

IVI 시스템 포렌식을 위한 파일 파편 유형 분류

이승민, 노경민, 정지현, 박수현, 조성제

WDSC 2025

2025.08.19

INDEX

01

서론

02

데이터셋 및 파편화

03

파일 파편 분류 모델

04

실험 및 평가

05

결론 및 향후 연구

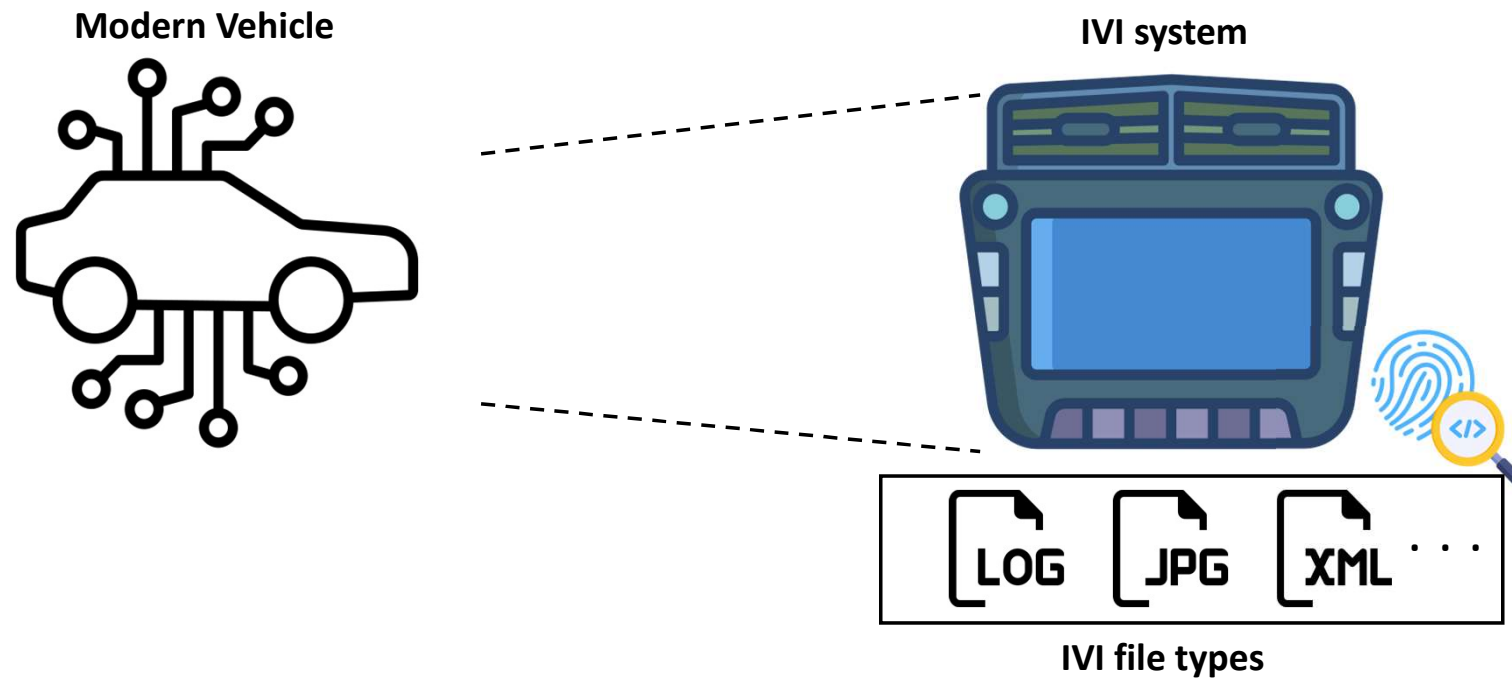


01

서론

Introduction

❖ IVI(In-Vehicle Infotainment) Forensic



관련 연구

- [8] K. Skračić et al.(2023) , “**Classification of low-and high-entropy file fragments using randomness measures and discrete Fourier transform coefficients**”, Vietnam journal of computer science

File Types	18 types
Model	RF, SVM, DNN
Score	F-score: 86.13%

- [11] K. Vulinović et al.(2019), “**Neural Networks for File Fragment Classification**”, IEEE MIPRO

