BMBAD BJESR Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches

Kötü Amaçlı Yazılım Türlerinin Tespitinde Kullanılan 1B Verilerin 2B Barkod Türlerine Dönüştürülerek Derin Ağlarla Analizlerinin Gerçekleştirilmesi

Conversion of 1D Data Used in Detection of Malware Types to 2D Barcode Types and Analysis with Deep Networks

[1]Mesut TOĞAÇAR

[1]Fırat Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Yönetim Bilişim Sistemleri, ELAZIĞ

[1]mtogacar@firat.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received : 9 March 2023
Accepted : 15 April 2023

Keywords:

Feature Extraction, Malware, 2D Barcode Types, Deep Learning

ABSTRACT

Malware is software designed to damage computer-based systems, obtain or modify important information. This type of software targets network environments where people interact. Smart devices used in these network environments have become one of the indispensable parts of our lives today. Recently, many artificial intelligence-based studies have been carried out in order to ensure the security of smart devices and to detect malicious software. The dataset of this study consists of text-based content containing hidden malware types. The proposed approach consists of a preprocessing step and a deep learning model. In the preprocessing step, two new datasets were obtained by transforming the text-based data into 2-dimensional barcode types. In the next step, the feature sets were extracted by training the datasets by the designed deep network model. In the last step, the feature sets were combined and the classification process was carried out using the Softmax method. Experimental analyzes showed that the proposed approach increased the overall performance and the overall accuracy in the classification process was 100%.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 9 Mart 2023
Kabul : 15 Nisan 2023

Anahtar Kelimeler:

Özellik Çıkarma, Kötü Amaçlı Yazılımlar, 2B Barkod Türleri, Derin Öğrenme

ÖZET

Kötü amaçlı yazılımlar bilgisayar tabanlı sistemlere zarar vermek, önemli bilgileri elde etmek veya değiştirmek amaçlı hazırlanmış yazılımlardır. Bu tür yazılımlar insanların etkileşim içerisinde olduğu ağ ortamlarını hedef alırlar. Bu ağ ortamlarında kullanılan akıllı cihazlar günümüzde hayatımızın vazgeçilmez parçalarından biri olmuştur. Akıllı cihazların güvenliğini sağlayabilmek, zararlı yazılımların tespitini gerçekleştirebilmek için son zamanlarda yapay zekâ tabanlı birçok çalışma gerçekleşmiştir. Bu çalışmanın veri kümesi gizlenmiş kötü amaçlı yazılım türlerini içerisinde barındıran metin tabanlı içeriklerden oluşmaktadır. Önerilen yaklaşım, önişlem adımından ve derin öğrenme modelinden oluşmaktadır. Önişlem adımında metin tabanlı veriler, 2-boyutlu barkod türlerine dönüştürülerek iki yeni veri kümesi elde edilmiştir. Bir sonraki adımda veri kümeleri tasarlanmış derin ağ modeli tarafından eğitilerek özellik setleri çıkartılmıştır. Son adımda özellik setleri birleştirilerek sınıflandırma süreci Softmax yöntemi kullanılarak gerçekleşmiştir. Deneysel analizler önerilen yaklaşımın genel performansı artırdığı görülmüştür ve sınıflandırma sürecinde genel doğruluk başarısı %100 olarak elde edilmiştir.

ORCID: [1]0000-0002-8264-3899

1. GİRİŞ

Son zamanlarda insanların internet ortamındaki etkileşimleri dikkate alındığında gözle görülebilir büyük bir artış yaşanmıştır. Bilgi tabanlı sistemlerin artması, dijital ortamların yoğun ilgi görmesi, sosyal medya üzerinden işlemlerin gerçekleşmesi, vb., durumlardan dolayı internet dünyası kötü amaçlı yazılımcıların hedefi haline gelmiştir. Kötü amaçlı yazılımcılar tarafından geliştirilen yazılımlar, çeşitli saldırı türleri ile sistemleri ele geçirmek, verileri yedeklemek veya değiştirmek, varlığını gizleyerek sürdürmek, kullanıcı hesaplarını ele geçirmek, ağ ortamına bağlı sistemlere zarar vermek, vb. amaçlar doğrultusunda tasarlanmıştır [1]. Panda Security'in 2018 raporuna göre bilgisayar korsanları tarafından günlük 230 bin kötü amaçlı yazılım tasarlanmaktadır. Bu sayı her geçen gün daha da artmaktadır. Kaspersky firmasının 2019 yılı raporuna göre kimlik avı dolandırıcılığında istenmeyen e-postalar ilk sırada yer almaktadır ve bu eğilimin yakın zamanda değişmesi mümkün değildir. Ayrıca siber saldırılara maruz kalan küçük şirketler ekonomik anlamda olumsuz etkilenmişlerdir. Accenture firması, web tabanlı saldırılara karşı her yıl yaklaşık 2.4 milyon dolar harcama yaparak tedbirler almıştır. Siber saldırılar sadece bir grup tarafından gerçekleşebilen eylemler olarak tanımlanmayıp, devletler tarafından da desteklenmektedir. Bu durumun en iyi örneği, İsrail tarafından desteklenmiş ve İran Nükleer Santrali'nin çalışmasını engellemek amacıyla geliştirilmiş Stuxnet adı verilen kötü amaçlı yazılımdır [2]. Kötü amaçlı yazılımların etkisini azalabilmek ve tespitini kolaylaştırmak için bu yazılımların türlerine göre gerçek zamanlı stratejiler geliştirilmektedir. Günümüzde yapay zeka tabanlı yaklaşımlar gerçek zamanlı sistemlere entegre edilerek işlemlerin daha kolaylaştırılması hedeflenmiştir. İnsanüstü karar verebilen bu sistemler gerçek zamanlı işlemlerde daha da etkili rol üstlenebilmektedir [3]. Literatür de bazı çalışmalar bu durumu destekler niteliktedir;

