

BLG447 VIZE PROJESİ – GLCM TABANLI SİBER TEHDİTLERİN GÖRSELLEŞTİRİLMESİ RAPORU

ZAHRA ESPARGHAMİ 230404904

MOHAMMADPARSA MALEK 230404929

NARDIN FARIDASL 230404956

PROJE TANIMI

Bu projede amaç, zararlı yazılım (malware) ailelerini otomatik olarak sınıflandıran bir sistem geliştirmektir. Günümüzde zararlı yazılımlar hızla çeşitlenmekte ve klasik imza tabanlı antivirüs yaklaşımları, özellikle yeni veya gizlenmiş (obfuscated) zararlı yazılımlar karşısında yetersiz kalabilmektedir. Bu nedenle son yıllarda, zararlı yazılım ikili dosyalarının görüntüye dönüştürülerek analiz edilmesi, alternatif ve etkili bir yaklaşım olarak öne çıkmaktadır.

Bu çalışmada MaleVis veri seti kullanılmıştır. MaleVis, zararlı yazılım ikili dosyalarının byte dizimlerinden oluşturulmuş gri seviye görüntülerden oluşan, çok sınıflı ve gerçekçi bir veri setidir. Her zararlı yazılım ailesi, ikili dosyanın yapısal özelliklerine bağlı olarak kendine özgü bir doku (texture) deseni üretmektedir. Bu nedenle görüntü işleme alanında yaygın olarak kullanılan doku tabanlı özellikler, malware ailelerini ayırt etmek için güçlü bir temsil sunmaktadır.

Proje kapsamında, MaleVis veri setindeki görüntülerden başta GLCM (Gray Level Co-occurrence Matrix) olmak üzere Wavelet, Gabor, LBP, HOG-özet ve global istatistik özelliklerini çıkarılmıştır. Elde edilen özellikler kullanılarak Weka ortamında çeşitli deneyler gerçekleştirilmiş; GLCM'nin açı ve mesafe parametrelerinin, farklı özellik gruplarının ve özellik birleşimlerinin sınıflandırma başarımına etkisi sistematik olarak incelenmiştir. Performans değerlendirmesinde yalnızca doğruluk (Accuracy) değil, sınıflandırma kalitesini daha iyi yansitan Cohen's Kappa metriği de kullanılmıştır.

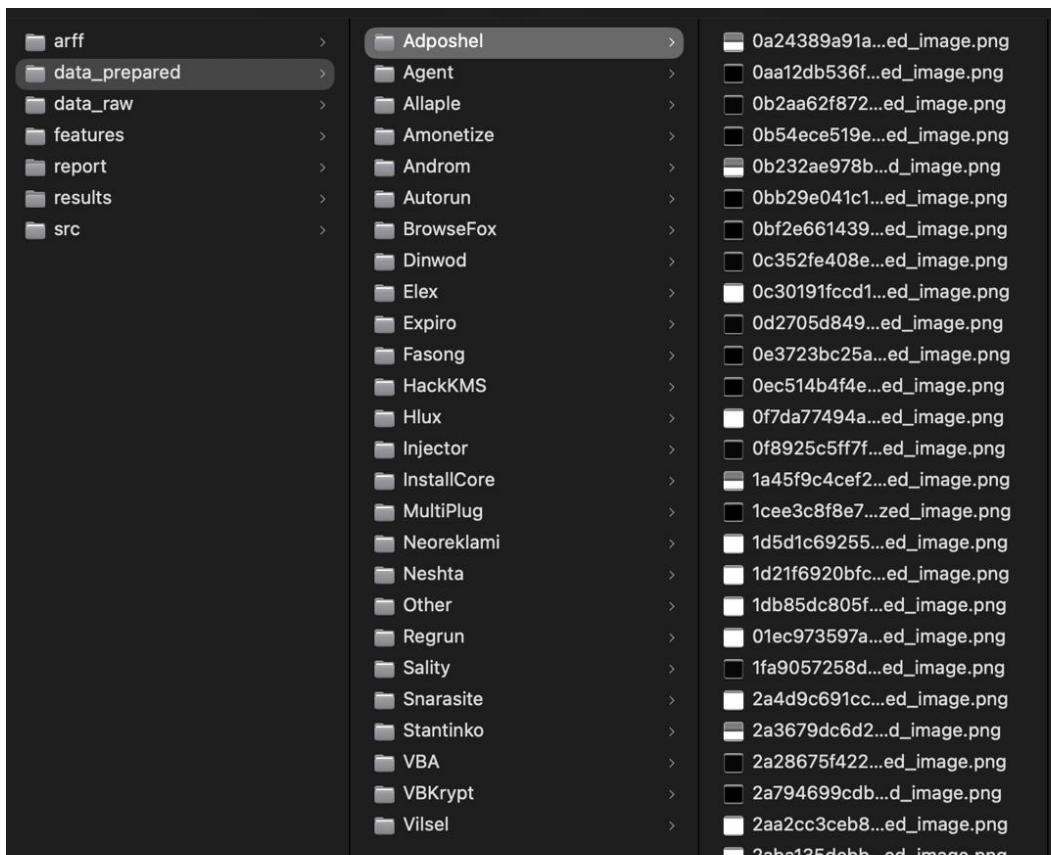
1. PROJENIN AMACI

Bu çalışma, MaleVis veri seti üzerinde görüntü tabanlı doku özelliklerinin zararlı yazılım ailelerini ayırt etmede etkili olduğunu ve özellikle GLCM tabanlı özelliklerin sınıflandırma başarımına önemli katkı sağladığını göstermektedir.

