Gümüşhane Üniversitesi Fen Bilimleri Dergisi

GUFBD / *GUJS* (2023) 13(1): 210-220 doi: 10.17714/gumusfenbil.1120895 Araştırma Makalesi / *Research Article*

Çekişmeli makine öğrenmesi saldırılarının rulman arıza teşhisindeki etkileri

The effects of adversarial machine learning attacks on bearing fault diagnosis

Mustafa Şinasi AYAS *[1], Selen AYAS [2]

[1]*Karadeniz Teknik Üniversitesi, Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, 61080, Trabzon*

[2]*Karadeniz Teknik Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 61080, Trabzon*

• Geliş tarihi / *Received*: 24.05.2022 • Kabul tarihi / *Accepted*: 08.12.2022

Öz

Bilgiye dayalı arıza teşhis yöntemleri, sırasıyla model tabanlı ve sinyal tabanlı teşhis yöntemlerinde gerekli olan kesin model ve sinyal kalıplarına ihtiyaç duymadıkları için daha fazla tercih edilir hale gelmiştir. Makine öğrenimi teknikleri, ham sinyallerden sağlık durumlarına bilgileri eşleyerek arıza teşhisinde dikkate değer sonuçlar sağlamaktadır. Ancak makine öğrenimi yöntemlerinin kullanıldığı diğer endüstriyel uygulamalarda olduğu gibi kötü niyetli saldırılara karşı zafiyetleri ortaya çıkmaktadır. Bu çalışmada erişime açık CWRU rulman sağlık durumu veri kümesindeki 10 farklı sağlık durumunu içeren titreşim sinyalleri 2B görüntülere çevrilmiş ve görüntülerin sınıflandırılması için kullanılan derin artık öğrenme (DRL) ağ modeline beyaz kutu çekişmeli saldırılarından Hızlı Gradyan İşareti Yöntemi (FGSM), Temel Yinelemeli Yöntem (BIM), İzdüşürülen Gradyan İniş (PGD) ve Carlini ve Wagner (CW) saldırıları uygulanmıştır. Uygulanan çekişmeli makine öğrenmesi saldırılarının etkisini incelemek için DRL modelinin dayanıklılığı analiz edilmiştir. Elde edilen sonuçlara göre uygulanan çekişmeli saldırılar DRL modelini kandırarak yanlış sonuç üretmesine yol açmış ve rulman arıza teşhis sınıflandırma doğruluğunu düşürmüştür. 2B görüntülere oldukça küçük bir pertürbasyon eklenmesi sonucu %99.98 olan sınıflandırma doğruluğu FGSM, BIM, PGD, ve CW saldırı yöntemleri ile sırasıyla %68,38, %61,75, %61,88 ve %63,31 değerine düşmüştür. Ulaşılan sonuçlar kullanılan çekişmeli makine öğrenmesi saldırı yöntemlerinin rulman arıza teşhis sınıflandırma doğruluğunu düşürmesi için büyük potansiyele sahip olduğunu göstermektedir.

Anahtar kelimeler: Arıza teşhis sistemi, Çekişmeli makine öğrenmesi, Derin artık öğrenme modeli

Abstract

Knowledge-based fault diagnosis methods have become more preferred as they do not need precise model and signal patterns required in model-based and signal-based diagnosis methods, respectively. Machine learning techniques provide remarkable results on fault diagnosis by mapping information from raw signals to health condition. However, their vulnerabilities against adversarial attacks arise as in the other industrial applications employing machine learning methods. In this study, the vibration signals containing 10 different health condition in the public CWRU bearing health condition dataset are converted into 2D images and Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD) and Carlini and Wagner (CW) white box adversarial attacks are applied into the deep residual learning (DRL) network model which classifies the images of rolling bearing. The robustness of the DRL model is analyzed to examine the effect of the implemented adversarial machine learning attacks. According to the obtained results, the adversarial attacks fooled the DRL model, causing it to produce misclassification results and so decrease the bearing fault diagnosis classification accuracy. As a result of injecting quite small perturbation to 2D images, the classification accuracy, which was 99.98%, is decreased to 68.38%, 61.75%, 61.88% and 63.31% by FGSM, BIM, PGD and CW attack methods, respectively. The achieved results show that the adversarial machine learning attack methods have great potential to reduce the accuracy of bearing fault diagnosis classification.

Keywords*: Adversarial machine learning, Deep residual learning model, Fault diagnosis system*

* Mustafa Şinasi AYAS; msayas@ktu.edu.tr

1. Giriş
1.Introduction

Endüstri 3.0'dan Endüstri 4.0'a geçişle birlikte endüstriyel süreçler daha akıllı olmaktadır. Özellikle her bir alt süreç durumunun izlenmesi, detaylandırılmış sensörler sayesinde daha kolay hale gelmektedir. Endüstriyel bir süreçte istenilen performansın elde edilmesi çok önemli olmasına rağmen, sürecin güvenilirliği de hem güvenlik hem de sürdürülebilirlik açısından aynı derecede önemlidir. Bu nedenle, arıza teşhisi, hem akademi hem de endüstri tarafından önemli derecede ilgi çeken bir süreç mühendisliği araştırma alanıdır (Park vd., 2020). Dönen makinelerin önemli bir parçası olan rulmanların arıza teşhisi, çekici araştırma alanlarından biridir. Rulman ile yataklanmış bir makinedeki rulmanda meydana gelen herhangi bir arızanın erken tespiti, tüm sürecin bozulmasını önleyebilir (Demir & Müştak, 2021; Zhao vd., 2021).