M. Akhtar ve ark. [4] botnet saldırılarını tespit etmek için hibrit bir derin öğrenme yaklaşımı önerdi. Önerdikleri yaklaşım evrişimsel sinir ağı (ESA) ile uzun-kısa süreli bellek (LSTM) modelinden oluşmuştu. Kötü amaçlı yazılım sınıflarının tespitinde önerdikleri yaklaşım ile %99 genel doğruluk başarısı elde etmişlerdi [4]. Ding Yuxin ve Zhu Siyi çalışmasında derin inanç ağlarını (DBN) kullanarak kötü amaçlı yazılımların tespitini başarılı bir şekilde gerçekleştirdi. Sınıflandırma sürecinde önerdikleri DBN modelini destek vektör makineleri (SVM), karar ağaçları ve k-en yakın komşu (kNN) yöntemleri ile kıyaslamışlar ve en iyi performansı DBN modeli ile yaklaşık %98 oranında genel doğruluk başarısı sağlamışlardı [5]. Vinayakumar Ravi ve ark. çalışmasında kötü amaçlı yazılımların tespiti için çok görünümlü dikkat tabanlı derin öğrenme modelini önerdiler. Önerilen yaklaşımı hem windows tabanlı verilerde hem de android tabanlı verilerde uyguladılar ve sırasıyla %98 ile %97 genel doğruluk başarısı elde ettiler [6]. J. Pavithra ve S.Samy çalışmasında web işlemlerinin yer aldığı veri kümesini kullanarak web saldırılarının tespitini gerçekleşmişlerdir. Önerdikleri yaklaşımda makine öğrenme yöntemlerini (SVM, rastgele orman) ve Bayes modelini kullandılar. Deneysel analizlerde en başarılı sonucu rastgele orman yöntemi vermiştir ve bu yöntem ile yaklaşık %99 genel doğruluk başarısı sağlamışlardır [7].

Bu çalışmada kötü amaçlı bellek analizi kayıtlarından oluşan veriler, barkod türlerine dönüştürülerek yeni veri kümeleri elde edildi. Bir sonraki aşamada veri kümeleri tasarlanmış ESA modeli ile eğitildi ve tür tabanlı özellik setleri elde edildi. Ardından tür tabanlı özellik setleri birleştirilerek sınıflandırma işlemi gerçekleşti.

Bu çalışmanın mevcut literatürden farkları şunlardır;

- Kötü niyetli yazılım verileri 1-boyutlu (1B) kayıtlardan oluşur ve 1B kayıtlar 1B-ESA modelleri kullanılarak analizleri gerçekleştirilir. Bu çalışma ile 1B kayıtlar 2-boyutlu (2B) barkod türlerine dönüştürülerek 2B-ESA modeliyle eğitimi gerçekleşti.
- Veri kümesindeki her bir kayıt okunabilir içeriklerden oluşmaktadır. Kayıtlar barkod türlerine dönüştürülerek bu durumun önüne geçildi ve veri güvenliği ön plana çıkartıldı.
- 2B-ESA modelinin sunmuş olduğu olanaklardan (tür tabanlı özellik seti elde etme, özellik birleştirme) yararlanıldı.

Bu çalışmanın amacı ve hedefleri şu şekilde özetlenir;

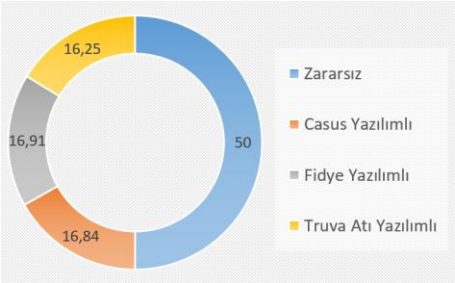
- Kötü amaçlı yazılım türlerini (casus yazılımı, fidye yazılımı, trojan) normal veri bilgilerinden ayırt edebilecek bir yaklaşım geliştirmeyi amaçlamıştır.
- Önişlem adımı olarak metin tabanlı verileri 2B barkod türlerine (veri matrisi, aztek) dönüştürerek elde edilen görüntü verileri, önerilen 2B-ESA modele girdi olarak verilerek başarılı bir sınıflandırma gerçekleştirmesi amaçlanmıştır.
- Ön işlem adımı sayesinde metin tabanlı veriler 2B-ESA modeli tarafından da eğitilebilecek,
- ESA modeli tarafından çıkartılan öznitelik setleri birleştirilerek sınıflandırma sürecine katkı sağlanması amaçlanmıştır.

Bu makale şu şekilde özetlenir; kullanılan materyal hakkında bilgiler Bölüm 2'de verilmiştir. Önerilen yaklaşımda kullanılan yöntemler ve modeller hakkında bilgiler Bölüm 3'te verilmiştir. Deneysel analizler ve analizlerin karşılaştırılması hakkında detaylı bilgiler Bölüm 4'te verilmiştir. Tartışma ve Sonuç bölümü makalenin son bölümünde yer almıştır.

2. CIC-MALMEM-2022 VERİ KÜMESİ

CIC-MalMem-2022, Kanada Siber Güvenlik Enstitüsü (CIC) araştırmacıları tarafından oluşturulmuş ve kötü amaçlı yazılım bellek analizi (MalMem) verilerini içeren açık erişimli bir veri kümesidir. Gizlenmiş kötü amaçlı

yazılımlar, tespit edilmekten veya yok edilmekten kaçınarak bellek gibi donanımsal yerlerde varlıklarını sürdürürler. Bu veri kümesi, ortalama bir kullanıcının bir kötü amaçlı yazılım saldırısı sırasında neleri çalıştıracağının daha doğru bir örneğini temsil etmektedir. CIC-MalMem-2022, iyi amaçlı veriler ve kötü amaçlı veriler olmak üzere dengeli bir dağılım gerçekleştirilerek oluşturulmuştur. Kötü amaçlı yazılım verilerinin dökümü incelendiğinde 29.298'i zararsız ve 29.298'i kötü amaçlı olmak üzere toplam 58.596 kayıt içermektedir. Kötü amaçlı yazılımlar; casus, fidye ve truva atı olmak üzere üç türden oluşmaktadır. Veri türlerinin içerdiği kayıt türlerine göre yüzdelik oranları Şekil 1'de gösterilmiştir. Her bir kayıt 57 adet öznitelikten oluşmaktadır ve bu özniteliklerin kategori ve sınıf öznitelikleri hariç diğer öznitelikler sayısal formatlardan oluşmaktadır [8].



Şekil 1. Veri kümesinin türleri ve yüzdelik dilimleri (%).

