile hesaplanır.

$$Doğruluk = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Duyarlılık metriği, pozitif olarak tahmin edilmesi gereken işlemlerin ne kadarının pozitif olarak tahmin edildiğini gösteren bir metriktir ve Eşitlik 2. ile hesaplanır.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (2)$$

Hassaslık metriği, derin öğrenme modelleri tarafından pozitif olarak değerlendirilen değerlerden gerçekten ne kadarının pozitif olduğunu gösteren bir metriktir ve Eşitlik 3. ile hesaplanır.

$$Hassaslık = \frac{TP}{TP+FP} \quad (3)$$

f1 skor metriği duyarlılık ve hassaslık metriklerinin harmonik ortalamasından oluşmaktadır. f1 skor metriği derin öğrenme yöntemlerinin başarımı değerlendirmek için yaygın olarak kullanılmaktadır (Das ve ark., 2022). f1 skor metriği Eşitlik 4. ile hesaplanır.

$$f1\ skor = \frac{2 \times Duyarlılık \times Hassalık}{Duyarlılık+Hassaslık} \quad (4)$$

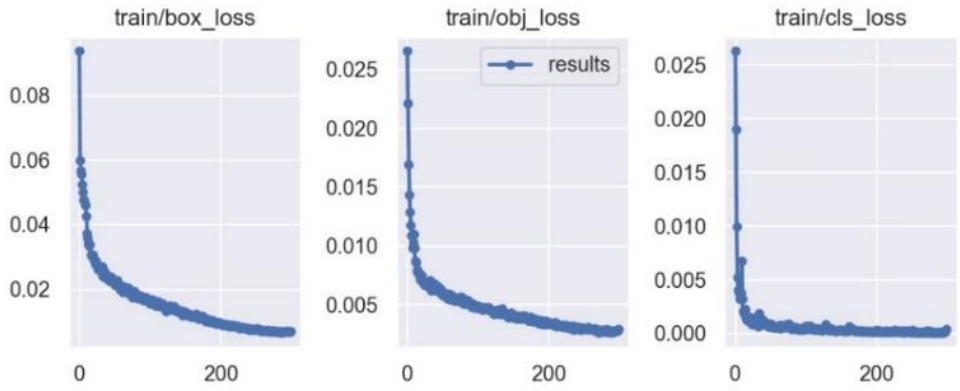
BULGULAR VE TARTIŞMA

Bu bölümde, nesne tespitinde kullanılacak olan görüntülerin çekişmeli saldırıya maruz kalıp kalmadığını tespit etmek için eğitilen ağların deneysel sonuçları sunulmaktadır. Eğitim ve test için ayrılan veri setinde orijinal görüntüler ve çekişmeli saldırı uygulanmış görüntüler bulunmaktadır. Çekişmeli saldırılı görüntüleri oluşturmak için FGSM, PGD ve APGD saldırıları kullanılmıştır. Çalışmada, Evrişimli sinir ağı olarak YOLO v5 ve Faster R-CNN modelleri kullanılmıştır. Modelleri eşit koşullarda test etmek için modellerin PyTorch versiyonları kullanılmıştır ve modeller Python 3.9 kullanılarak eğitilip test edilmiştir. Deneysel çalışmalar i9 12. nesil 3.19 GHz veri yolu hızına sahip, 64

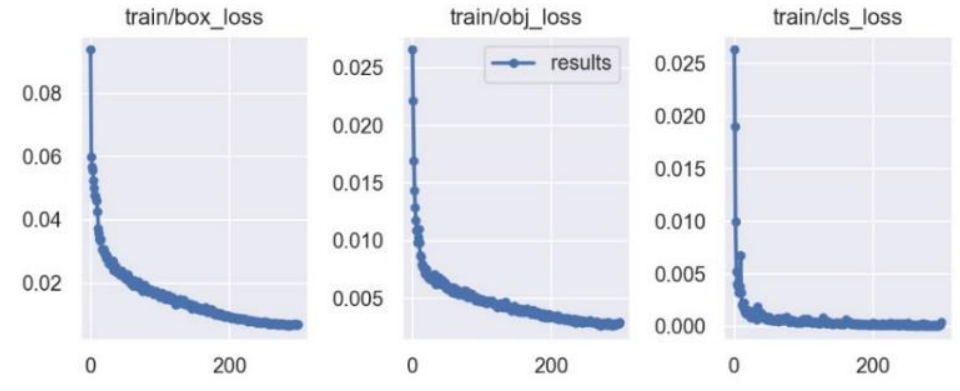
GB 3200 MHz RAM, 2 TB SSD ve 12 GB NVIDIA GeForce ekran kartına sahip bir bilgisayar kullanılarak yapılmıştır.