Arıza teşhis yöntemleri model tabanlı, sinyal tabanlı, bilgi tabanlı ve hibrit arıza teşhisi olarak sınıflandırılabilir (Gao vd., 2015). Bu yöntemlerden bilgi tabanlı olanlar, sırasıyla model tabanlı ve sinyal tabanlı yöntemlerde gerekli olan kesin model bilgisi ve işaret örüntüsüne ihtiyaç duymadıkları için daha fazla tercih edilir hale gelmiştirler. Bilgiye dayalı yöntemlerde, ham ölçülen veriler ile sürecin durumu arasında bir ilişki yürütmek için uzun vadeli sürekli izlemeden toplanan büyük veriler kullanılır (Jia vd., 2018). Endüstri 4.0, veri toplama prosedürünü hızlandırmakta ve araştırmacıları, arıza teşhisinde büyük verilerin potansiyelini kullanmaya yönlendirmektedir (Chen & Lin, 2014). Rulman arızası açısından, doğrusal olmayan ve dengesiz kararlılık özelliklerine sahip titreşim sinyalleri bu büyük veriyi oluşturmaktadır. Titreşim sinyallerindeki zorlukların üstesinden gelmek için, rulmanın sağlık durumunun izleme sürecinde yaygın olarak makine öğrenmesi (ML) ve daha çok derin öğrenme (DL) algoritmaları kullanılır (Zhao vd., 2021).

DL yaklaşımları, titreşim sinyallerinden sağlık durumuna bilgileri haritalayarak rulman arıza teşhisinde memnun edici sonuçlar sağlamaktadırlar. Ancak, DL yaklaşımlarını kullanan diğer endüstriyel uygulamalarda olduğu gibi böyle bir arıza teşhis sisteminde art niyetli saldırılara karşı artan güvenlik açıkları sorunu ortaya çıkmaktadır (Kumar vd., 2020). DL tabanlı sistemlere karşı gerçekleştirilen art niyetli saldırılara Çekişmeli Makine Öğrenmesi (AML) saldırıları adı verilir. AML'nin amacı, DL modeline hafif pertürbasyon ekleyerek sınıflandırıcıyı nominal kararını değiştirmeye gizlice zorlamaktır. Bu nedenle, yanlış sınıflandırma sayısı artabileceği için modelin performansı düşebilmektedir (Anthi vd., 2021).

1.1. Literatür taraması
1.1.Literature review

Literatür çalışmaları arıza teşhis ve çekişmeli makine öğrenmesi çalışmaları olmak üzere iki alt bölümde incelenmiştir.

1.1.1. Arıza teşhis çalışmaları
1.1.1. Fault diagnosis studies

Wen vd. (2017) arıza teşhisi için LeNet-5'ten ilham alan evrişimli sinir ağı (CNN) tabanlı bir yaklaşım önerdiler. Çalışmalarında bir motor rulmanının halka açık veri kümesinden alınan titreşim sinyallerini 2B gri seviyeli görüntülere dönüştürdüler. Çalışmanın sonucunda, rulman arıza teşhisinde 2B görüntüleri kullanan CNN tabanlı yaklaşımlarının etkinliğini vurguladılar. Transfer öğrenme (TL) tekniği Guo vd. (2018) için rulman arıza teşhisi çalışmalarına ilham kaynağı olmuştur. Çalışmada iki aşamalı bir yaklaşımla derin bir evrişimli TL ağı sunulmuştur. Aşamalardan birinde, rulmanın sağlık durumu, ham sinyallerin ayırt edici özelliklerini öğrenen 1D CNN kullanılarak sınıflandırılmıştır. Rulmanın titreşim sinyallerini kullanan TL tabanlı başka bir yöntem Wen vd. (2019) tarafından tanıtılmıştır. Çalışmada, ham titreşim sinyallerinden ayırt edici özellikleri çıkarmak için üç katmanlı bir seyrek otomatik kodlayıcı kullanılmıştır. Çalışmanın sonuçları, sunulan tekniğin klasik ML yöntemlerine göre daha iyi sınıflandırma doğruluğu sağladığını göstermektedir. Liu vd. (2019) bir arıza teşhis çözümü olarak derin bir çekişmeli alan uyarlama modeli sunmuştur. Daha değerli rulman arızası özelliklerini çıkarmak için derin yığın otomatik kodlayıcı ve özellik öğrenme birleştirilmiştir. Bu çalışmada, geleneksel ML ve DL yöntemlerine kıyasla oldukça iyi sağlık sınıflandırması sonuçları elde edilmiştir. Chen vd. (2020) döngüsel spektral analizi içeren DL tabanlı bir teknik tanıtmıştır. Önce farklı rulman arıza tiplerinin ayırt edici modelleri elde edilmiş, daha sonra rulmanın sağlık durumunu sınıflandırmak için önerilen bir CNN modeli uygulanmıştır. Derin bir orman ve bir CNN modeli içeren hibrit bir teknik Xu vd. (2021) tarafından sunulmuştur. Derin orman sınıflandırıcısı, titreşim sinyallerinden üretilen zaman-frekans

görüntülerini kullanan CNN modeli tarafından çıkarılan ayırt edici özelliklerle beslenmiştir. Bir CNN modeli ve bir uzun kısa süreli bellek (LSTM) modeli, rulman arıza teşhisinde ham titreşim sinyallerini kullanan uçtan uca bir DL yaklaşımı olarak birleştirilmiştir. 2D gri seviyeli görüntülerle beslenen bir CNN modeli Zhao vd. (2021) tarafından sunulmuştur. Çalışmada, titreşim sinyallerini 2D görüntülere dönüştürmek için bir sinyalden görüntüye haritalama tekniği kullanılmıştır. Rulman arıza teşhis için gerçekleştirdiğimiz önceki çalışmada (Ayas & Ayas, 2021), titreşim sinyallerinden oluşturulan 2B görüntüler ve rulman sağlık durumu arasında uçtan uca eşlemeyi öğrenmek için oluşturulmuş derin artık öğrenme (DRL) tabanlı bir model önerilmiştir. Kapsamlı literatür taraması için Park vd. (2020) tarafından sunulmuş olan derleme makalesi incelenebilir.