2. VERİ SETİ

Bu çalışmada MaleVis dataset'i kullanılmıştır. MaleVis, Hacettepe Üniversitesi tarafından oluşturulan ve 9.100 adet RGB görsel içeren bir malware görüntü veri setidir. Veri seti 26 farklı sınıf içerir: bunlardan 25'i malware (zararlı yazılım) ailelerini, 1'i ise clean/benign (zararsız/temiz) sınıfını temsil eder. Her sınıf, farklı malware türlerini temsil eden görüntü örnekleri içerir ve bu yapı sayesinde çok sınıflı sınıflandırma problemlerini çalışmak için uygundur.

ÖZELLİK	AÇIKLAMA
Data Set Adı	MaleVis
Sınıf Sayısı	26
Toplam görüntü	9.100
Her sınıfta görüntü sayısı	350
Kaynak	Kaggle



```

1 import os
2 import cv2
3 import numpy as np
4 import pandas as pd
5 from tqdm import tqdm
6
7 from skimage.feature import graycomatrix, graycoprops, local_binary_pattern
8 from skimage.filters import gabor
9 from skimage import img_as_ubyte
10 import pywt
11 import arff
12
13 # ====== PATHS ======
14 BASE_DIR = os.path.abspath(os.path.join(os.path.dirname(__file__), '..'))
15 DATASET_DIR = os.path.join(BASE_DIR, 'data_prepared')
16
17 FEATURE_DIR = os.path.join(BASE_DIR, 'features')
18 ARFF_DIR = os.path.join(BASE_DIR, 'arff')
19
20 # ARFF subfolders
21 ARFF_ANGLES_DIR = os.path.join(ARFF_DIR, 'angles')
22 ARFF_DISTS_DIR = os.path.join(ARFF_DIR, 'dists')
23 ARFF_PROP_ANGLE_DIR = os.path.join(ARFF_DIR, 'prop_angle')
24 ARFF_PROP_DIST_DIR = os.path.join(ARFF_DIR, 'prop_dist')
25 ARFF_SETS_DIR = os.path.join(ARFF_DIR, 'sets')
26
27 # CSV subfolders (new)
28 CSV_SETS_DIR = os.path.join(FEATURE_DIR, 'sets_csv')
29 CSV_ANGLES_DIR = os.path.join(FEATURE_DIR, 'angles_csv')
30 CSV_DISTS_DIR = os.path.join(FEATURE_DIR, 'dists_csv')
31 CSV_PROPS_DIR = os.path.join(FEATURE_DIR, 'props_csv')
32 CSV_PROP_ANGLE_DIR = os.path.join(FEATURE_DIR, 'prop_angle_csv')
33 CSV_PROP_DIST_DIR = os.path.join(FEATURE_DIR, 'prop_dist_csv')
34
35 for d in [
36     ARFF_ANGLES_DIR, ARFF_DISTS_DIR, ARFF_PROP_ANGLE_DIR,
37     CSV_SETS_DIR, CSV_ANGLES_DIR, CSV_DISTS_DIR, CSV_PROPS_DIR, CSV_PROP_A
38 ]:
39     os.makedirs(d, exist_ok=True)
40
41 # ====== PARAMETERS ======
42 IMAGE_SIZE = (256, 256)
43
44 GLOM_ANGLES = [0, np.pi/4, np.pi/2, 3*np.pi/4] # PDF required
45 ANGLE_NAMES = ['0°', '45°', '90°', '135°']
46
47 GLOM_DISTS = [1, 2, 3, 4] # richer
48 GLOM_BASE_PROPS = ['contrast', 'dissimilarity', 'homogeneity', 'energy', 'corr']
49 GLOM_EXTRA_KEYS = ['entropy', 'max_probability', 'mean_i', 'mean_j', 'var_i', 'var_j']
50
51 LBP_RADIUS = 2
52 LBP_POINTS = 8 * LBP_RADIUS
53
54 # Gabor bank
55 GABOR_THETAS = [0, np.pi/4, np.pi/2, 3*np.pi/4]
56 GABOR_FREQS = [0.1, 0.2]

```

Name	Type	Size	Value
DATASET_DIR	str	68	/Users/zahra.es/Documents/image processing/MalingFinal/data_prepared
FEATURE_DIR	str	63	/Users/zahra.es/Documents/image processing/MalingFinal/features
GABOR_FREQS	list	2	[0.1, 0.2]
GABOR_THETAS	list	4	[0, 0.7853981633974483, 1.5707963267948966, 2.356194490192345]
GLCM_ANGLES	list	4	[0, 0.7853981633974483, 1.5707963267948966, 2.356194490192345]
GLCM_BASE_PROPS	list	6	['contrast', 'dissimilarity', 'homogeneity', 'energy', 'correlation', 'corr']
GLCM_DISTS	list	4	[1, 2, 3, 4]
GLCM_EXTRA_KEYS	list	6	['entropy', 'max_probability', 'mean_i', 'mean_j', 'var_i', 'var_j']
IMAGE_SIZE	tuple	2	(256, 256)
LBP_POINTS	int	1	16
LBP_RADIUS	int	1	2