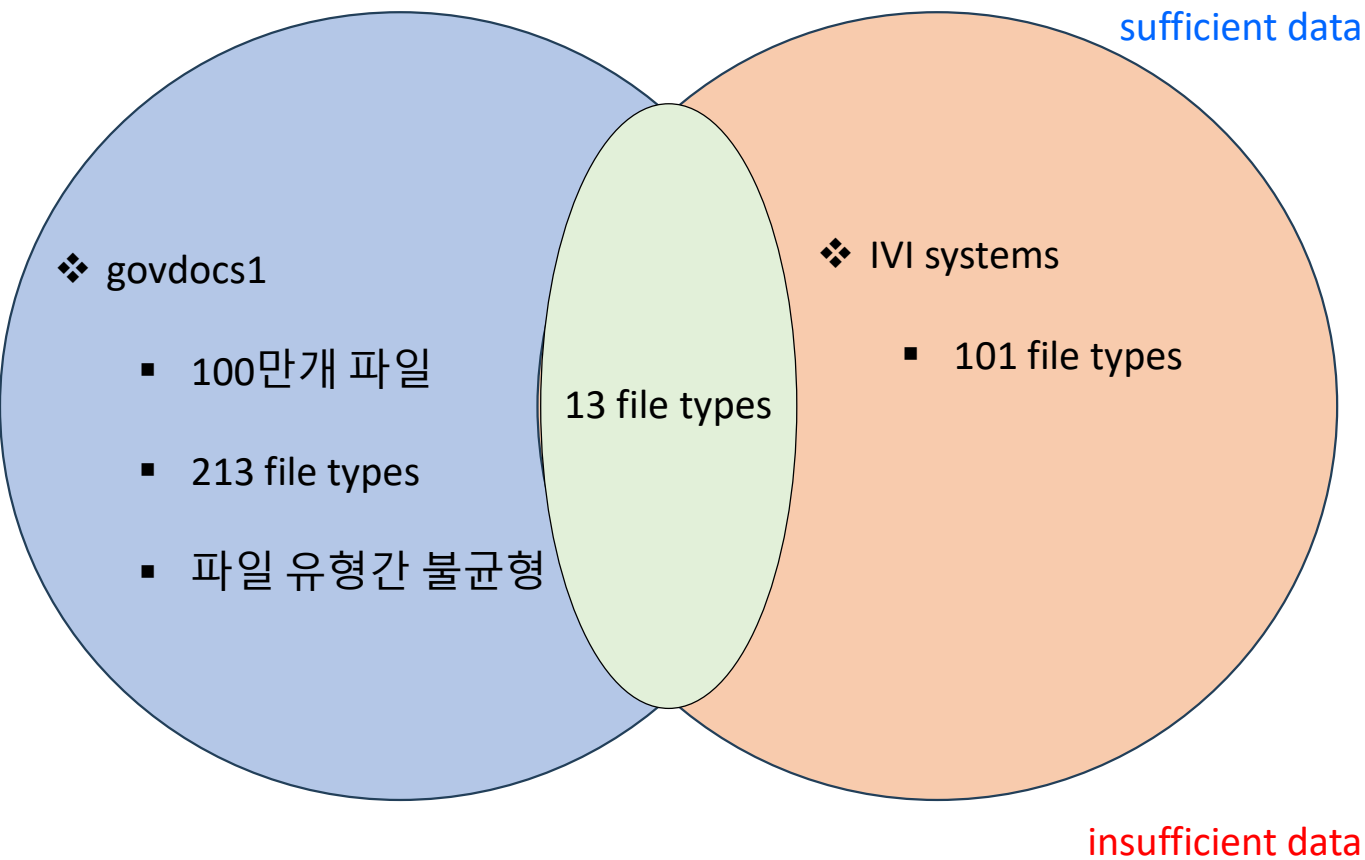
File Types	20 types
Model	CNN, FFANN
Score	F-score: 87.72%