Deneysel analizlerin donanımsal gereksinimleri de dikkate alınarak orijinal veri kümesinin tamamı bu çalışma da kullanılmadı. Veri kümesinin her bir türünden 500 kayıt rastgele seçilerek yeni bir veri kümesi oluşturuldu. Bu çalışma için kullanılan veri kümesi türleri ve kullanılan kayıt sayısı hakkında bilgiler Tablo 1'de verilmiştir.

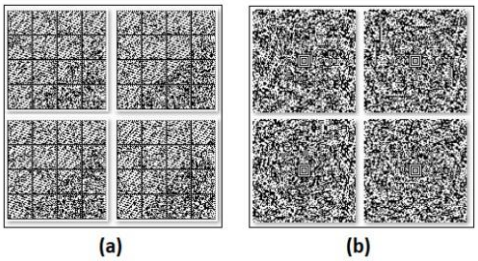
Tablo 1. Bu çalışmanın deneysel analizlerinde kullanılan veri kümesi türleri ve istatistik bilgisi.

Veri Türü	Kayıt Sayısı
Zararsız	500
Casus yazılım	500
Fidye yazılım	500
Truva atı yazılım	500
Toplam	2000

3. YÖNTEMLER VE YAKLAŞIMLAR

3.1. Veri Matrisi ve Aztek Kodu

Veri matrisleri büyük miktarda veriyi kodlamak için tercih edilen, siyah beyaz kare modüller şeklinde kodlayan iki boyutlu bir matris barkod türüdür. Kısacası çok yüksek veri saklama yoğunluğuna sahip 2B barkod sembolojisidir. Veri matrisleri genellikle kare şeklinde iki boyutlu bir görüntü ile temsil edilir; bazen dikdörtgen görüntülerden de oluşabilmektedir. Veri matrislerinin her bir noktası bit ile temsil edilir. Tek bir veri matrisi 2335 alfa sayısal karakteri destekler veya 1556 bayta kadar veri tutabilir. Bu kapasite veri sıkıştırma algoritmaları ile birlikte daha da artabilmektedir. Veri matrislerinin hata düzeltme seviyeleri de mevcuttur. Yaklaşık %25'e kadar okunmayan veri matrisi barkodunu veri kaybı söz konusu olmadan yeniden yüklemek ve okumak mümkündür. Veri matrisleri genel olarak posta hizmetleri, tıbbi/sağlık endüstrisi, genel lojistik amaçlar, belge yönetim uygulamalarında sıkça tercih edilmektedir [9]. Veri matrisi örnekleri Şekil 2(a)'da gösterilmiştir.
Aztek kodu, piramidin havadan görünümüne benzeyen merkezinde bulucu kodu sayesinde bu isim verilmiştir. Araç tescil belgelerinde, seyahat belgeleri, uçak biletlerinde sıkça kullanılan bu kodlama, 2B bir matris barkod türüdür. Çoğu 2B matris kodlarının aksine aztek kodu ile tasarlanmış bir barkod, kenarında sessiz bir gölge barındırmaz. Böylece daha küçük alanları daha verimli kullanarak büyük verileri depolama özelliğine sahip olurlar. Ayrıca veri büyüklüğüne bağlı olarak aztek kodların boyutları da değişebilir ve potansiyel olarak çok miktarda bilgileri tutabilirler. Aztek kodların QR kodlardan üç köşesinde üç bulucu desen bulunmaz, merkezinde bulucu desen yer alır [10]. Aztek kodları 3832 numerik değerleri destekler ve 1914 bayta kadar veri tutabilir [9]. Aztek kodunu temsil eden örnek Şekil 2(b)'de gösterilmiştir.

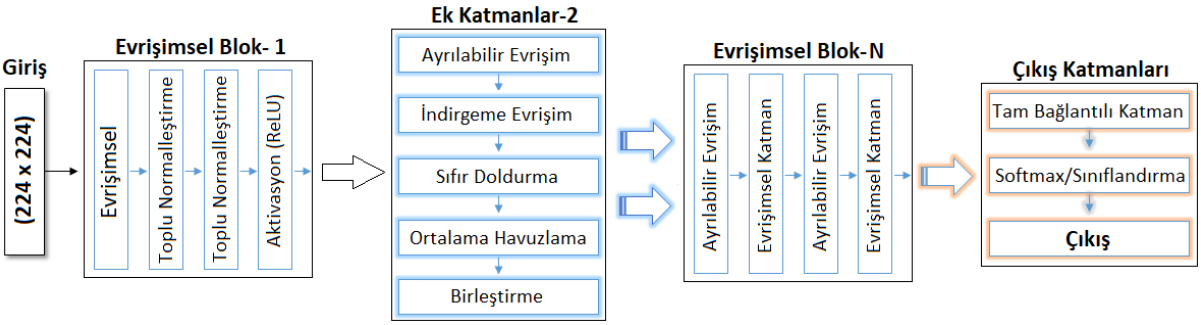


Şekil 2. Veri kümesinden elde edilmiş 4 kayıt örneğinin barkod gösterimi; a) veri matrisi, b) aztek kodu.