Console output (partial):

```

Extracting Adobelot: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Alapie: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Amontize: 100% [██████████] 350/350 [03:56<00:00, 1.481t/s]
Extracting Androm: 100% [██████████] 350/350 [03:56<00:00, 1.491t/s]
Extracting Androsoft: 100% [██████████] 350/350 [03:56<00:00, 1.491t/s]
Extracting BrowsFox: 100% [██████████] 350/350 [04:02<00:00, 1.441t/s]
Extracting Dimwod: 100% [██████████] 350/350 [03:57<00:00, 1.471t/s]
Extracting Emsisoft: 100% [██████████] 350/350 [03:57<00:00, 1.471t/s]
Extracting Exploit: 100% [██████████] 350/350 [03:55<00:00, 1.481t/s]
Extracting Fasong: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Fazit: 100% [██████████] 350/350 [03:55<00:00, 1.481t/s]
Extracting Haux: 100% [██████████] 350/350 [03:56<00:00, 1.481t/s]
Extracting Injector: 100% [██████████] 350/350 [03:56<00:00, 1.481t/s]
Extracting InstallCore: 100% [██████████] 350/350 [03:55<00:00, 1.481t/s]
Extracting Interceptor: 100% [██████████] 350/350 [03:54<00:00, 1.481t/s]
Extracting Neorealm: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Neshta: 100% [██████████] 350/350 [03:57<00:00, 1.481t/s]
Extracting Nightraven: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Regent: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Sality: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Sparware: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]
Extracting Stuxnet: 100% [██████████] 350/350 [03:55<00:00, 1.491t/s]
Extracting VBA: 100% [██████████] 350/350 [03:55<00:00, 1.491t/s]
Extracting VBKrypt: 100% [██████████] 350/350 [03:55<00:00, 1.491t/s]
Extracting Visel: 100% [██████████] 350/350 [03:54<00:00, 1.491t/s]

```

IPython Console: Update available, Inline, Conda: anaconda3 (Python 3.10.14), LSP: Python, Line 232, Col 34, UTF-8, LF, RW, Mem 77%

3. ÖN İŞLEME (PREPROCESSING)

- Tüm görüntüler grayscale olarak işlendi
- Görüntüler 256×256 boyutuna yeniden ölçeklendirildi
- Bozuk(okunamayan) dosyalar filtreldi

4. ÖZELLİK ÇIKARIMI (FEATURE EXTRACTION)

KULLANILAN ÖZELLİK GRUPLARI:

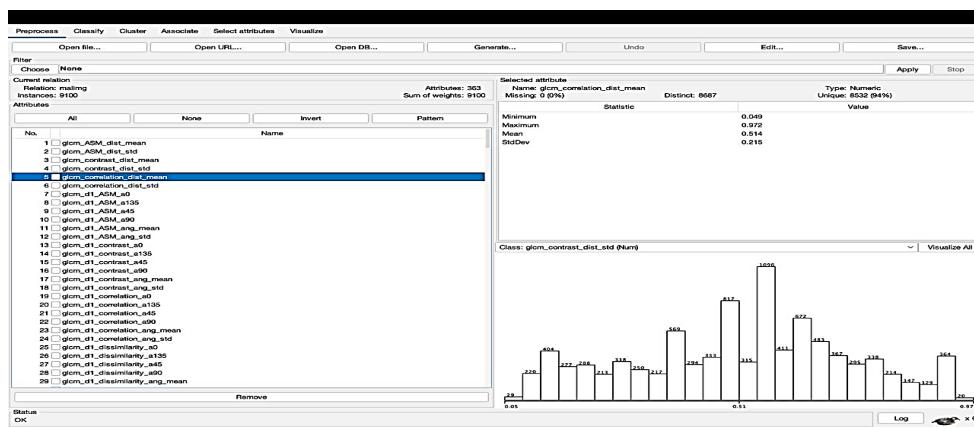
1. GLCM (Texture)

- 4 açı: 0° , 45° , 90° , 135°
- 4 mesafe: $d = 1, 2, 3, 4$
- Temel + türetilmiş özellikler

2. Wavelet (Haar DWT)

3. Gabor filtreleri
4. LBP (özet)
5. HOG (özet)
6. Global istatistikler.

Özellik Grubu	Açıklama	Yaklaşık Özellik Sayısı
GLCM	Doku (açı + mesafe)	~80
Wavelet	Frekans bilgisi	16
Gabor	Yönlü doku	16
LBP	Yerel doku (özet)	4
HOG	Kenar/yapı (özet)	4
Global	İstatistiksel	10
Toplam		~130+



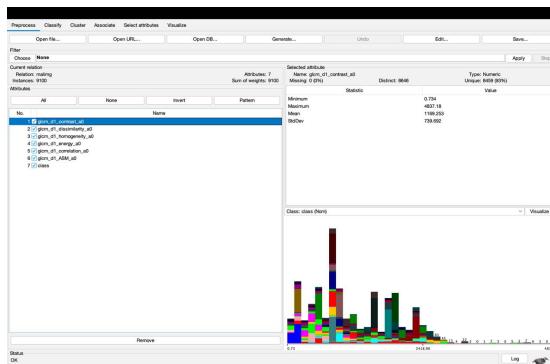
5.DENEYLER VE SONUÇLAR (RESULTS)