Modellerin Eğitilmesi

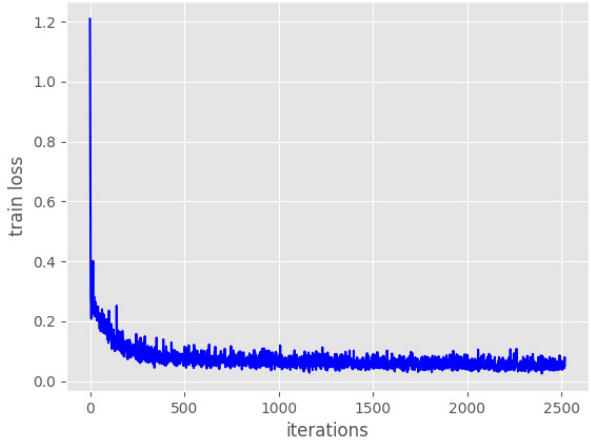
Modellerin eğitimi transfer öğrenmeli ve transfer öğrenmesiz olarak ayrı ayrı gerçekleştirilmiştir. Derin ögrenme modelleri 200 epok eğitilmiştir. Transfer öğrenmeli ve transfer öğrenmesiz eğitilen YOLO v5 modellerinin eğitimler sırasında elde ettiği yitim değerleri sırasıyla Şekil 3 ve 4'te sunulmuştur. Transfer öğrenmeli ve transfer öğrenmesiz eğitilen Faster R-CNN modellerinin eğitimler sırasında elde ettiği yitim değerleri ise sırasıyla Şekil 5 ve 6'da sunulmuştur.



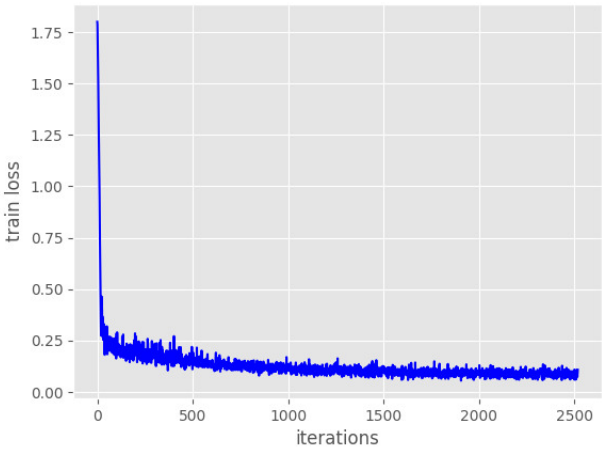
Şekil 3. Transfer öğrenmeli eğitilen YOLO V5 modelinin eğitimdeki yitim değerleri



Şekil 4. Transfer öğrenmesiz eğitilen YOLO V5 modelinin eğitimdeki yitim değerleri



Şekil 5. Transfer öğrenmeli eğitilen Faster R-CNN modelinin eğitimdeki yitim değerleri



Şekil 6. Transfer öğrenmesiz eğitilen Faster R-CNN modelinin eğitimdeki yitim değerleri

Yitim değeri derin ögrenme modellerinin eğitim başarımını göstermektedir (Moustapha ve ark., 2022). Yitim değerlerinin 0'a çok yakın olması derin ögrenme modellerinin başarılı bir şekilde eğitildiğini göstermektedir (Hu ve ark., 2018). Sekil 3-6'daki yitim değerlerinden görüldüğü üzere transfer öğrenmeli ve transfer öğrenmesiz eğitilen YOLO v5 ve Faster R-CNN modellerinin başarılı bir şekilde eğitildiği anlaşılmaktadır.

Deneysel Sonuçlar ve Tartışma

Test verisi üzerinde transfer öğrenme ile eğitilen YOLO v5 ve Faster R-CNN modellerinin saldırı tespiti başarımı performans metriklerine göre incelenmiştir ve elde edilen sonuçlar Çizelge 2'de sunulmuştur. Çizelge 2'de sunulduğu üzere transfer öğrenmeli YOLO v5 ve Faster R-CNN modelleri 200 orijinal ve 600 saldırılı görüntü üzerinde test edilmiştir. Çizelge 2'de transfer öğrenme ile eğitilen YOLO v5 ve Faster R-CNN modelleri sırasıyla 0.965 ve 0.971 f1 skoru elde etmiştir. Çizelgeden Faster R-CNN modelinin daha başarılı olduğu açıkça görülmektedir. Öte yandan, duyarlılık metriğine göre transfer öğrenmeli YOLO v5 modelinin daha başarılı olduğu görülmektedir.