3.2. Tasarlanmış Derin Ağ Modeli

ESA modelleri genel olarak üç bileşenden oluşur; Evrişimsel katman, havuzlama katmanı ve tam bağlantılı katman. Evrişimsel katmanlar girdi görüntülerini bir filtre aracılığıyla dolaştırarak özniteliklerin çıkartılmasını sağlar. Bunu gerçekleştirirken çıktı olarak aktivasyon haritalarını oluştururlar. Aktivasyon haritaları havuzlama katmanı kullanılarak daha düşük boyutlu özniteliklere dönüştürülür [11]. Bu süreç ESA modelinin mimari yapısına bağlı olarak belirli aralıklarla tekrarlanır. Havuzlama katmanı bu sayede ESA modelinin aşırı öğrenmesini engeller. Toplu normalleşme (batch normalization), bir önceki katmanın öğrenmesini beklemeden eş zamanlı olarak öğrenme olanağı sağlar [12]. Girdi verisi iki boyutlu matristen oluşmaktadır. Bu matris çekirdek olarak ifade edilmektedir ve ana çekirdek iki küçük çekirdeğe bölünecek şekilde ayrılması işlemine ayrılabilir evrişimsel (separable convolution) denilmektedir [13]. İndirgeme (reduction) evrişimsel katman sayesinde girdi boyutları bir sonraki katmana düşürülerek aktarılması sağlanır. Sıfır doldurma (zero padding) bir önceki katman çıkış boyutunu korumaya müsaade eder. Yani, çıkış görüntüsünün kenarlarına tamamı sıfır değerine sahip bir piksel kenarlığı eklenmesini sağlar [14]. ESA modelinin son aşamasında tam bağlantılı katmanlar kullanılır ve tam bağlantılı katmanlar önceki katmanlardan çıkartılmış nöronları/ düğümleri tek bir katmanda düzleştirir. Yani nöronlar çıkış katmanına düzleştirilerek bağlanır. Son adımda girdi türlerinin sınıflandırılma süreci ile ilgili olasılık değerlerin hesaplanması gerçekleşir [15]. Genel olarak ESA modellerinde farklı bir sınıflandırıcı kullanılmadığı takdirde tercih edilen sınıflandırıcı fonksiyonu Softmax'tır.
Softmax fonksiyonu ESA modellerinin ham çıktısını normalleştirmek olasılık değerleri elde etmek için son katmanda genellikle tercih edilir. Softmax fonksiyonun matematiksel işlevini gösteren denklem aşağıda verilmiştir. Denklem 1 incelendiğinde; x_i bir makine öğrenimi sonucu elde edilmiş ağın çıktılarını temsil eder. $i = 1,..,N$ arası örneklem değerlerinden oluşur. P_i değişkeni, i. elemanın olasılık değerini temsil eder [16].

$$P_i = \frac{e^{x_i}}{\sum_{k=1}^{N} e^{x_k}} \tag{1}$$

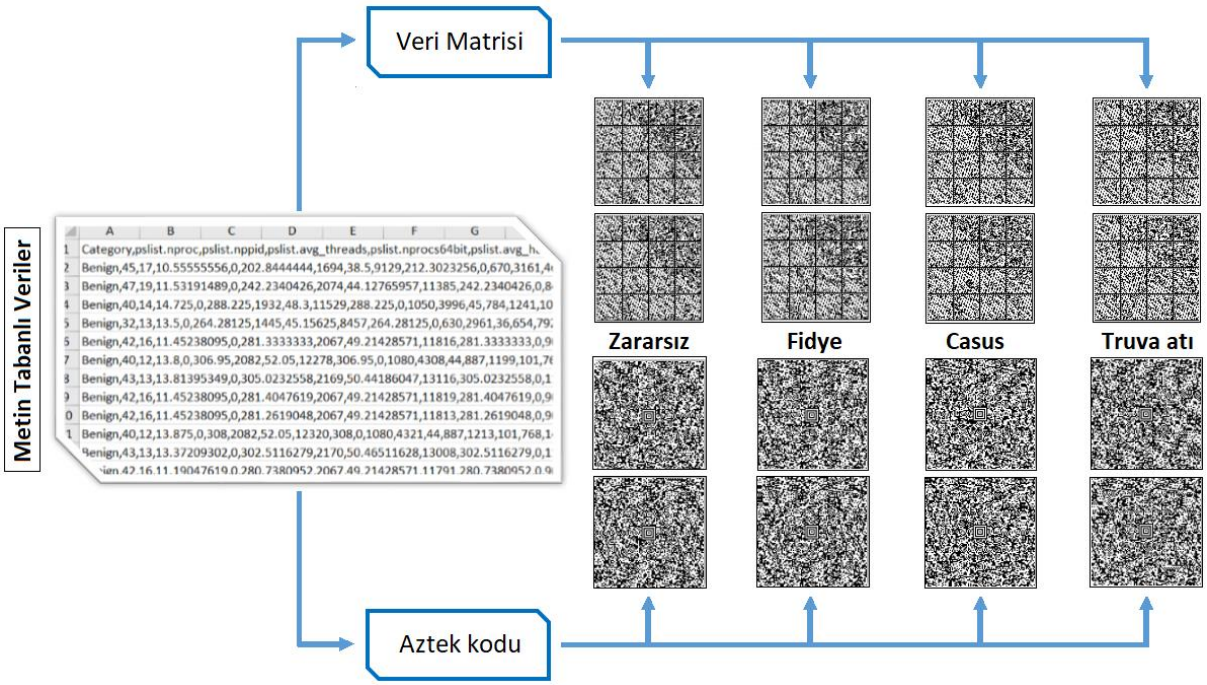


Şekil 3. Tasarlanmış ESA modelinin katman yapısı.

Bu çalışmanın analizleri için Keras Kütüphanesi kullanılarak özgün bir ESA modeli tasarlanmıştır. Tasarlanan ESA modeli Şekil 3'te gösterildi. Önerilen modelin girdi çözünürlüğü 224×224'tür ve İmageNet veri kümesi kullanılarak önceden eğitilmiş ağırlık parametreleri elde edilerek tasarlandı. Öğrenme oranı 1e-4 tercih edildi ve optimizasyon yöntemi olarak stokastik gradyan inişi (SGD) seçildi. Tam bağlantılı katman her bir girdi verisi için 1000 öznitelik verecek şekilde ayarlandı. Sınıflandırma sürecinde çıkış, $tür\ sayısı\ (zararsız, casus, fidye, truva\ atı) \times görüntü\ sayısı$ öznitelik seti verecek şekilde oluşturuldu.