.GLCM

➤ GLCM AÇI KARŞILAŞTIRMASI

1.GLCM_D1.ARFF :ATTRIBUTES:CONTRAST-DISSIMILARITY-HOMOGENEITY-ENERGY-CORRELATION-ASM

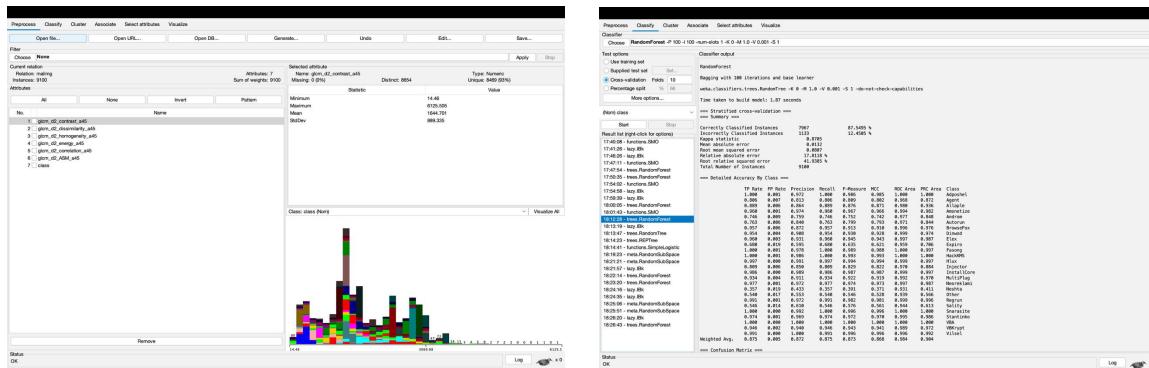
Açı	Mesafe	Classify	Accuracy (%)	Kappa
0°	d1	Randomforest:	87.7363%	0.8725
		Function-SMO:	47.6593%	0.4557
		Lazy-IBK	86.7802%	0.8625
45°	d1	Randomforest:	87.1868%	0.8667
		Function-SMO:	45.0879%	0.4289
		Lazy-IBK	85.7802%	0.8521
90°	d1	Randomforest:	87.044%	0.8653
		Function-SMO:	44.7802%	0.4257
		Lazy-IBK	85.9341%	0.8537
135	d1	Randomforest:	87.3846%	0.8688
		Function-SMO:	43.0549%	0.4078
		Lazy-IBK	85.7692%	0.852



2.GLCM_D2.ARFF :

Attributes:contrast-dissimilarity-homogeneity-energy-correlation-ASM

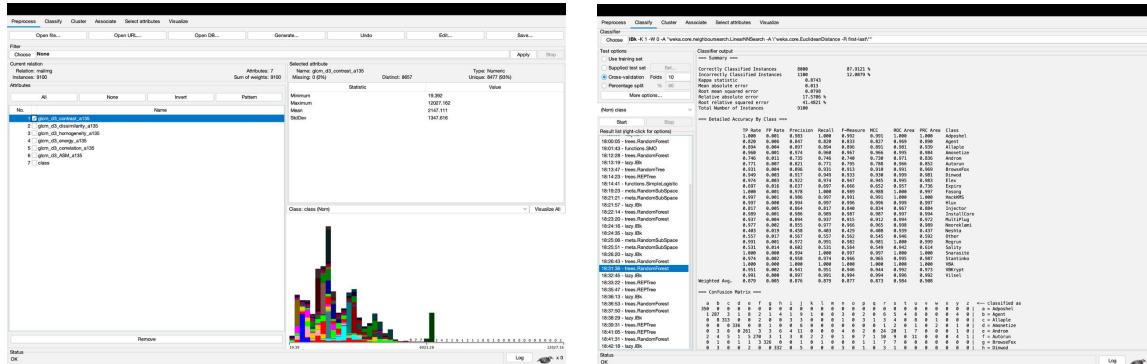
Açı	Mesafe	Classify	Accuracy (%)	Kappa
0°	d2	Randomforest:	87.5497%	0.8705
		Meta randomsubspace	83.9341%	0.8329
		Lazy-IBK	86.8022%	0.8627
45°	d2	Randomforest:	87.1867%	0.8667
		Meta randomsubspace	83.6374%	0.8298
		Lazy-IBK	85.7802%	0.8521
90°	d2	Randomforest:	86.9341%	0.8642
		Meta randomsubspace	86.011%	0.8233
		Lazy-IBK	85.3597%	0.8572
135°	d2	Randomforest:	87.4736%	0.8688
		Meta randomsubspace	83.6044%	0.8595
		Lazy-IBK	85.7692%	0.852



3.GLCM_D3.ARFF :

Attributes:contrast-dissimilarity-homogeneity-energy-correlation-ASM

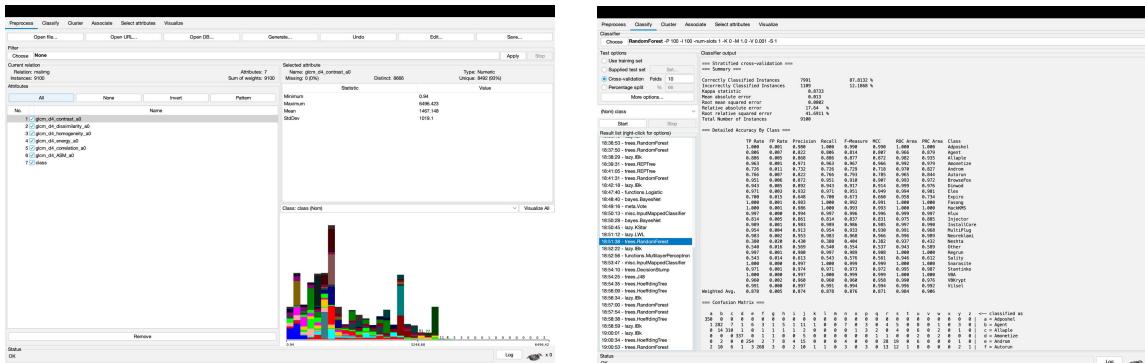
Açı	Mesafe	Classify	Accuracy (%)	Kappa
0°	d3	Randomforest:	87.9121%	0.8743
		Trees.REPTree	82.1538%	0.8144
		Lazy-IBK	86.7253%	0.8615
45°	d3	Randomforest:	86.5824%	0.8625
		Trees.REPTree	81.022%	0.8026
		Lazy-IBK	85.0879%	0.8449
90°	d3	Randomforest:	86.8462%	0.8632
		Trees.REPTree	81.3516%	0.8031
		Lazy-IBK	84.8681%	0.8426
135°	d3	Randomforest:	86.4945	0.8595
		Trees.REPTree	81.3077%	0.8056
		Lazy-IBK	85.15385	0.8456