1.1.2. Çekişmeli makine öğrenmesi çalışmaları
1.1.2. Adversarial machine learning studies

Sinir ağlarındaki güvenlik açıklarından ilk kez 2013 yılında Szegedy vd. (2013) tarafından bahsedilmiştir. Ağın tahmin hatasını maksimize ederek hesaplanan, güçlükle algılanabilen bir pertürbasyon enjekte ederek bir görüntünün yanlış sınıflandırabileceğini vurguladılar. Hesaplanan pertürbasyonu meşru görüntüye enjekte ederek, insan gözünün algılayamayacağı yanıltıcı bir görüntü elde edilebileceği bu çalışmada gösterilmiştir. Ardından, 2015 yılında Goodfellow vd. (2015), geri yayılım kullanılarak verimli bir şekilde hesaplanan gradyana göre çekişmeli örnekler oluşturabilen Hızlı Gradyan İşaret Yöntemi'ni (FGSM) sunmuşturlar. FGSM'nin farklı bir biçimi olarak, daha fazla çekişme çeşitliliği arayan Hızlı Gradyan Değeri (FGV) yöntemi 2016 yılında Rozsa vd. (2016) tarafından tanıtılmıştır. FGV yöntemi, ham bir gradyan kullanır ve yalnızca gradyanın işaretini kullanan FGSM'ye kıyasla daha az dokusal bozulma ile sonuçlanan daha etkili pertürbasyonlar ürettiği belirtilmiştir. Yanlış sınıflandırma için yeterli minimum pertürbasyonu üreterek derin sinir ağlarını kandıran DeepFool isimli yaklaşım Moosavi-Dezfooli vd. (2016) tarafından önerilmiştir. FGSM'nin yinelemeli bir versiyonu olan Temel Yinelemeli Yöntem (BIM) Kurakin vd. (2016) tarafından sunulmuştur. BIM'in farklı bir formu olarak bilinen izdüşürülen gradyan inişi (PGD), Madry vd. (2016) tarafından en güçlü "birinci dereceden çekişmeli" saldırı olarak sunulmuştur. PGD, ağ ile ilgili birinci dereceden bilgileri kullanır.

AML uygulamaları ile ilgili çalışmalar son yıllarda hızla artmaktadır. 2018 yılında Suciu vd. (2018), doğrusal bir sınıflandırıcıya sahip Android kötü amaçlı yazılım algılama sistemine karşı zehirlenme saldırısı başlatmışlardır. 2019 yılında Kuppa vd. (2019), denetimsiz aykırılık dedektörüne karşı bir gri kutu saldırısı başlatmıştırlar ve dikkate değer bir sonuç elde etmiştirler. 2020'de Xu vd. (2020), hem n-güçlü düğümleri hem de FGSM yöntemini kullanarak kötü amaçlı yazılım sınıflandırıcısına karşı bir gri kutu saldırısı başlatmışlardır. 2021'de Vakhshiteh vd. (2021) yüz tanıma sistemlerine karşı gerçekleştirilen çekişmeli saldırılar hakkında kapsamlı bir derleme çalışması sunmuşturlar. Makalede giriş görüntüleri doğal görünse bile yüz tanıma sistemlerinin AML'ye karşı savunmasız olduğu vurgulanmaktadır.

Kurakin vd. (2018) tarafından fiziksel dünyada AML tekniklerinin mümkün olduğu ele alınmıştır. Basılı çekişmeli örnekler, bir cep telefonu kamerası kullanılarak bir ImageNet sınıflandırıcısına beslenmiştir. Sonuçlar, çekişmeli örneklerin çoğunun yanlış sınıflandırıldığını göstermektedir. Brown vd. (2017), evrensel çekişmeli görüntü yamaları oluşturmak için bir yöntem önermişlerdir. Oluşturulan yamalar herhangi bir resme yapıştırılabilir ve yazdırılabilir. Bu çalışmada, yeni çekişmeli görüntünün, yapıştırılan yama hedef nesneden daha küçük olsa bile sınıflandırıcıyı aldattığı vurgulanmıştır. Sayles vd. (2021), oküler artefaktlarla hedef nesne yerine nesneyi aydınlatan ışık seviyesini değiştirmiştir. İnsan gözü tarafından algılanamayan fiziksel çekişmeli görüntüler yanlış sınıflandırılmıştır.

Çekişmeli saldırılar ve savunma mekanizmalarındaki ilerlemelerle ilgili kapsamlı literatür taraması için Akhtar vd. (2021) tarafından sunulmuş olan derleme makalesi incelenebilir.

1.2. Çalışmanın motivasyonu ve katkısı
1.2.Motivation and contribution of the study

Bu çalışmadaki temel amaç, makine öğrenmesi tabanlı rulman arıza teşhis yaklaşımlarının AML saldırılarına karşı potansiyel bir hedef olup olmadıklarını incelemektir. Bu kapsamda önceki çalışmamızda (Ayas & Ayas, 2021) rulman arıza teşhisi için önerdiğimiz değiştirilmiş DRL modeli ele alınmıştır. Önerilen DRL modelinin performansı, 10 farklı sağlık durumunu içeren erişime açık bir veri seti üzerinde test edilmiş ve ilgili çalışmada güncel teşhis yöntemleriyle karşılaştırılmıştır. Bu karşılaştırma sonucunda değiştirilmiş DRL modelinin

mevcut yöntemlerden daha iyi performans gösterdiği ve ortalama %99,98 doğrulukla çalıştığı gözlemlenmiştir. Bu çalışma kapsamında değiştirilmiş DRL modelinin AML saldırıları karşısındaki dayanıklılığı incelenmiştir. FGSM, BIM, PGD ve Carlini ve Wagner (CW) saldırıları çalışma kapsamında incelenen AML saldırılarıdır.

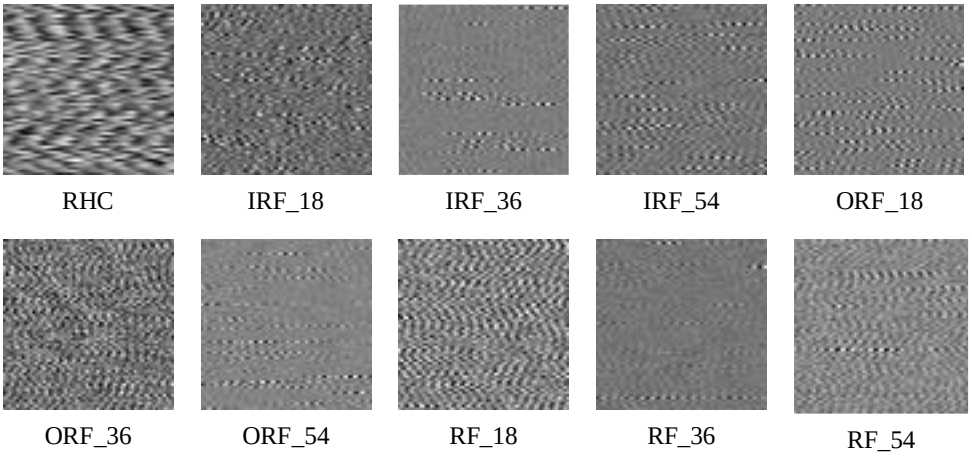
2. Materyal ve metot
2. Material and method