Çizelge 2. Transfer Öğrenmeli Yöntemlerin Başarımları

CNN	Görüntü Türleri	Veri Sayısı	TP	TN	FP	FN	Doğruluk	Duyarlılık	Hassaslık	f1 skor
YOLO v5	Orijinal	200	183	0	0	17	0.915	1.000	0.915	0.956
YOLO v5	Saldırılı	600	555	0	0	37	0.938	1.000	0.938	0.968
YOLO v5	Orijinal ve Saldırılı	800	738	0	0	54	0.932	1.000	0.932	0.965
Faster R-CNN	Orijinal	200	189	0	1	10	0.945	0.995	0.950	0.972
Faster R-CNN	Saldırılı	600	566	0	7	27	0.943	0.988	0.954	0.971
Faster R-CNN	Orijinal ve Saldırılı	800	755	0	8	37	0.944	0.990	0.953	0.971

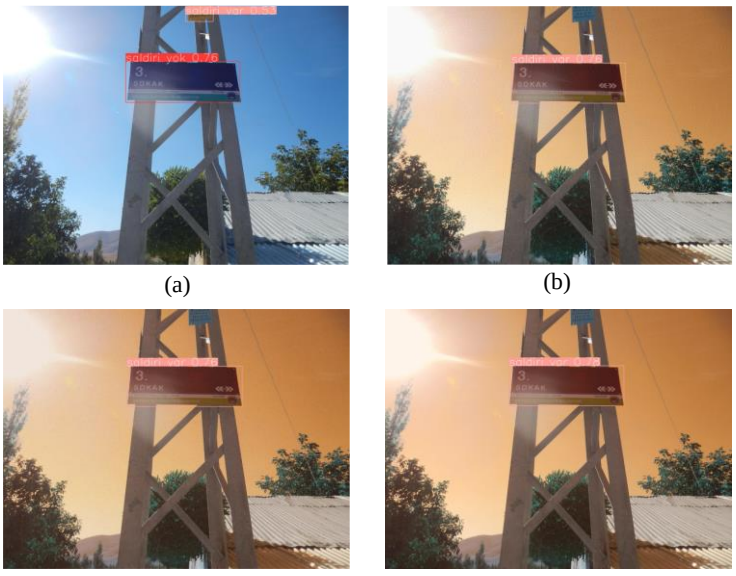
Test verisi üzerinde transfer öğrenmesiz eğitilen YOLO v5 ve Faster R-CNN modellerinin saldırı tespiti başarımları ise Çizelge 3'te gösterilmiştir. Çizelge 3'te görüldüğü üzere transfer öğrenmesiz eğitilen YOLO v5 ve Faster R-CNN modelleri sırasıyla 0.960 ve 0.964 f1 skoru elde etmiştir. Çizelgeden Faster R-CNN modelinin transfer öğrenmesiz eğitim sonucunda YOLO v5 modeline göre daha başarılı olduğu açıkça görülmektedir. Diğer yandan, duyarlılık metriğine göre ise transfer öğrenmesiz YOLO v5 modelinin daha başarılı olduğu görülmektedir.

Çizelge 3. Transfer Öğrenmesiz Yöntemlerin Başarımları

CNN	Görüntü Türleri	Veri Sayısı	TP	TN	FP	FN	Doğruluk	Duyarlılık	Hassaslık	f1 skor
YOLO v5	Orijinal	200	183	0	0	17	0.915	1.000	0.915	0.956
YOLO v5	Saldırılı	600	555	0	0	45	0.925	1.000	0.925	0.961
YOLO v5	Orijinal ve Saldırılı	800	738	0	0	62	0.923	1.000	0.923	0.960
Faster R-CNN	Orijinal	200	191	0	4	9	0.936	0.979	0.955	0.967
Faster R-CNN	Saldırılı	600	557	0	11	32	0.928	0.981	0.946	0.963
Faster R-CNN	Orijinal ve Saldırılı	800	748	0	15	41	0.930	0.980	0.948	0.964

Çizelge 2 ve 3 incelendiğinde test veri seti üzerinde en başarılı sonucu 0.971 f1 skoru ile transfer öğrenme ile eğitilen Faster R-CNN modeli elde etmiştir. Transfer öğrenmeli YOLO v5 modeli ise 0.965 f1 skoru ile ikinci en başarılı sonucu elde etmiştir. Ayrıca, transfer öğrenmeli YOLO v5 ve Faster R-CNN modelleri transfer öğrenmesiz eğitilen modellerine kıyasla daha iyi sonuçlar elde etmişlerdir.