4.GLCM_d4.arff :

Attributes:contrast-dissimilarity-homogeneity-energy-correlation-ASM

Açı	Mesafe	Classify	Accuracy (%)	Kappa
0°	d4	Randomforest:	87.8132%	0.8733
		Trees.HoeffdingTree	46.1648%	0.4401
		Lazy-IBK	86.3956%	0.8585
45°	d4	Randomforest:	86.9451%	0.8642
		Trees.HoeffdingTree	50.0659%	0.4807
		Lazy-IBK	84.9341%	0.8433
90°	d4	Randomforest:	86.5604%	0.8602
		Trees.HoeffdingTree	47.7143%	0.4562
		Lazy-IBK	84.9121%	0.8431
135°	d4	Randomforest:	86.7033%	0.8617
		Trees.HoeffdingTree	50.7912%	0.4882
		Lazy-IBK	84.956%	0.8435



➤ GLCM MESAFE KARŞILAŞTIRMASI

ATTRIBUTES: CONTRAST-DISSIMILARITY-HOMOGENEITY-ENERGY-CORRELATION-ASM(MEAN-STD)

Classify: Randomforest

Mesafe (d)	Accuracy (%)	Kappa
1	89.8901%	0.8949
2	89.8571%	0.8945
3	89.6484%	0.8923
4	89.2967%	0.8887

Confusion Matrix		Confusion Matrix	
		a	b
0	c	d	e
1	d	e	f
2	e	f	g
3	f	g	h
4	g	h	i
5	h	i	j
6	i	j	k
7	j	k	l
8	k	l	m
9	l	m	n
10	m	n	o
11	n	o	p
12	o	p	q
13	p	q	r
14	q	r	s
15	r	s	t
16	s	t	u
17	t	u	v
18	u	v	w
19	v	w	x
20	w	x	y
21	x	y	z
22	y	z	← classified as
23	z	← classified as	a = Adrophit
24	0	1	b = Agent
25	1	0	c = Allegro
26	0	0	d = Androm
27	0	1	e = Auturon
28	1	0	f = Automax
29	0	0	g = BrodusFox
30	0	1	h = Dimod
31	1	0	i = Elex
32	0	0	j = Exstro
33	0	1	k = Fasong
34	1	0	l = HackMS
35	0	0	m = Injetor
36	0	1	n = Injector
37	1	0	o = MultiPcore
38	0	0	p = Neorekami
39	0	1	q = Neostream
40	1	0	r = Other
41	0	0	s = Regim
42	0	1	t = Recon
43	1	0	u = Sality
44	0	0	v = Srasilite
45	0	1	w = Statinice
46	1	0	x = VBRopt
47	0	0	y = VKrypt
48	0	1	z = Vilset
49	1	0	← classified as
50	0	0	a = Adrophit
51	0	1	b = Agent
52	1	0	c = Allegro
53	0	0	d = Androm
54	0	1	e = Auturon
55	1	0	f = Automax
56	0	0	g = BrodusFox
57	0	1	h = Dimod
58	1	0	i = Elex
59	0	0	j = Exstro
60	0	1	k = Fasong
61	1	0	l = HackMS
62	0	0	m = Injetor
63	0	1	n = Injector
64	1	0	o = MultiPcore
65	0	0	p = Neorekami
66	0	1	q = Neostream
67	1	0	r = Other
68	0	0	s = Regim
69	0	1	t = Recon
70	1	0	u = Sality
71	0	0	v = Srasilite
72	0	1	w = Statinice
73	1	0	x = VBRopt
74	0	0	y = VKrypt
75	0	1	z = Vilset

== Confusion Matrix ==																							classified as				
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	classified as
350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	a = Adolphus		
0	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	b = Bungle		
0	2	317	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	c = Amnetize		
0	2	0	337	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	d = Amonetize		
0	2	0	2	275	2	2	3	1	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = Androm		
0	2	0	0	286	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	f = Androm		
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = BrownFox		
0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	h = Dimwood		
0	3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	i = Driller		
0	3	6	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	j = Expilo		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	k = Fasong		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	l = HackX99		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	m = InstalCore		
0	0	4	5	0	3	1	4	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n = Injector		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	o = Necroklam		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p = Neuron		
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q = Neshta		
1	2	5	3	34	12	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	r = Neutron		
0	0	4	0	16	22	9	2	1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	s = Other		
1	5	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	t = Neutron		
1	5	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	u = Salty		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	v = Snarstine		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	w = Stantinko		
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	x = Vrykrypt		
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	y = Vrykrypt		
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	z = Vilsel		
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	== classified as		
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	== classified as
350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	a = Adolphus	
3	295	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	b = Bungle	
0	10	316	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	c = Amnetize	
0	1	0	0	337	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	d = Amonetize		
0	5	5	0	268	1	1	2	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = Androm		
2	0	2	0	286	2	2	3	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	f = Androm		
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	g = BrownFox		
0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	h = Dimwood		
0	3	6	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	i = Driller		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	j = Expilo		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	k = Fasong		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	l = HackX99		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	m = HackX99		
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n = InstalCore		
0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	o = Installer		
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p = MultiPlus		
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q = Neutron		
0	1	3	5	2	26	11	5	2	3	29	1	0	2	10	0	0	4	149	42	1	36	0	3	1	2	r = Other	
0	0	8	0	8	10	3	2	19	0	1	0	3	1	5	2	41	214	25	0	2	0	0	0	0	s = Other		
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	t = Neutron		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	u = Salty		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	v = Snarstine		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	w = Stantinko		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	x = Vrykrypt		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	y = Vrykrypt		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	z = Vilsel		
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	== classified as		