Çekişmeli makine öğrenmesi saldırılarının rulman arıza teşhisindeki etkilerini incelemek için çalışma kapsamında kullanılan veri kümesinin detayları, bu veri kümesindeki sağlık durumlarını teşhis etmek için kullandığımız değiştirilmiş DRL modelinin mimarisi ve değiştirilen DRL modeline uygulanan AML saldırıları alt başlıklarda detaylandırılmıştır.

2.1. Veri kümesi tanıtımı
2.1. Dataset description

Çalışmada rulman titreşim sinyallerini içeren yaygın olarak kullanılan erişime açık Case Western Reserve University (CWRU) veri kümesi kullanılmıştır. Veri kümesi normal sağlıklı duruma (RHC) ek olarak, iç yuva arızası (IRF), dış yuva arızası (ORF) ve bilye arızası (RF) olmak üzere üç farklı arıza durumu içermektedir. Ayrıca, her arızalı durumun 0,18 mm, 0,36 mm ve 0,54 mm olmak üzere üç farklı büyüklük düzeyi vardır. Dolayısıyla veri setinde toplam on farklı sağlık durumu bulunmaktadır. Bu on farklı koşulun titreşim sinyalleri, 12 kHz örnekleme frekansı ile dört farklı yük senaryosu (0, 1, 2 ve 3 hp) altında toplanmıştır.

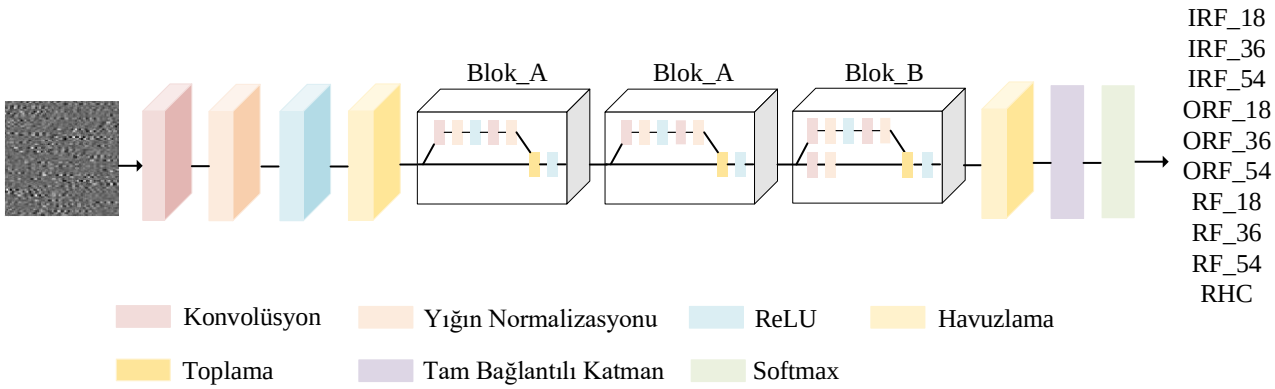
Veri kümesinde bulunan 1B ham titreşim sinyalleri ilk olarak 64×64 boyutunda 2B görüntüye dönüştürülmüştür. 1B sinyaller üzerinde uygulanan veri ön işlemenin işlem adımları önceki çalışmamızda (Ayas & Ayas, 2021) detaylandırılmıştır. Çalışma kapsamında kullanılan derin sinir ağı modelinin eğitiminde her bir sağlık durumu için 560 görüntü olmak üzere toplam 5600 görüntü, test aşamasında ise her bir sağlık durumu için 160 görüntü olmak üzere toplam 1600 görüntü kullanılmıştır. On farklı sağlık durumu için elde edilen bazı örnek görüntüler Şekil 1'de verilmiştir.



Şekil 1. Rulmanın 10 farklı sağlık durumu için 1B ham titreşim sinyallerinden dönüştürülen görüntüler
Figure 1. *The images converted from 1D vibration signals for 10 different health conditions of the bearing*

2.2. Değiştirilmiş DRL sınıflandırıcısı
2.2. Modified DRL classifier

Şekil 2, özellik gösterimi, DRL ve sınıflandırma ağları yapılarından oluşan DRL ağının mimarisini göstermektedir. Özellik gösterimi ağı, gri tonlamalı görüntüleri özellik haritası olarak temsil etmek için giriş olarak almaktadır. Ardından, DRL, arıza teşhis görevini yerine getirmektedir. Son olarak, DRL ağının son katmanındaki özellik haritaları, rulmanın arıza durumunu belirlemek için sınıflandırma ağına giriş olarak verilmektedir.



Şekil 2. Önerilen değiştirilmiş DRL ağ mimarisi ve blok yapıları
Figure 2. *The proposed modified DRL network architecture and its block design*