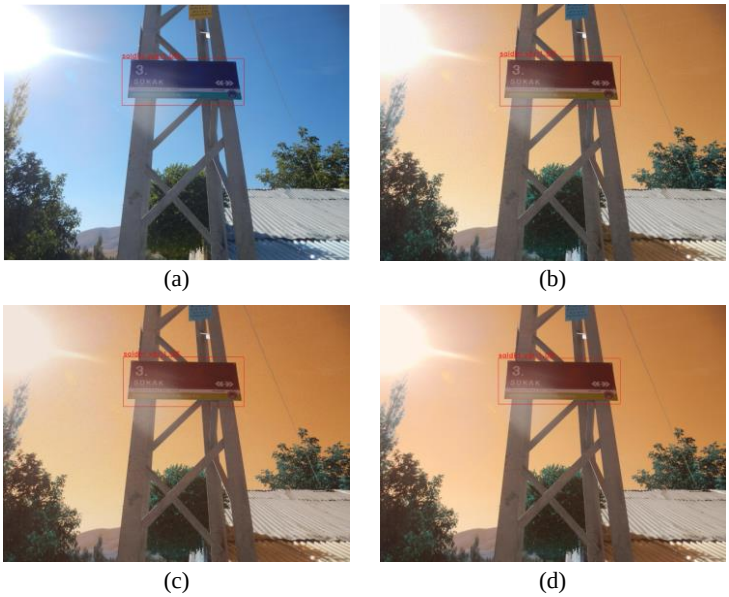
Yöntemlerin başarımını daha iyi irdelemek için bir test görüntüsü üzerinde transfer öğrenmeli YOLO v5 ve Faster R-CNN modellerinin saldırı tespitleri gerçekleştirilmiştir. Tespit sonuçları sırasıyla Şekil 7 ve 8'de gösterilmektedir. Şekil 7a ve 8a'da ağların orijinal görüntüdeki saldırı tespiti sonuçları gösterilirken Şekil 7b-d ve 8b-d'de ağların sırasıyla FGSM, PGD ve APGD saldırıları uygulanmış görüntülerdeki tespitleri gösterirmiştir. Şekillerden görüldüğü üzere transfer öğrenmeli eğitilen YOLO v5 ve Faster R-CNN modelleri görüntüde saldırı olup olmadığını başarılı bir şekilde tespit etmiştir. Ayrıca şekillerden Faster R-CNN modelinin YOLO v5 modeline kıyasla daha yüksek güven skoru ile saldırı tespiti yaptığı açıkça görülmektedir.



(a) (b)

(c) (d)

Şekil 7. Transfer öğrenmeli eğitilen YOLO v5 modelinin örnek test görüntülerinde saldırı tespiti sonucu: (a) orijinal görüntü, (b) FGSM uygulanmış görüntü, (c) PGD uygulanmış görüntü (d) APGD uygulanmış görüntü



(a) (b)

(c) (d)

Şekil 8. Transfer öğrenmeli eğitilen Faster R-CNN modelinin örnek test görüntülerinde saldırı tespiti sonucu: (a) orijinal görüntü, (b) FGSM uygulanmış görüntü, (c) PGD uygulanmış görüntü (d) APGD uygulanmış görüntü

SONUÇ

Bu çalışmada, nesne tespitinin daha yüksek doğrulukta yapılabilmesi için nesne tespitinde kullanılacak görüntülerin çekişmeli saldırıya maruz kalıp kalmadığı tespit edilmiştir. Çalışma, nesne tespitinde kullanılmak üzere seçilecek olan derin öğrenme modellerinin belirlenmesine yardımcı olmak amacıyla gerçekleştirilmiştir. Bu amaçla çalışma kapsamında, orijinal ve saldırılı görüntülerden oluşan veri seti kullanılarak YOLO v5 ve Faster R-CNN modelleri transfer öğrenmeli ve transfer öğrenmesiz olarak eğitilmiştir. Eğitilen ağlar daha sonra test verisi kullanılarak test edilmiştir. Deneysel sonuçlar eğitilen modellerin yüksek başarı ile görüntülere çekişmeli saldırıların (FGSM, PGD ve APGD) uygulanıp uygulanmadığını tespit edebildiğini göstermiştir. Deneysel sonuçlara göre en yüksek başarıyı 0.971 f1 skoru ile transfer öğrenmeli Faster R-CNN modelinin elde ettiği görülmüştür.