➤ GLCM FARKLI ÖZELLİKLERİN MEAN-STD'LERİ MESAFEYE GÖRE

Classify:Randomforest

Özellik	İşlem	Accuracy (%)	Kappa
Contrast	Mean:	60.3846%	0.588
	Std:	56.4066%	0.5466
Energy	Mean:	64.6374%	0.6322
	Std:	58.989%	0.5735
Entropy	Mean:	59%	0.5736
	Std:	52.4725%	0.5057
Correlation	Mean:	58.6593%	0.5701
	Std:	50.3956%	0.4841
ASM	Mean:	64.2857%	0.6986
	Std:	60.6703%	0.591
homogeneity	Mean:	60.6813%	0.5911
	Std:	47.4615%	0.4536
Max-probability	Mean:	65.9451%	0.6458
	Std:	61.7473	0.6022
Dissimilarity	Mean:	60.2747%	0.5869
	Std:	54.6124	0.528

➤ GLCM ALL

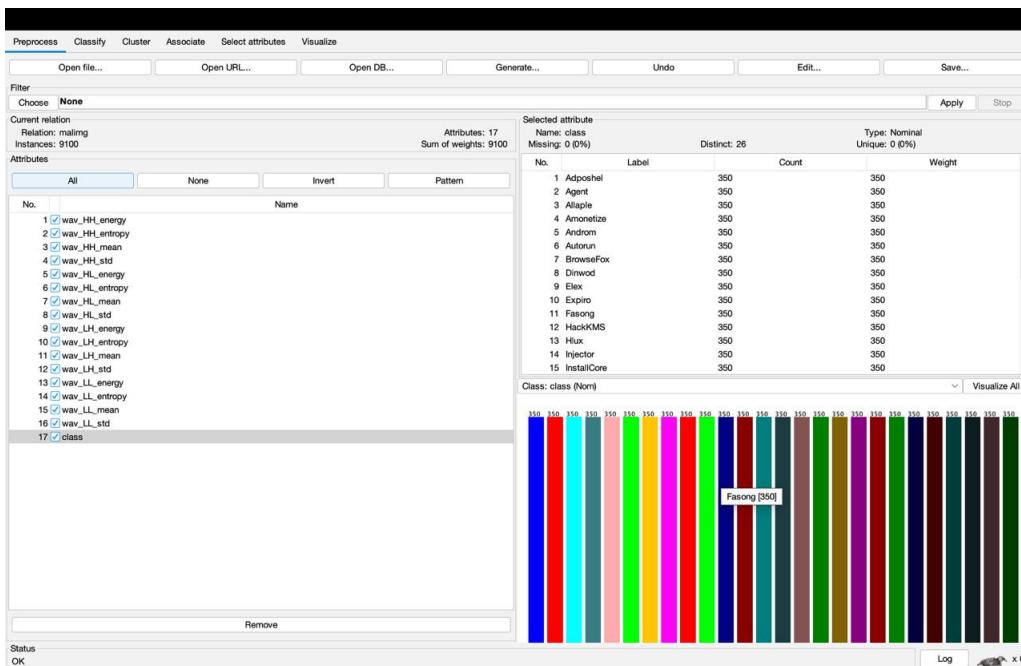
Classify	Accuracy (%)	Kappa
Random-forest	92.6374%	0.9234
Function-SMO	84.8703%	0.8406
Lazy-IBK	92.2637%	0.9195
Tree-REPTree	85.7692%	0.852

.Wavelet(DWT)

➤ WAVELET ONLY

Özellik	Band	Accuracy (%)	Kappa
Energy	LL-LH-HL-HH	91.0 769%	0.9072
Entropy	LL-LH-HL-HH	91.0769%	0.9072

NOT: Burda önce bantları ayrı ayrı çalıştırarak aynı sonucu elde ettiğimiz için genel bir tablo da topladık.