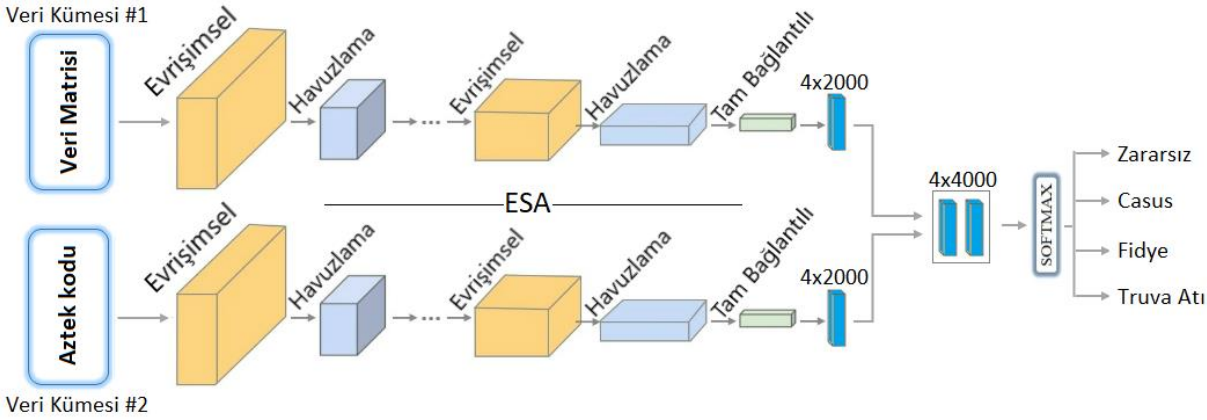
3.3. Önerilen Yaklaşım

Önerilen yaklaşım, metin içerikli verileri 2B barkod türlerine dönüştürerek 2B-ESA modeli tarafından başarılı bir şekilde eğitilmesini sağlamak amacıyla tasarlanmıştır. Literatürdeki çalışmalarda kullanılan 1B veriler, 1B-ESA modelleri tarafından eğitilerek analizler gerçekleştirilmiştir. Bu çalışma da ise hedef, 1B veri kümesini 2B görüntü verilerine dönüştürmektir ve ardından 2B olarak tasarlanmış ESA modeliyle eğitilmesini başarılı bir şekilde gerçekleştirmektir. Buradan elde edilecek başarı diğer 2B-ESA modelleri tarafından da başarılı bir şekilde kullanabileceği öngörülmektedir. Önerilen yaklaşım iki aşamadan oluşmaktadır;
Birinci aşamada 1B metin tabanlı kayıtlar 2B barkod türlerine dönüştürülerek 2B görüntüler elde edildi. 2B barkod türlerine dönüştürme işleminde iki teknik kullanıldı. Bu teknikler; aztek kodu (aztec code) [17] ve veri matrisi (data matrix) [18]. 1B veri kümesinin her satırı sırasıyla Python dilinde çağrılarak aztek ve veri matrisi teknikleri ile derlendi. Ardından her bir kayıt satırı için barkod türlerine ait 2B görüntüler elde edildi. Bu aşama sonunda iki adet veri kümesi oluşturuldu (aztek ve veri matrisi teknikleri aracılığıyla). Birinci aşamada elde edilen görüntü kümesine ait örnek görüntüler Şekil 4'te gösterilmiştir.



Şekil 4. Metin tabanlı kayıtların 2B Barkod türlerine dönüştürülmesi işlemi.

İkinci aşamada, 2B-veri kümeleri (veri matrisi ve aztek kodu ile oluşturulmuş veri kümesi) tasarlanmış ESA modeli tarafından eğitildi ve modelin tam bağlantılı son katmanından tür tabanlı öznitelik setleri çıkartıldı. Ardından özellik birleştirme (feature fusion) tekniği ile iki özellik seti birleştirildi. Burada amaç genel doğruluk performansını artırmaktı. Ardından birleştirilmiş öznitelik seti Softmax tarafından yeniden sınıflandırıldı. İkinci aşamada gerçekleştirilen işlemler ve önerilen yaklaşımın genel tasarımı Şekil 5'te gösterilmiştir.



Şekil 5. Önerilen yaklaşımın genel tasarımı.

4. DENEYSEL ANALİZLER

Bu çalışmada önişlem adımları ve tasarlanmış ESA modeli Pyhton diliyle kodlandı ve analizler için Google Colab sunucusu kullanıldı. Python kodları Jupyter Notebook platformu üzerinden derlendi. Analiz ölçümlerinde karmaşıklık matrisi tercih edildi [19]. Karmaşıklık matrisinin hesaplanmasında kullanılan metrikler şunlardır; geri çağırma (G_çağ), kesinlik (kes), f-skor (f-skr) ve doğruluk (doğ) [20]. Metriklerin hesaplanma işlemleri için Denklem 2-5'te yer alan matematiksel formüller kullanıldı. İlgili denklemlerde; (D): doğru, (Y): yanlış, (N): negatif, (P): pozitif anlamı taşımaktadır [21, 22]. Tasarlanmış ESA modeli için tercih edilen parametre değerleri Tablo 2'de verilmiştir.

$$\text{G\_çağ} = \frac{DP}{DP+YN} \quad (2)$$

$$\text{Kes} = \frac{DP}{DP+YP} \quad (3)$$

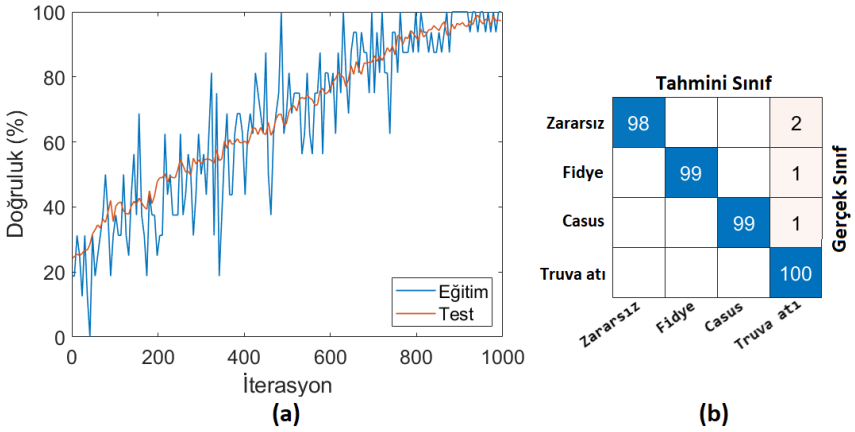
$$\text{F-skr} = \frac{2\times DP}{2\times DP+YP+YN} \quad (4)$$

$$\text{Doğ.} = \frac{DP+DN}{DP+DN+YP+YN} \quad (5)$$

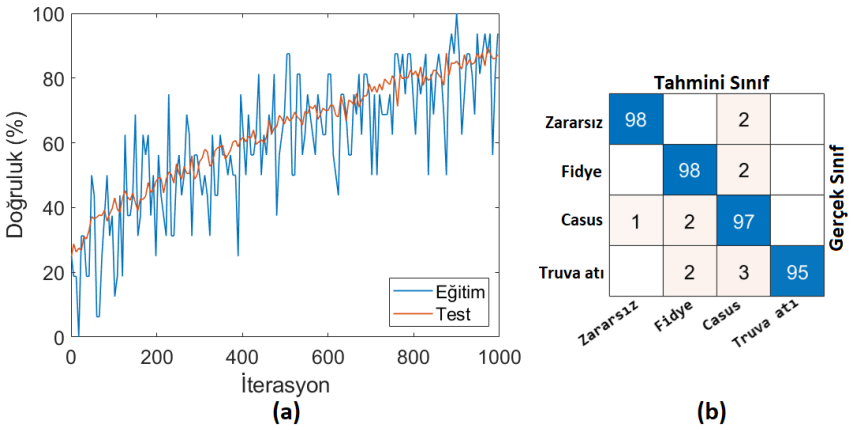
Tablo 2. ESA modeli için tercih edilmiş parametre değerleri.