Özellik gösterimi ağının girişi olan 64×64 gri tonlamalı görüntüye, *x* ve *y* boyutunda 2 adım ve 3 dolgu kullanılarak her biri 7×7 boyutunda 64 farklı filtre ile konvolüsyon işlemi uygulanır. Konvolüsyon katmanını, yığın normalizasyonu katmanı ve ReLU aktivasyon fonksiyonu takip etmektedir. Özellik gösterimi ağı, 3×3 boyutunda 2 adım ve 1 dolgu kullanılan maksimum havuzlama gerçekleştiren bir havuzlama katmanı ile sona ermektedir. Sonuç özellik haritaları daha sonra DRL ağına beslenir. Bloklarda kullanılan konvolüsyon katmanları 3×3 filtre boyutuna sahiptir. Bir blokta kullanılacak olan filtre sayısını tasarlamak için benimsenen yaklaşım, elde edilen özellik haritası boyutu 2 adım kullanılarak yarıya indirildiğinde filtre sayısının 2 katına çıkarılmasıdır. Blok_B'de gösterildiği gibi özellik haritalarını toplamak için artık öğrenme şemasına atlamalı bağlantı yapısı eklenmiştir. Özellik haritaları aynı boyutta olduğundan dolayı Blok_A doğrudan kullanılırken, Blok_B yapısında boyutu artırmak için 1×1 boyutlu konvolüsyon kullanılarak giriş özellik haritasına bir projeksiyon şeması uygulanır. Sınıflandırma ağı, boyutu son konvolüsyon katmanının çıkışına eşit olan ve aynı boyutta adım kullanan bir ortalama havuzlama katmanıyla başlamaktadır. Sonuç özellik haritaları, 10 yollu tam bağlantılı bir katmandan geçirilerek softmax katmanına giriş olarak uygulanır. Softmax katmanının çıkışı ise ilgili giriş görüntüsünün rulman sağlık durumu sınıfını vermektedir. Önerilen modelin katman konfigürasyonları Tablo 1'de verilmiştir.

Tablo 1. Önerilen değiştirilmiş DRL modelinin katman konfigürasyonları: a, adım boyutu; d, dolgu miktarı
Table 1. *The layer configurations of the proposed modified DRL model: a, stride; d, padding*

Katman İsmi		**Model**
Konvolüsyon		7x7, 64, a:2, d:3
Havuzlama		3x3 maksimum havuzlama, a:2, d:1
Blok_A	Konvolüsyon1	3x3, 64, a:1, d:1
	Konvolüsyon2	3x3, 64, a:1, d:1
Blok_A	Konvolüsyon1	3x3, 64, a:1, d:1
	Konvolüsyon2	3x3, 64, a:1, d:1
Blok_B	Konvolüsyon1	3x3, 128, a:2, d:1
	Konvolüsyon2	3x3, 128, a:1, d:1
Havuzlama		8x8 ortalama havuzlama, a:8, d:0
Tam bağlantılı katman		128x10 tam bağlantı

2.3. Çekişmeli makine öğrenmesi saldırıları
2.3. Adversarial machine learning attacks

Çekişmeli saldırılar, genellikle bir optimizasyon algoritması tarafından oluşturulan algılanması zor bir pertürbasyonun girdi görüntüsüne eklenip çekişmeli görüntü ya da örneğin oluşturulmasıyla gerçekleştirilir. Hafifçe bozulan test girdi örnekleri eğitim örnekleri ile eğitilmiş derin sinir ağları modellerini yüksek bir güven skoru ile yanlış tahmin yapmaya kandırabilirler. Çalışma kapsamında değiştirilmiş DRL modeline AML saldırısı uygulamak için kullanılan saldırı yöntemleri alt başlıklarda detaylandırılmıştır.

2.3.1. Hızlı gradyan işareti yöntemi saldırısı
2.3.1. Fast gradient sign method attack

FGSM, Goodfellow vd. (2014) tarafından önerilen tek adım algoritmasıdır. FGSM'nin temel özelliği, giriş gradyan yönü boyunca ε miktarı kadar sadece bir adımda $\mathbf{x}$ giriş örneklerini bozmaktır ve bu nedenle FGSM düşük hesaplama maliyetine sahiptir. Verimliliğine karşın FGSM yöntemi tarafından üretilen pertürbasyonlar genellikle çok yüksek olmaktadır. Verilen bir $\mathbf{x}$ giriş örneği için, FGSM yöntemi (1) eşitliğini kullanarak çekişmeli örnekler oluşturmaktadır.

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \varepsilon \, \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\text{h}(\mathbf{x}_{\text{adv}}), \text{y})) \tag{1}$$

Burada, $\mathbf{x}_{\text{adv}}$ oluşturulan çekişmeli örneği, ∇_{x} gradyan vektörü, ε pertürbasyon miktarını ve $\mathcal{L}(.)$, h sınıflandırıcısını kullanan modelin kayıp fonksiyonunu ifade etmektedir.

2.3.2. Temel yinelemeli yöntem saldırısı
2.3.2. Basic iterative method attack

FGSM yönteminin yinelemeli versiyonu olan BIM yöntemi Kurakin vd. (2018) tarafından önerilmiştir. FGSM'den farklı olarak BIM giriş görüntüsünü daha küçük adım boyutlarıyla yinelemeli bir şekilde bozmaktadır. Özyinelemeli bir yöntem olan BIM çekişmeli örnekleri (2) eşitliği kullanılarak oluşturmaktadır.

$$\mathbf{x}^{t} = \mathbf{x}^{t-1} + \alpha \, \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\text{h}(\mathbf{x}^{t-1}), \text{y})) \tag{2}$$

Burada α adım boyutu ve $\mathbf{x}^{t}$ t. adımdaki ($\mathbf{x}^{0} = \boldsymbol{x}$) çekişmeli örneği ifade etmektedir. Adım boyutu genellikle bozulmanın tüm T miktarı için $\varepsilon/T \leq \alpha \leq \varepsilon$ aralığında seçilmektedir.

2.3.3. İzdüşürülen gradyan iniş saldırısı
2.3.3. Projected gradient descent attack

Madry vd. (2017) tarafından önerilen PGD saldırısı, $\mathbf{x}$ giriş görüntüsünü küçük adım boyutlarıyla T adımı sayısınca bozmaktadır. Bozulmanın her bir adımından sonra PGD çekişmeli örneği eğer ötesine geçerse $\mathbf{x}$ örneğini ε-küresine geri yansıtmaktadır.

$$\mathbf{x}^{t} = \prod_{\varepsilon}\left(\mathbf{x}^{t-1} + \alpha \, \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\text{h}(\mathbf{x}^{t-1}), \text{y}))\right) \tag{3}$$

Burada α; adım boyutu, $\prod(.)$; izdüşürme fonksiyonunu ve $\mathbf{x}^{t}$; t. adımdaki ($\mathbf{x}^{0} = \boldsymbol{x}$) çekişmeli örneği ifade etmektedir. BIM yönteminden farklı olarak PGD $\mathbf{x}^{t} = \mathbf{x} + U^{d}(-\varepsilon, \varepsilon)$ için rastgele bir başlangıç noktası kullanmaktadır. Burada, $U^{d}(-\varepsilon, \varepsilon)$, $-\varepsilon$ ve ε aralığında tekdüze dağılımı ifade etmektedir. PGD yöntemi en güçlü birinci derecen saldırı yöntemi olarak kabul edilmektedir.