.Özellik birleşimi

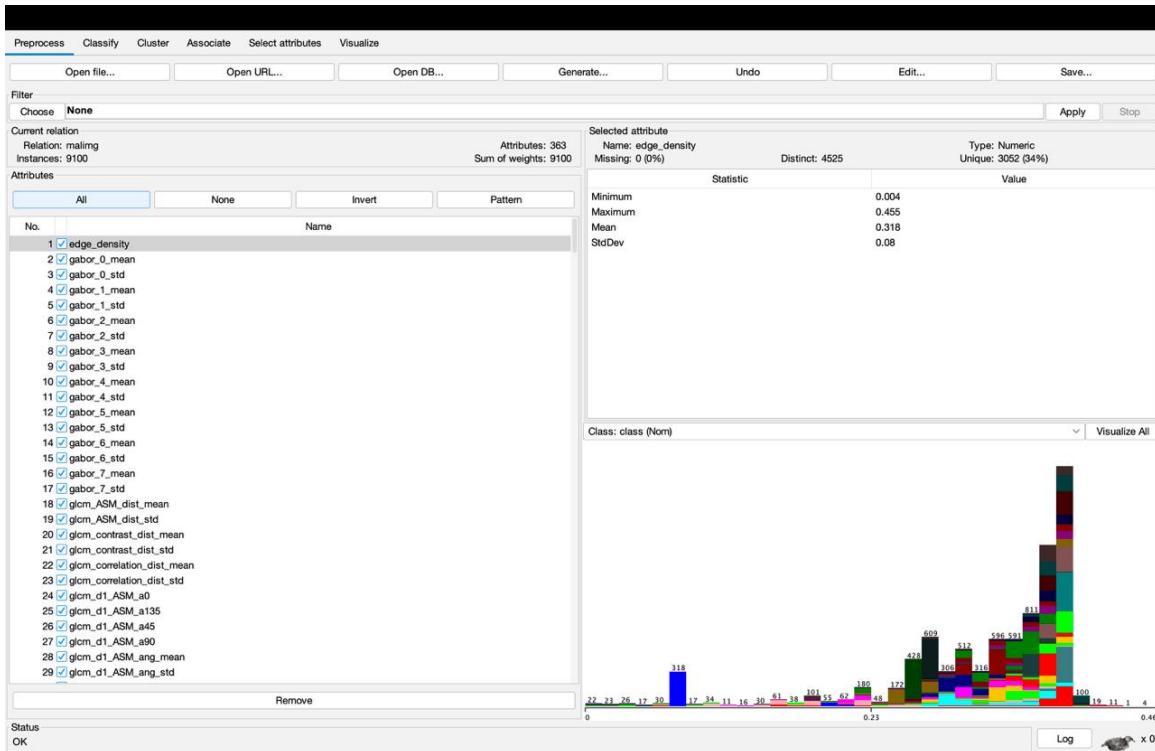
Classify: Randomforest

Feature Set	Özellik Sayısı	Accuracy (%)	Kappa
GLCM only	313	92.6373%	0.9234
GLCM + Wavelet	313+16=329	92.6044%	0.9231
GLCM + Gabor	313+16=329	92.7912	0.925
Wavelet only	17	91.0769%	0.9072
Gabor only	17	89.6703%	0.8926
GLCM-LBP-HOG	321	92.7253%	0.9243
GLCM-Wavelet Gabor-LBP-HOG Global	363	92.6154%	0.9232
Global only	11	91.33033%	0.9067
HOG only	5	82.8462%	0.8216
LBP only	5	74.2418%	0.7321

FULL:GLCM-WAVELET GABOR-LBP-HOG GLOBAL FARKLI CLASSIFFYLER İLE

Özellik sayısı:363

Classify	Accuracy (%)	Kappa
Randomforest	92.6154%	0.9232
RepTree	85.5934%	0.8502
IBK	92.5275%	0.9223
SMO	86.7473%	0.8622
Naivebayes	53.967	0.5641



6.*BONUS: DOMAIN DÖNÜŞÜM YÖNTEMİ

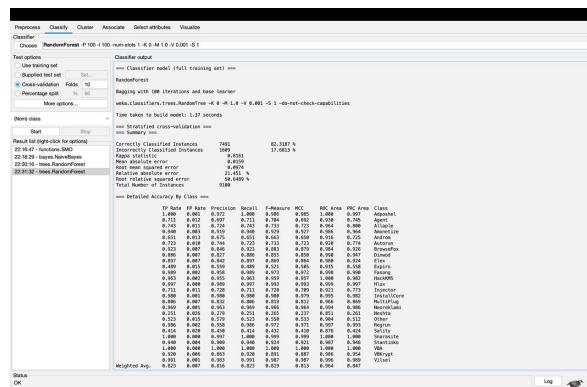
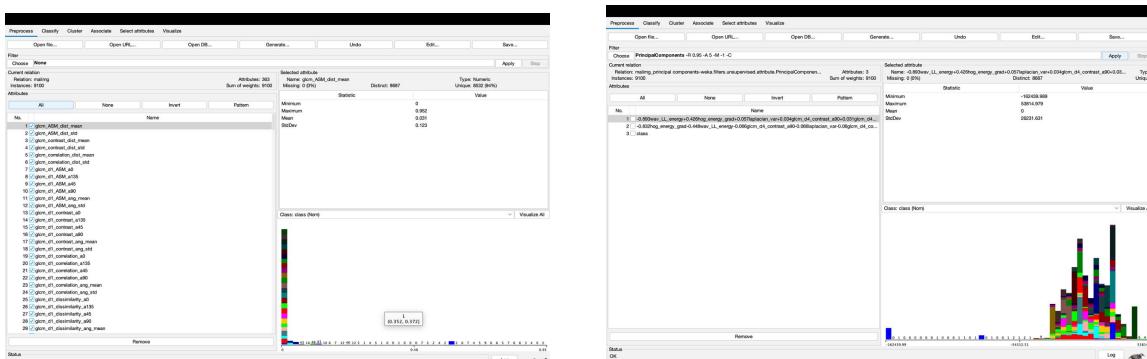
Transfer yöntemi	Classify	Accuracy (%)	Kappa
Wavelet	Randomforest	91.0769%	0.9072
	Bayes.Naivebayes	60.7143%	0.5914
	Function-SMO	66.48355	0.6514

BAYES.NAIVEBAYES İLE OLUŞTURDUGUMUZ TÜM SINIF VE ÖZELLİKLER TEK TABLODA :

7. PCA / FEATURE SELECTION

Boyut indirgeme amacıyla PCA yöntemi uygulanmış ve toplam özellik sayısı XXX'ten YY'ye düşürülmüştür. PCA sonrası elde edilen doğruluk ve Kappa değerlerinin PCA öncesine yakınlığı gözlemlenmiş, böylece daha az sayıda özellik ile benzer sınıflandırma başarımı elde edilebileceği gösterilmiştir.

Yöntem	Özellik	Accuracy (%)	Kappa
PCA öncesi	363	92.6154%	0.9232
PCA sonrası	3	82.3187%	0.175713



8.TARTIŞMA VE ANALİZ

Bu çalışmada MaleVis veri seti kullanılarak, zararlı yazılım ailelerinin görüntü tabanlı doku özellikleri ile sınıflandırılması kapsamlı bir şekilde incelenmiştir. Elde edilen deneyel sonuçlar, özellikle GLCM tabanlı özelliklerin malware sınıflandırmasında yüksek ayırt ediciliğe sahip olduğunu açıkça göstermektedir. GLCM'nin farklı açı ve mesafe parametreleri için yapılan deneyler, sınıflandırma başarımının bu parametrelere bağlı olarak değiştiğini, ancak genel olarak tüm açılarda ve mesafelerde istikrarlı ve yüksek performans elde edildiğini ortaya koymuştur.