Parametre	Tercih / Değer
İterasyon sayısı	1000
Öğrenme oranı	1e-4
Optimizasyon	SGD
Sınıflandırıcı	Softmax
Aktivasyon fonksiyonu	ReLU
Evrişimsel katman sayısı	36
Evrişimsel katman filtre boyutu	(3×3) ve (5×5)
Havuzlama katman filtre boyutu	(3×3)
Donanım kaynağı	Tekli GPU
Mini – topluluk (mini-batch)	16
Eğitim oranı: test oranı	0.8:0.2

Deneysel analizler iki adımdan oluşmuştur. Birinci adımda bu çalışma için tasarlanmış 2B-ESA modeli tarafından 2B-veri kümelerinin (veri matrisi ve aztek kodu) eğitimi gerçekleşti. Veri matrisi ile oluşturulmuş veri kümesinin eğitim-test başarı grafikleri ve karmaşıklık matrisi Şekil 6'da gösterilmiştir. Aztek kodu ile oluşturulmuş veri kümesinin eğitim-test başarı grafikleri ve karmaşıklık matrisi Şekil 7'de gösterilmiştir. 2B- veri kümelerinin tasarlanmış ESA modeli ile eğitiminden elde edilmiş metrik sonuçları Tablo 3'te verilmiştir. Tablo 3'teki sonuçlar incelendiğinde veri matrisi ön işleme adımıyla elde edilmiş veri kümesinin ESA modeli tarafından eğitimi aztek koduyla işlenmiş veri kümesine göre daha başarılı sonuç verdiği gözlemlenmiştir. Veri matrisi ile işlenmiş veri kümesinin ESA modeliyle eğitim sonucunun genel doğruluk başarısı %99'du ve aztek kodu ile işlenmiş veri kümesinin ESA modeliyle eğitim sonucunun genel doğruluk başarısı %97'ydi.



Şekil 6. Veri matrisi ile oluşturulmuş veri kümesinin ESA modeli ile eğitim analizi; a) doğruluk grafiği, b) karmaşıklık matrisi.

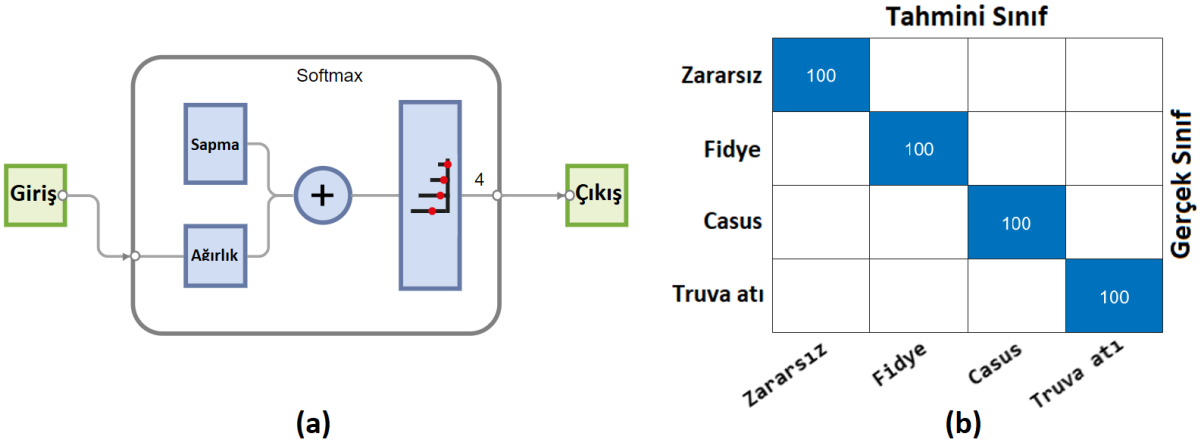


Şekil 7. Aztek kodu ile oluşturulmuş veri kümesinin ESA modeli ile eğitim analizi; a) doğruluk grafiği, b) karmaşıklık matrisi.

Tablo 3. Önerilen ESA modeli ile elde edilmiş karmaşıklık matrisi sonuçları.

Ön işlem	Sınıf	Kes	G_çağ	F-skr	Doğ. (%)	Genel Doğ. (%)
Veri matrisi	Zararsız	0,98	1,0	0,99	99,50	99,0
	Fidye	0,99	1,0	0,99	99,75	
	Casus	0,99	1,0	0,99	99,75	
	Truva atı	1,0	0,96	0,98	99,0	
Aztek kodu	Zararsız	0,98	0,99	0,98	99,25	97,0
	Fidye	0,98	0,96	0,97	98,50	
	Casus	0,97	0,93	0,95	97,50	
	Truva atı	0,95	1,0	0,97	98,75	

Deneysel analizin ikinci adımında ESA modelinin son katmanından çıkartılmış tür tabanlı öznitelik seti her iki veri kümesi için birleştirilerek ($2x$ [$sınıf\ sayısı \times görüntü\ sayısı$]) önerilen yaklaşımın genel performansı artırılması amaçlanmıştır. Her bir veri kümesinden çıkartılmış 4×2000 öznitelik setleri birleştirilerek 4×4000 boyutunda öznitelik seti oluşturuldu. Ardından softmax yöntemi ile birleştirilmiş öznitelik seti yeniden sınıflandırıldı. Softmax yönteminin mimari yapısı ve sınıflandırma sonucunda elde edilmiş karmaşıklık matrisi Şekil 8'de gösterilmiştir. Deneysel analizin ikinci adımında öznitelik setlerin birleştirilmesi genel performansı artırdığı gözlemlenmiştir ve %100 genel doğruluk başarısı elde edilmiştir. Bu adımda elde edilmiş metrik sonuçları Tablo 4'te verilmiştir.