2.3.4. Carlini ve wagner saldırısı
2.3.4. Carlini and wagner attack

CW saldırısı (Carlini vd., 2017) çekişmeli örnekler oluşturmak için önerilen en gelişmiş algoritmayı temsil etmektedir. CW saldırısının ℓ_2 ve ℓ_∞ olmak üzere iki versiyonu bulunmaktadır. Madry vd. (2017) çalışmalarında önerdiği üzere hedeflenen CW saldırısının ℓ_∞ versiyonu yinelemeli bir şekilde PGD algoritması ile (4) ve (5) eşitliği kullanılarak çözülmektedir.

$$\mathbf{x}^{t} = \prod_{\varepsilon}\left(\mathbf{x}^{t-1} - \alpha \, \text{sign}\left(\nabla_{\mathbf{x}} \hat{f}(\mathbf{x}^{t-1})\right)\right) \tag{4}$$

$$\hat{f}(\mathbf{x}^{t-1}) = max\left(\boldsymbol{z}_{y}(\mathbf{x}^{t-1}, \boldsymbol{\theta}) - \boldsymbol{z}_{y_{max \neq y}}(\mathbf{x}^{t-1}, \boldsymbol{\theta}), -\kappa\right) \tag{5}$$

Burada $\hat{f}(.)$, sınırlandırılmış optimizasyon problemi için asıl kayıp fonksiyonunun yerine geçen kayıp fonksiyonunu ifade etmektedir. $\boldsymbol{z}_{y}$, y sınıfına göre lojit değerleri, $\boldsymbol{z}_{y_{max \neq y}}$, diğer sınıfların maksimum lojit değerlerini ve κ, saldırının güvenini kontrol eden bir parametreyi ifade etmektedir.

3. Bulgular ve tartışma
3. Results and discussion

Bu bölümde FGSM, BIM, PGD ve CW çekişmeli saldırıları uygulandıktan sonra rulmanın 10 farklı sağlık durumunun DRL modeli kullanılarak sınıflandırma doğrulukları incelenecektir. Önerilen değiştirilmiş DRL modeli, CWRU veri kümesi ile eğitilmiş ve herhangi bir saldırı uygulanmadan önce %99,98 doğruluk elde edilmiştir (Ayas & Ayas, 2021). Önerilen değiştirilmiş DRL modelinin eğitiminde kullanılan hiperparametrelerin seçimi Tablo 2'de verilmiştir. Deneylerde kullanılan AML yaklaşımlarının parametreleri ise Tablo 3'te verilmiştir. Uygulanan çekişmeli saldırılar hedeflenmeyen saldırılar olup belirli bir sınıf etiketi vermeksizin DRL modelinin sadece yanlış sınıflandırma sonucu üretmesini sağlamaktadır. Bunun yanı sıra, deneyler esnasında sadece test örnekleri AML saldırı yöntemleri kullanılarak bozulmuştur.

Tablo 2. DRL modelinin eğitiminde kullanılan parametreler
Table 2. *The parameters used in the training of DRL model*

Parametreler	Değerler
Optimizasyon yöntemi	Momentumlu stokastik gradyan inişi
Öğrenme katsayısı	10^{-4}
Momentum	0.9
Yığın boyutu	4
Epok sayısı	50

Tablo 3. AML saldırılarının parametre ayarlamaları: α, adım boyutu; ε, maksimum pertürbasyon; i, iterasyon sayısı
Table 3. *The settings of AML attacks:* α*, step size;* ε*, maximum perturbation*; *i, iteration number*

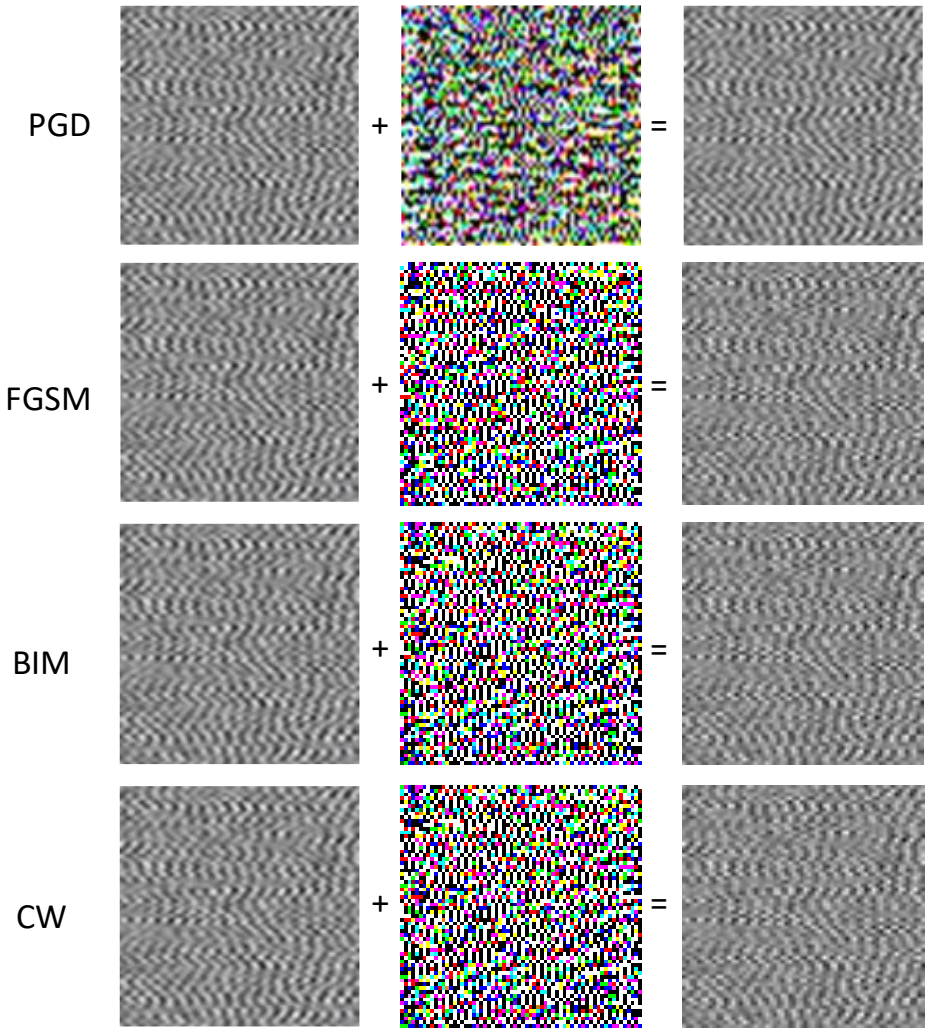
Saldırı yöntemi	Parametreler
FGSM	α=0.0005, ε=$(1/255, 2/255, 3/255)$, i=1
BIM	α=0.0005, ε=$(1/255, 2/255, 3/255)$, i=40
PGD	α=0.0005, ε=$(1/255, 2/255, 3/255)$, i=40
CW	α=0.0005, ε=$(1/255, 2/255, 3/255)$, i=250, κ=0