GLCM açı karşılaştırması sonuçları incelendiğinde, 0° , 45° , 90° ve 135° yönleri arasında Random Forest sınıflandırıcısı için doğruluk ve Kappa değerlerinin birbirine oldukça yakın olduğu görülmüştür. Bu durum, zararlı yazılım görüntülerinin dokusal yapısının yön bağımsız olarak güçlü ayırt edici özellikler barındırdığını göstermektedir. Özellikle $d = 1$ mesafesinde elde edilen yaklaşık %87–88 doğruluk ve 0.86–0.87 aralığındaki Kappa değerleri, temel GLCM özelliklerinin tek başına dahi anlamlı sonuçlar üretebildiğini kanıtlamaktadır.

Mesafe bazlı analizlerde, GLCM özelliklerinin mean ve standart sapma (std) ile özetlenmesi sonucunda sınıflandırma başarımının daha da arttığı gözlemlenmiştir. $d = 1, 2, 3$ ve 4 mesafeleri için doğruluk oranlarının yaklaşık %89–90 seviyelerine ulaştığı ve Kappa değerlerinin 0.88–0.89 aralığında seyrettiği görülmüştür. Bu sonuç, farklı mesafelerden elde edilen dokusal bilgilerin birleştirilmesinin (mean–std temsili) daha zengin ve güçlü bir özellik kümesi oluşturduğunu göstermektedir.

Tek özellik analizleri incelendiğinde, contrast, energy, entropy, ASM ve max-probability gibi bazı GLCM özelliklerinin diğerlerine kıyasla daha yüksek performans sunduğu belirlenmiştir. Bununla birlikte, hiçbir tekil özelliğin, çoklu özellik birleşimleri kadar yüksek doğruluk sağlayamadığı görülmüştür. Bu durum, malware görüntülerinin karmaşık yapısı nedeniyle tek bir doku ölçütünün tüm sınıfları ayırt etmeye yetersiz kaldığını, buna karşın özellik birleşimlerinin sınıflandırma başarımını belirgin şekilde artırdığını göstermektedir.

Özellik birleşimi (feature fusion) deneyleri, çalışmanın en güçlü sonuçlarını ortaya koymuştur. GLCM özelliklerine Wavelet, Gabor, LBP, HOG ve global istatistik özelliklerinin eklenmesiyle oluşturulan tam birleşik (full fusion) özellik kümesi, yaklaşık %92.6 doğruluk ve 0.92'nin üzerinde Kappa değeri ile en yüksek performansı sağlamıştır. Bu sonuç, farklı özellik gruplarının birbirini tamamlayıcı nitelikte olduğunu ve birlikte kullanıldıklarında sınıflandırma başarımını artırdığını göstermektedir.

Farklı sınıflandırıcılar karşılaştırıldığında, Random Forest ve IBk algoritmalarının en yüksek performansı sunduğu, SMO ve özellikle Naive Bayes'in ise daha düşük doğruluk değerleri ürettiği gözlemlenmiştir. Bu durum, yüksek boyutlu ve karmaşık özellik uzaylarında ağaç tabanlı ve örnek tabanlı yöntemlerin daha başarılı olduğunu göstermektedir. Naive Bayes'in görece düşük performansı ise özellikler arasındaki bağımsızlık varsayıminın bu problem için uygun olmamasıyla açıklanabilir.

Son olarak PCA ile gerçekleştirilen boyut indirgeme deneyleri, toplam özellik sayısının önemli ölçüde azaltılmasına rağmen sınıflandırma başarımının büyük oranda korunduğunu göstermiştir. Bu bulgu, MaleVis veri setinde kullanılan özelliklerin önemli bir kısmının bilgi açısından redundant olduğunu ve daha az sayıda özellik ile de benzer performans elde edilebileceğini ortaya koymaktadır.

Genel olarak değerlendirildiğinde, bu çalışma görüntü tabanlı doku özelliklerinin malware sınıflandırmasında etkili ve güvenilir bir yaklaşım sunduğunu, özellikle GLCM tabanlı özelliklerin ve özellik birleşimlerinin yüksek sınıflandırma başarımı sağladığını göstermektedir.

9.SONUÇ

- Bu projede, MaleVis veri seti kullanılarak zararlı yazılım ailelerinin görüntü tabanlı doku özellikleri ile sınıflandırılması gerçekleştirilmiştir.
- Zararlı yazılım ikili dosyalarının gri seviye görüntülere dönüştürülmesi sayesinde, farklı malware ailelerinin ayırt edici dokusal desenler oluşturduğu gözlemlenmiştir.
- GLCM tabanlı özellikler, açı ve mesafe parametrelerinden bağımsız olarak yüksek sınıflandırma başarımı sağlamış ve malware sınıflandırmasında en etkili özellik grubu olarak öne çıkmıştır.
- Farklı özellik gruplarının birlikte kullanıldığı feature fusion yaklaşımı, tekil özelliklere kıyasla daha yüksek doğruluk ve Kappa değerleri üretmiştir.
- Random Forest ve IBk sınıflandırıcıları, diğer yöntemlere göre daha başarılı sonuçlar vermiştir.
- PCA ile boyut indirgeme sonrasında, daha az sayıda özellik kullanılarak benzer sınıflandırma performansı elde edilebildiği gösterilmiştir.
- Elde edilen sonuçlar, görüntü işleme tabanlı yaklaşımların zararlı yazılım sınıflandırmasında etkili ve uygulanabilir olduğunu ortaya koymaktadır.