Şekil 8. Birleştirilmiş öznitelik setinin softmax yöntemi ile sınıflandırılması; a) softmax'ın işlevi, b) karmaşıklık matrisi.

Tablo 4. Önerilen yaklaşım ile elde edilmiş karmaşıklık matrisinin metrik sonuçları.

Model	Sınıf	Kes	G_çağ	F-skr	Genel Doğ. (%)
Önerilen yaklaşım	Zararsız	1,0	1,0	1,0	100
	Fidye	1,0	1,0	1,0	
	Casus	1,0	1,0	1,0	
	Truva atı	1,0	1,0	1,0	

5. TARTIŞMA VE SONUÇ

Yakın zamanda teknolojik gelişmeler ile beraber insanların internet dünyasıyla etkileşimi artmıştır. Bununla beraber kötü amaçlı yazılımcıların çeşitli senaryolar ortaya koyarak internet kullanıcılarını kendi amaçları doğrultusunda tuzağa çekme oranlarında da artış yaşanmıştır. Dolayısıyla ağ ortamında çok sayıda kullanıcıyı eş zamanlı kontrol edebilmek ve güvenliğini sağlayabilmek için birçok analiz gerçekleşmiştir. Özellikle eşzamanlı kontrollerin zaman, hız ve doğruluk parametrelerini verimli bir şekilde kullanabilmek önem arz etmektedir. Yapay zekâ yaklaşımları eş zamanlı kontrolleri gerçekleştirebilmek için son zamanlarda birçok çalışmada tercih edilmiştir. Bu çalışma, kötü amaçlı yazılımların tespitinde metin tabanlı verilerin 2B görüntülere dönüştürülerek 2B-ESA modeliyle eğitimini başarılı bir şekilde ayırt edebildiğini göstermiştir. Ayrıca hızdan kazanç ve zamandan tasarruf sağlamak amacıyla orijinal veri kümesinin tüm kayıtları değil de belirli bir oranda (her bir sınıf için 500 kayıt) kayıtları rastgele seçilerek eğitilmiştir. Son aşamada ESA modelinin son katmanından tür tabanlı öznitelik setleri çıkartılarak birleştirilmiş ve softmax ile yeniden sınıflandırılmıştır. Sonuç olarak %100 genel performans

başarısı sağlanmıştır. Önerilen yaklaşım, kötü amaçlı yazılımların tespitin de başarılı bir rol üstlenmiştir. Önerilen yaklaşımın ana katkıları şunlardır;

- 1B verilerin 2B-ESA modelleri ile eğitilmesinin önü açılmıştır. Bunu sağlamak için 1B veriler 2B barkod türlerine dönüştürülmüştür.
- 2B-ESA modeli ile başarılı sonuçlar elde edilmiştir.
- 1B veri kayıtları 2B barkod türlerine dönüştürülerek veri güvenliği önplanda tutulmuştur.
- Tasarlanmış ESA modelinin son katmanın da tür/sınıf tabanlı özellik setleri elde edilmiştir.
- Özellik setlerini birleştirerek (feature fusion), önerilen yaklaşımın performans artısını sağlamıştır.

Önerilen yaklaşımın sınırlılıkları arasında donanımsal yetersizlikten dolayı veri kümesinin tüm kayıtları analizlerde kullanılmadı. Ayrıca uçtan uca bir model olmaması belirli aşamalarda zaman kaybına yol açmıştır.

Bu çalışmada kullanılan veri kümesi 2022 yılında yayınlanmış ve aynı veri kümesini kullanan diğer çalışmalar karşılaştırılması Tablo 5'te gösterilmiştir. Louk ve ark. [23] çalışmasında ağaç tabanlı topluluk öğrenme yöntemlerini kullanarak analizlerini gerçekleştirmiştir. Klasik yöntemlerle analizlerin gerçekleştirilmesi sonucu en iyi performansı XGBoost ve rastgele orman yöntemleri ile elde etmiştir. Bu yöntemler ile yaklaşık %99 genel doğruluk başarısı elde etmişlerdir. Dener ve ark. [24] çalışmasında ağaç tabanlı topluluk öğrenme yöntemleri ile makine öğrenme yöntemlerini kullanmışlar. Çalışmalarında klasik yöntemler ile veri kümesinin tamamını analiz ettiler ve analizler sonucunda en iyi performansı lojistik regresyon yöntemi ile sağladılar. Bu yöntem ile %99,97 genel doğruluk başarısı elde etmişlerdir. Louk [23] ve Dener [24] çalışmalarında veri kümesini zararsız ve zararlı yazılımlar olmak üzere ikili sınıflandırma gerçekleştirmiştir. Bu çalışmanın analizlerinde ise zararsız veriler ile beraber kötü amaçlı yazılımların (casus, fidye, Truva atı) verileri kullanıldı. Mezina ve Burget [25], genişletilmiş 1B-ESA tabanlı model tasarladılar. Sınıflandırıcı olarak softmax yöntemini kullandılar. Sonuç olarak önerdikleri yaklaşımda %83,53 genel doğruluk performansı elde ettiler. Mezina ve Burget [25] önerdikleri yaklaşımda geleneksel yöntemler içerdiği için yenilikçi yönü sınırlı kalmıştır. Talukder ve ark. [26] veri kümesinin sayısını azaltmak ve dengelemek için SMOTE tekniğini kullandılar. Özellik seçimi için XGBoost yöntemini ve sınıflandırma sürecinde rastgele orman yöntemini kullandılar. İkili olarak gerçekleştirdikleri sınıflandırma (zararlı ve zararsız) işleminde %100 genel doğruluk performansı elde ettiler.

Tablo 5. Önerilen yaklaşımın diğer çalışmalar ile karşılaştırılması.