02

데이터셋 및 파편화

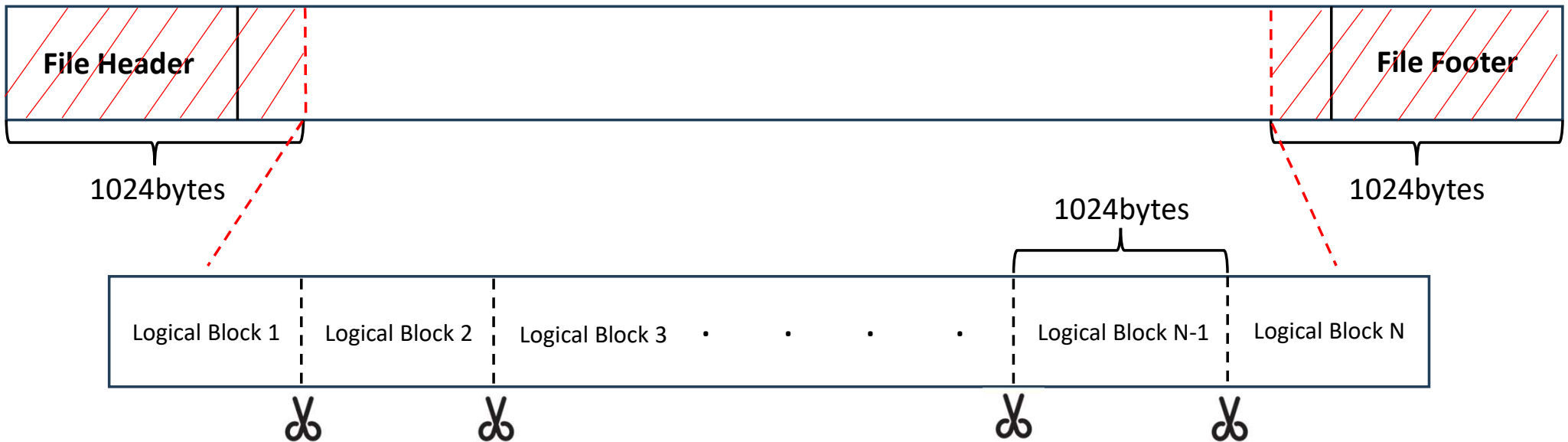
Data Collection



Type	Count
.html	207,857
.jpg	109,164
.txt	74,022
.gif	36,132
.xml	32,929
.gz	13,703
.log	9,643
.swf	3,458
.java	288
.tmp	137
.zip	10
.ttf	7
.bin	1

Fragmentation

ex) TEST.jpg



03

파일 파편 분류 모델

Feature Extraction

❖ Entropy

- 파일 파편의 정보량
- 암호화/압축을 측정하는 지표
- [14] A. Mantovani et al.(2020), “Prevalence and impact of low-entropy packing schemes in the malware ecosystem”, NDSS
if Entropy value ≥ 7.0
 -> **High Entropy**
else -> **Low Entropy**

❖ N-gram

- 특정 바이트 시퀀스(sequence)가 자주 나타나는지 확인
- ex) .txt -> “0D 0A”(줄 바꿈 문자)