Tablo 4, farklı pertürbasyon, ε, değerlerine bağlı olarak tüm saldırı yöntemlerinin uygulanmasıyla DRL modeli ile elde edilen sınıflandırma doğruluklarının karşılaştırmalı sonuçlarını göstermektedir. Burada ε değeri başlangıçta 1/255 olarak seçilmiş ve 1/255 değeri ile artırılarak ε parametresinin sınıflandırma doğruluğuna etkisi analiz edilmiştir. Rulman arıza teşhis sistemi tarafından ölçülen ham titreşim sinyallerinden dönüştürülen görüntüler 8 bit olduklarından dolayı dinamik aralığın 1/255'inin altındaki tüm bilgilerin göz ardı edilmesini önlemek için ε artışı 1/255'ten büyük olacak şekilde seçilmiştir.

Tablo 4'teki sonuçlar incelendiğinde beklenildiği üzere tüm yöntemler için çekişmeli pertürbasyon değerinin artışı sınıflandırma doğruluğunda büyük bir düşüşe neden olmaktadır. Tek iterasyonda hesaplanan FGSM saldırısı 1/255 pertürbasyon değerinde modelin sınıflandırma doğruluğunu %99,98'den %68,38'e düşürdüğü ve çoklu iterasyona sahip BIM, PGD ve CW saldırıları ise sınıflandırma doğruluğunu sırasıyla %61,75, %61,88 ve %63,31'e düşürdüğü görülmektedir. ε pertürbasyon değerindeki artış bütün saldırılar için sınıflandırma doğruluğunu azaltmıştır. Değiştirilmiş DRL modeli FGSM saldırısına karşı daha dayanıklı olup ε=3/255 için %31,81 sınıflandırma doğruluğu elde edilmiştir. Üretilen pertürbasyon ve oluşturulan çekişmeli örnekler Şekil 3'te verilmektedir. Şekilden de görüldüğü üzere oluşturulan çekişmeli görüntüler orijinal görüntülere oldukça benzemektedir.

Tablo 4. Farklı ε değerlerine bağlı AML saldırılarının sınıflandırma doğruluğuna etkisi
Table 4. *The effect of AML attacks on classification accuracy wıth different values of* ε

Saldırı yöntemi	$\varepsilon = 1/255$	$\varepsilon = 2/255$	$\varepsilon = 3/255$
FGSM	%68,38	%47,81	%31,81
BIM	%61,75	%34,63	%10,06
PGD	%61,88	%35,19	%14,00
CW	%63,31	%37,38	%10,00



Şekil 3. Önerilen değiştirilmiş DRL modelini kandırmak için kullanılan çekişmeli makine öğrenmesi saldırı yöntemlerinin 0_18ball sınıfı için örnek görüntüleri. Sol: temiz görüntü, orta: çekişmeli makine öğrenmesi yöntemleriyle oluşturulan çekişmeli pertürbasyonlar, sağ: çekişmeli örnekler

Figure 3. *Examples of adversarial machine learning attacks for 0_18ball class to fool the proposed modified DRL model. Left: clear images, middle: adversarial perturbations generated by the adversarial machine learning attack methods, right: adversarial samples*

4. Tartışma ve sonuçlar

4. Discussion and conclusions

Bu çalışmada rulman arıza teşhis sistemlerinde kullanılan DL modellerinin maruz kalacağı AML saldırıları karşısındaki davranışı incelenmiştir. Bu kapsamda değiştirilmiş DRL modeli kullanan bir teşhis sistemine FGSM, BIM, PGD ve CW saldırıları uygulanarak modelin sınıflandırma doğruluğuna etkileri gözlemlenmiştir. Elde edilen bulgular, teşhis sistemine uygulanan bu saldırılardan BIM yönteminin modelin sınıflandırma doğruluğunu %61,75 oranına düşürerek en çok yanlış teşhise neden olduğunu ve buna karşın FGSM yönteminin ise %68,38 oranına düşürerek en dayanıklı çekişmeli makine öğrenmesi yöntemi olduğunu göstermiştir. Deneysel sonuçlar, AML'nin mevcut DL modellerine uygulanmasının mümkün olduğunu ve oldukça küçük pertürbasyonlar için bile, modelin teşhis sistemindeki sınıflandırma doğruluğunu azalttığını ve yanlış teşhislere neden olduğunu göstermektedir. Bunun yanı sıra, oluşturulan çekişmeli örneklerin orijinal görüntülerden insan gözüyle net bir şekilde ayrıştırılamadığı sonucuna varılmıştır. Gelecek çalışmalar, DL modellerini çekişmeli örneklere karşı savunmak için en etkili yaklaşımlardan biri olan çekişmeli eğitimin rulman arıza teşhis sistemlerinde DL modellerine uygulanan AML saldırılarına karşı etkisini incelemeyi içermektedir.