Makale	**Sınıf Sayısı**	**Yöntem / Model**	**Genel Doğ.** (%)
M.H. Louk ve ark. [23]	2	Ağaç Tabanlı Topluluk Öğrenimi	99
M. Dener ve ark. [24]	2	Topluluk Öğrenme / Makine Öğrenme	99,97
A. Mezina ve R. Burget [25]	4	Genişletilmiş 1B- ESA	83,53
M.A. Talukder ve ark. [26]	2	SMOTE & Makine Öğrenme	100
Önerilen Yaklaşım	4	2B Barkod dönüştürme / Tasarlanmış ESA / Özellik birleştirme	100

Gelecek çalışmada, metin tabanlı verilerin 2B diğer barkod yöntemleri kullanılarak analizleri gerçekleştirilecektir. Barkod türleri arasında en iyi performansı verebilen yöntem belirlenerek çeşitli 2B-ESA modelleriyle eğitimi gerçekleştirilecektir. Son adımda ise ESA modellerinden çıkartılmış öznitelik setleri meta-sezgisel algoritmalara girdi olarak verilerek en verimli öznitelik setleri seçilecek ve sınıflandırılacaktır.

Yazar Katkıları

Mesut Toğaçar: Fikrin ortaya çıkmasını sağlamış, deneysel analizleri gerçekleştirmiş ve makaleyi yazmıştır.

Çıkar Çatışması

Makale yazarı, herhangi bir çıkar çatışması olmadığını beyan eder.

KAYNAKÇA

[1] M.D. Yılmaz "Malware classification with using deep learning", Comput. Informatics, vol. 2, no. 2, pp. 21–40, 2022.

[2] U.H. Tayyab, F.B. Khan, M.H. Durad, A. Khan, and Y.S. Lee "A Survey of the Recent Trends in Deep Learning Based Malware Detection", J. Cybersecurity Priv., vol. 2, no. 4, pp. 800–829, 2022.

[3] M. Toğaçar "Siber Saldırılara Karşı Kullanılan Makine Öğrenme Yöntemlerinin Web Uygulamalarında Güvenlik Etkinliğinin Ölçümü", Gazi Üniversitesi Fen Bilim. Derg. Part C Tasarım ve Teknol., vol. 9, no. 4, pp. 608–620, 2021.

[4] M.S. Akhtar and T. Feng "Detection of Malware by Deep Learning as CNN-LSTM Machine Learning Techniques in Real Time", Symmetry (Basel)., vol. 14, no. 11, pp. 2308, 2022.

[5] D. Yuxin and Z. Siyi "Malware detection based on deep learning algorithm", Neural Comput. Appl., vol. 31, no. 2, pp. 461–472, 2019.

[6] V. Ravi, M. Alazab, S. Selvaganapathy, and R. Chaganti "A Multi-View attention-based deep learning framework for malware detection in smart healthcare systems", Comput. Commun., vol. 195, pp. 73–81, 2022.

[7] J. Pavithra and S. Selvakumara Samy "A Comparative Study on Detection of Malware and Benign on the Internet Using Machine Learning Classifiers", Math. Probl. Eng., vol. 2022, pp. 1–8, 2022.

[8] T. Carrier, P. Victor, A. Tekeoğlu, and A. Lashkari "Malware Memory Analysis", UNB, 2022.

[9] D. G. Tec-it "Data Matrix (ECC200) - 2D Barcode", 2022.

[10] C. Center "Aztec Codes", url: https://www.cognex.com/resources/symbologies/2-d-matrix-codes/aztec-codes, (Erişim Tarihi: 12/03/2023).

[11] A. Zafar "A Comparison of Pooling Methods for Convolutional Neural Networks", Appl. Sci., vol. 12, no. 17, p. 8643, 2022.

[12] C. Garbin, X. Zhu, and O. Marques "Dropout vs. Batch Normalization: An Empirical Study of Their Impact to Deep Learning", Multimed. Tools Appl., vol. 79, no. 19, pp. 12777–12815, 2020.

[13] T. Huang, J. Chen, and L. Jiang "DS-UNeXt: depthwise separable convolution network with large convolutional kernel for medical image segmentation", Signal Image Video Process, 2022.

[14] M.R.A. Bacha, A. Oukebdane, and A. Hafid Belbachir "Implementation of the zero-padding interpolation technique to improve angular resolution of X-ray tomographic acquisition system", Pattern Recognit. Image Anal., vol. 26, no. 4, pp. 817–823, 2016.

[15] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study", Neural Comput. Appl., vol. 34, no. 7, pp. 5321–5347, 2022.

[16] G.C. Cardarilli et al. "A pseudo-softmax function for hardware-based high speed image classification", Sci. Rep., vol. 11, no. 1, p. 15307, 2021.

[17] D. Alimov "Aztec Code generator", url: https://pypi.org/project/aztec-code-generator/, (Erişim Tarihi: 21/02/2023).

[18] X. Ling Yushulx "Barcode Image Composer", url: https://pypi.org/project/barcode-image-composer/, (Erişim Tarihi: 21/02/2023).

[19] E. Başaran "Classification of white blood cells with SVM by selecting SqueezeNet and LIME properties by mRMR method", Signal Image Video Process., vol. 16, no. 7, pp. 1821–1829, 2022.

[20] A. Çalışkan "Detecting human activity types from 3D posture data using deep learning models", Biomed. Signal Process. Control, vol. 81, p. 104479, 2023.

[21] A. Ari "Multipath feature fusion for hyperspectral image classification based on hybrid 3D/2D CNN and squeeze-excitation network", Earth Sci. Informatics, vol. 16, no. 1, pp. 175–191, 2023.

[22] M. Toğaçar, B. Ergen, and Z. Cömert "Detection of weather images by using spiking neural networks of deep learning models", Neural Comput. Appl., vol. 33, no. 11, pp. 6147–6159, 2021.

[23] M.H.L. Louk and B.A. Tama "Tree-Based Classifier Ensembles for PE Malware Analysis: A Performance Revisit", Algorithms, vol. 15, no. 9, p. 332, 2022.

[24] M. Dener, G. Ok, and A. Orman "Malware Detection Using Memory Analysis Data in Big Data Environment", Appl. Sci., vol. 12, no. 17, p. 8604, 2022.

[25] A. Mezina and R. Burget "Obfuscated malware detection using dilated convolutional network", 14th Int. Congr. Ultra Mod. Telecommun. Control Syst. Work., p. 110–115, 2022.

[26] M.A. Talukder et al. "A dependable hybrid machine learning model for network intrusion detection", J. Inf Secur. Appl., no. 72, p. 103405, 2023.