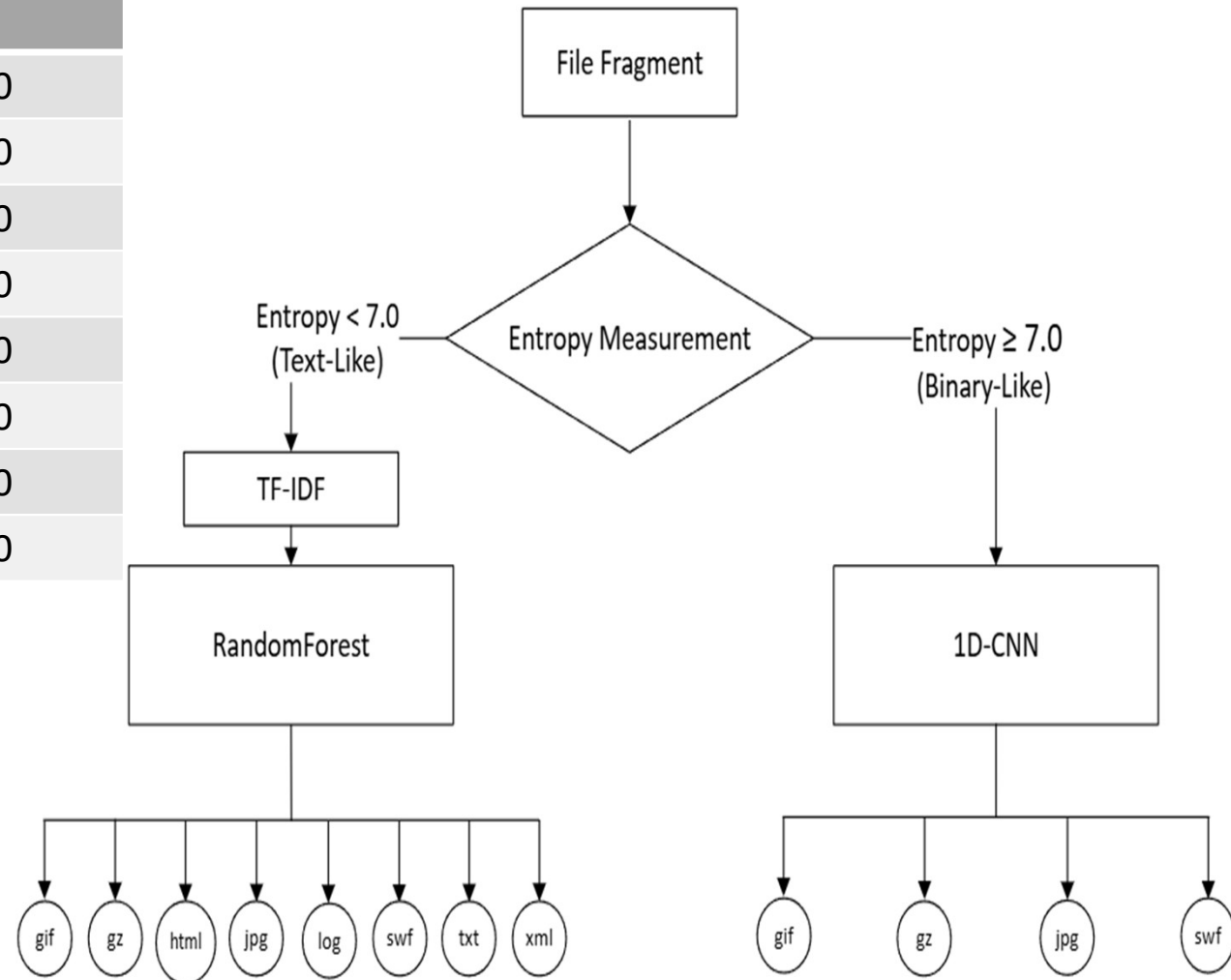


❖ TF-IDF

- 단어 중요도 측정 기법
- 파일 파편 -> 문서
N-gram(3~5bytes) -> 단어

Flowchart

Type	Train	Test
gif	50,000	30,000
gz	50,000	30,000
html	50,000	30,000
jpg	50,000	30,000
log	50,000	30,000
swf	50,000	30,000
txt	50,000	30,000
xml	50,000	30,000



Classification Models

Low Entropy Classification Model

RF(Random Forest)

❖ Features: TF-IDF

Offset(h)	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	Decoded text
00000000	38	2E	30	33	3A	33	31	3A	30	33	2E	38	39	23	61	6E	8.03:31:03.89#an
00000010	74	63	6E	23	41	43	55	3A	20	45	6C	3A	20	6E	6F	20	tcn#ACU: El: no
00000020	66	61	75	6C	74	20	20	20	20	20	20	20	20	20	20	20	fault
00000030	20	20	20	20	20	2C	20	6E	6F	20	66	61	75	6C	74	0A	, no fault.
00000040	32	30	30	33	2E	30	39	38	2E	30	33	3A	33	31	3A	30	2003.098.03:31:0
00000050	33	2E	39	30	2F	61	6E	74	65	6E	6E	61	2F	41	43	4B	3.90/antenna/ACK