Teşekkür
Acknowledgement

Yazarlar, 2219 programı kapsamında verdikleri destek için TÜBİTAK'a ve makalenin inceleme ve değerlendirme aşamasında yapmış oldukları katkılardan dolayı editör ve hakemlere teşekkür etmektedir.

Yazar katkısı
Author contribution

Yazarlar çalışmaya eşit oranda katkıda bulunmuştur.

Etik beyanı
Declaration of ethical code

Bu makalenin yazarları, bu çalışmada kullanılan materyal ve yöntemlerin etik kurul izni ve / veya yasal-özel izin gerektirmediğini beyan etmektedir.

Çıkar çatışması beyanı
Conflicts of interest

Yazarlar herhangi bir çıkar çatışması olmadığını beyan etmektedir.

Kaynaklar
References

Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access, 9*, 155161-155196. https://doi.org/ 10.1109/ACCESS.2021.3127960

Anthi, E., Williams, L., Rhode, M., Burnap, P., & Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in industrial control systems. *Journal of Information Security and Applications*, *58*, 102717. https://doi.org/10.1016/j.jisa.2020.102717

Ayas, S., & Ayas, M. S. (2022). A novel bearing fault diagnosis method using deep residual learning network. *Multimedia Tools and Applications, 81,* 22407–22423. https://doi.org/10.1007/s11042-021-11617-1

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Chen, X. W., & Lin, X. (2014). Big data deep learning: challenges and perspectives. *IEEE Access*, *2*, 514-525. https://doi.org/10.1109/ACCESS.2014.2325029

Chen, Z., Mauricio, A., Li, W., & Gryllias, K. (2020). A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks. *Mechanical Systems and Signal Processing*, *140*, 106683. https://doi.org/10.1016/j.ymssp.2020.106683

Demir, H. G., & Müştak, O. (2021). Rulman hasarlarının titreşim ve gürültü analizi ile tespiti. *Avrupa Bilim ve Teknoloji Dergisi*, *25*, 571-581. https://doi.org/10.31590/ejosat.869285

Gao, Z., Cecati, C., & Ding, S. X. (2015). A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, *62*(6), 3757-3767. https://doi.org/10.1109/TIE.2015.2417501

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

Guo, L., Lei, Y., Xing, S., Yan, T., & Li, N. (2018). Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data. *IEEE Transactions on Industrial Electronics*, *66*(9), 7316-7325. https://doi.org/10.1109/TIE.2018.2877090

Jia, F., Lei, Y., Guo, L., Lin, J., & Xing, S. (2018). A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing, 272*, 619-628. https://doi.org/10.1016/j.neucom.2017.07.032

Khorram, A., Khalooei, M., & Rezghi, M. (2021). End-to-end CNN+LSTM deep learning approach for bearing fault diagnosis. *Applied Intelligence*, *51*(2), 736-751. https://doi.org/10.1007/s10489-020-01859-1

Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., & Xia, S. (2020). Adversarial machine learning-industry perspectives. *IEEE Security and Privacy Workshops (SPW)* (pp. 69-75), USA. https://doi.org/10.1109/SPW50608.2020.00028

Kuppa, A., Grzonkowski, S., Asghar, M. R., & Le-Khac, N. A. (2019). Black box attacks on deep anomaly detectors. *14th International Conference on Availability, Reliability and Security* (pp. 1-10), United Kingdom. https://doi.org/10.1145/3339252.3339266

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. Roman V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (ss. 99-112). Chapman and Hall/CRC.

Liu, Z. H., Lu, B. L., Wei, H. L., Chen, L., Li, X. H., & Rätsch, M. (2019). Deep adversarial domain adaptation model for bearing fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(7), 4217-4226. https://doi.org/10.1109/TSMC.2019.2932000

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

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2574-2582), USA.

Park, Y. J., Fan, S. K. S., & Hsu, C. Y. (2020). A review on fault detection and process diagnostics in industrial processes. *Processes, 8*(9), 1123. https://doi.org/10.3390/pr8091123

Rozsa, A., Rudd, E. M., & Boult, T. E. (2016). Adversarial diversity and hard positive generation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 25-32), USA.

Sayles, A., Hooda, A., Gupta, M., Chatterjee, R., & Fernandes, E. (2021). Invisible perturbations: physical adversarial examples exploiting the rolling shutter effect. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 14666-14675), USA.

Suciu, O., Marginean, R., Kaya, Y., Daume III, H., & Dumitras, T. (2018). When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. *27th USENIX Security Symposium* (pp. 1299-1316), USA.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199.*

Vakhshiteh, F., Nickabadi, A., & Ramachandra, R. (2021). Adversarial attacks against face recognition: A comprehensive study. *IEEE Access*, *9*, 92735-92756. https://doi.org/10.1109/ACCESS.2021.3092646

Wen, L., Li, X., Gao, L., & Zhang, Y. (2017). A new convolutional neural network-based data-driven fault diagnosis method. *IEEE Transactions on Industrial Electronics*, *65*(7), 5990-5998. https://doi.org/10.1109/TIE.2017.2774777

Wen, L., Gao, L., & Li, X. (2019). A new deep transfer learning based on sparse auto-encoder for fault diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(1), 136-144. https://doi.org/10.1109/TSMC.2017.2754287

Xu, P., Kolosnjaji, B., Eckert, C., & Zarras, A. (2020). Manis: Evading malware detection system on graph structure. *35th Annual ACM Symposium on Applied Computing* (pp. 1688-1695), Czech Republic.

Xu, Y., Li, Z., Wang, S., Li, W., Sarkodie-Gyan, T., & Feng, S. (2021). A hybrid deep-learning model for fault diagnosis of rolling bearings. *Measurement*, *169*, 108502. https://doi.org/10.1016/j.measurement.2020.108502

Zhao, J., Yang, S., Li, Q., Liu, Y., Gu, X., & Liu, W. (2021). A new bearing fault diagnosis method based on signal-to-image mapping and convolutional neural network. *Measurement*, *176*, 109088. https://doi.org/10.1016/j.measurement.2021.109088