Gelecek çalışmalarda nesne tespitine yardımcı olmak amacıyla gürültü, geometrik ve filtreleme gibi saldırı türlerinin de tespit çalışmaları gerçekleştirilebilir. Ayrıca, önerilen yöntem nesne tespiti sistemlerinin model seçimi aşamasına entegre edilebilir.

Çıkar Çatışması

Makale yazarları aralarında herhangi bir çıkar çatışması olmadığını beyan ederler.

Yazar Katkısı

Yazarlar makaleye eşit oranda katkı sağlamış olduklarını beyan eder.

KAYNAKLAR

Amit, Y., Felzenszwalb, P., & Girshick, R. (2020). Object detection. *Computer Vision: A Reference Guide*, 1-9.

Ayas, M. S., Ayas, S., & Djouadi, S. M. (2022, July). Projected Gradient Descent Adversarial Attack and Its Defense on a Fault Diagnosis System. In *2022 45th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 36-39). IEEE.

Balamurugan, T., & Gnanamanoharan, E. (2023). Brain tumor segmentation and classification using hybrid deep CNN with LuNetClassifier. *Neural Computing and Applications*, *35*(6), 4739-4753.

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Croce, F., & Hein, M. (2020, November). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning* (pp. 2206-2216). PMLR.

Das, S. D., Basak, A., & Dutta, S. (2022). A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles. *Neurocomputing*, *491*, 607-620.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185-9193).

Du, X., Song, L., Lv, Y., & Qiu, S. (2022). A Lightweight Military Target Detection Algorithm Based on Improved YOLOv5. *Electronics*, *11*(20), 3263.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Govindarajulu, Y., Amballa, A., Kulkarni, P., & Parmar, M. (2023). Targeted Attacks on Timeseries Forecasting. *arXiv preprint arXiv:2301.11544*.

Gu, X., Li, S., Ren, S., Zheng, H., Fan, C., & Xu, H. (2022). Adaptive enhanced swin transformer with U-net for remote sensing image segmentation. *Computers and Electrical Engineering*, *102*, 108223.

Guo, Z., Wang, C., Yang, G., Huang, Z., & Li, G. (2022). Msft-yolo: Improved yolov5 based on transformer for detecting defects of steel surface. *Sensors*, *22*(9), 3467.

Hu, K., Zhang, Z., Niu, X., Zhang, Y., Cao, C., Xiao, F., & Gao, X. (2018). Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function. *Neurocomputing*, *309*, 179-191.

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. *Procedia Computer Science*, *199*, 1066-1073.

Jocher, G., Nishimura, K., Mineeva, T., & Vilariño, R. (2020). Yolov5. *Code repository https://github.com/ultralytics/yolov5*.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

Längkvist, M., Kiselev, A., Alirezaie, M., & Loutfi, A. (2016). Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, *8*(4), 329.

Liu, H., Yu, Y., Liu, S., & Wang, W. (2022). A Military Object Detection Model of UAV Reconnaissance Image and Feature Visualization. *Applied Sciences*, *12*(23), 12236.

Liu, S., Wu, H., Lee, H. Y., & Meng, H. (2019, December). Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 312-319). IEEE.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing*, *438*, 14-33.

Moustapha, M., Tasyurek, M., & Ozturk, C. (2022). A Novel YoloV5 Deep Learning Model for Handwriting Detection and Recognition. *International Journal on Artificial Intelligence Tools*. doi:10.1142/S0218213023500161

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

Shelatkar, T., Urvashi, D., Shorfuzzaman, M., Alsufyani, A., & Lakshmanna, K. (2022). Diagnosis of brain tumor using light weight deep learning model with fine-tuning approach. *Computational and Mathematical Methods in Medicine*, *2022*.

Tasyurek, M., & Gul, E. (2023). A new deep learning approach based on grayscale conversion and DWT for object detection on adversarial attacked images. *The Journal of Supercomputing*, 1-34.

Terzi, R., Azginoglu, N., & Terzi, D. S. (2022). False positive repression: Data centric pipeline for object detection in brain MRI. *Concurrency and Computation: Practice and Experience*, *34*(20), e6821.

Wang, Y., Hao, Z., Zuo, F., & Pan, S. (2021, September). A fabric defect detection system based improved yolov5 detector. In *Journal of Physics: Conference Series* (Vol. 2010, No. 1, p. 012191). IOP Publishing.

Zhang, Y., Jiang, Z., Villalba, J., & Dehak, N. (2020, October). Black-Box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples. In *Interspeech* (pp. 4238-4242).