High Entropy Classification Model

1D-CNN(Convolutional Neural Network)

❖ Features: Automatic

Offset(h)	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	Decoded text
00000000	D1	03	2A	17	A6	EE	EE	B2	2D	CF	20	F7	D1	DD	61	76	Ñ.*.¡i!-İ ÷ÑYav
00000010	45	D5	C1	CF	C7	2F	DC	AE	F7	F7	98	2D	A0	63	E1	4F	EÖÁİÇ/Ü÷÷~ - cảo
00000020	35	AA	B2	7A	02	9A	F6	BA	86	52	1E	57	C7	61	D5	03	5^z.šš°+R.WÇaÖ.
00000030	BC	EA	6D	C7	19	6F	AE	BB	07	B7	9C	2C	56	06	43	D8	4êmÇ.©»..œ,V.CØ
00000040	99	2B	38	20	8D	41	5A	BB	56	C2	86	D4	60	7C	B7	CD	™+8 .AZ»VÂ+Ô` ·Í
00000050	BC	9A	CD	47	B0	0D	EA	F4	EC	3D	AD	37	F0	4C	69	36	4šÍG°.êôì=.7ðLi6

04

실험 및 평가

Entropy Classification Result

Type	High Entropy		Low Entropy	
	Train	Test	Train	Test
gif	46,255	29,727	3,745	273
gz	49,639	29,850	361	150
html	28	0	49,972	30,000
jpg	49,265	29,401	735	599
log	0	0	50,000	50,000
swf	44,162	26,471	5,838	3,529
txt	0	0	50,000	30,000
xml	0	0	50,000	30,000

❖ High Entropy: gz, jpg, gif, swf

❖ Low Entropy: log, txt, xml, html, swf, gif, jpg, gz

Low Entropy Classification Result

Accuracy	gif	gz	html	jpg	log	swf	txt	xml
gif	25.7	47.3	0	0.3	0	7.9	18.8	0.1
gz	5.6	82.7	0	1.1	0	10.5	0	0
html	0.2	0.2	91.6	0	0.4	0	7.6	0
jpg	3.5	63.7	0	15.4	0.2	14.1	1.6	1.4
log	0	0	0.3	0	94.4	0	5.3	0
swf	10.4	8.2	0.4	1.7	0.1	78.7	0.5	0
txt	0	0	1.3	0	2.4	0	96.2	0
xml	0	0	1.9	0	0.2	0	2.4	95.5

❖ Low Entropy Result

- txt, log, gz, html, xml은 높은 성능
- gif, jpg, swf은 낮은 성능
- gif의 경우 gz으로 오분류(47.3%), jpg는 gz으로 오분류(63.7%)



- ✓ gif, jpg, gz, swf는 대부분 High Entropy
- ✓ 충분한 학습 파편이 주어지지 않음

High Entropy Classification Result

	gif	gz	jpg	swf
gif	88.8	7.3	0.4	3.5
gz	3.1	84.2	1	11.7
jpg	1	4.9	80.9	13.1
swf	3.1	44.8	21	31.1

❖ High Entropy Result

- gif, jpg 높은 성능
- swf은 대부분의 파편이 High Entropy 분류되었지만 31.1%, gz으로 오분류(44.8%)



05

결론 및 향후 연구

결론 및 향후 연구

Conclusion

- 파일 파편 분류를 위해 Entropy를 기준으로 Hybrid Model을 구성, **F1-score: 81.5%**
- 7개 클래스(txt, xml, log, html, gz, gif, jpg)에 대해 우수한 정확도
- Entropy 분류로 인한 모델 별 파편 수 부족, 적합한 특징 정보 필요

Future Work

- Entropy 분류를 고려한 데이터 확장, SSD 환경을 고려한 데이터 셋 구축
- 제안 모델의 최적화 작업

Acknowledgement

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 -대학
ICT연구센터(ITRC)의 지원(IITP-2025-RS-2023-00259967, 50%)과
산업통상자원부(MOTIE)와 한국에너지기술평가(KETEP) 지원을 받아 수행한
연구임(No.RS-2021-KP002461, 50%).

Thank you

Q&A
