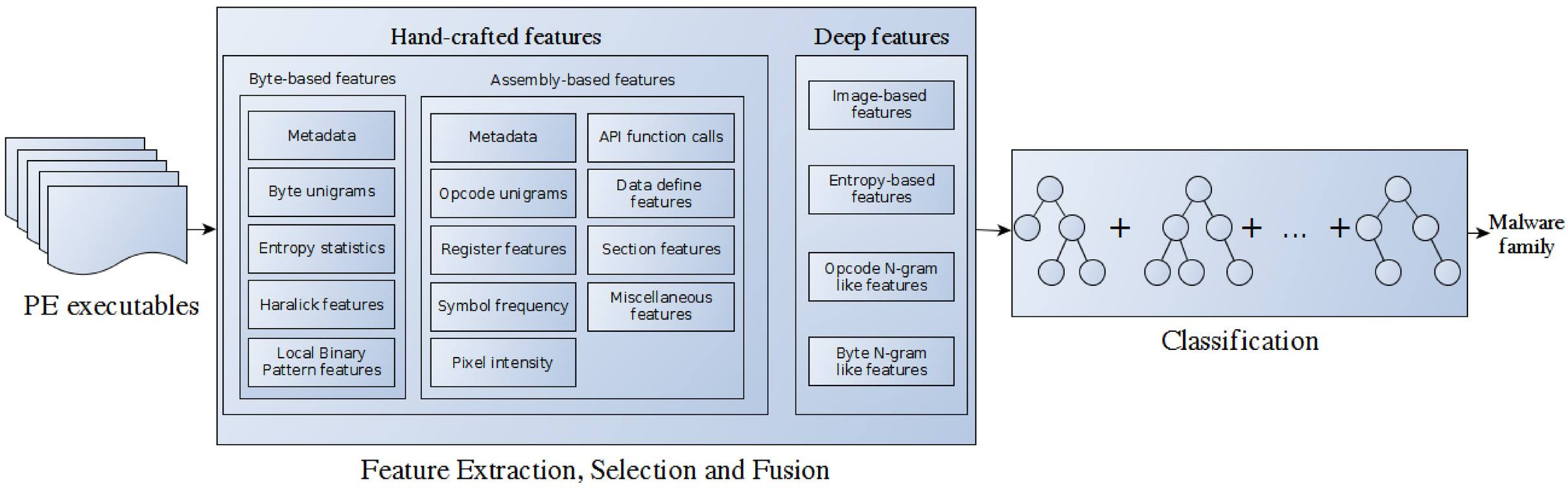
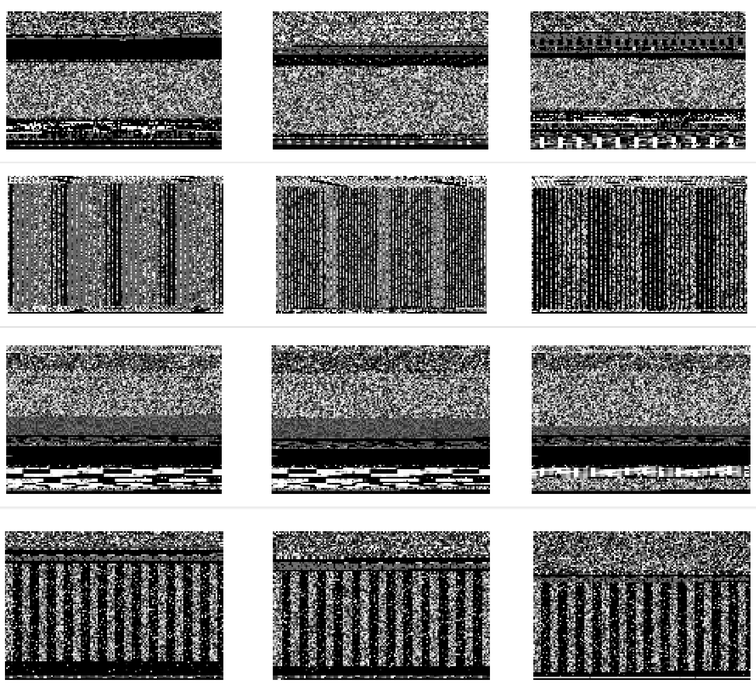
Fusing Feature Engineering and Deep Learning: A Case Study for Malware Classification

Dr Daniel Gibert, Dr Jordi Planes, Dr Carles Mateu, Dr Quan Le

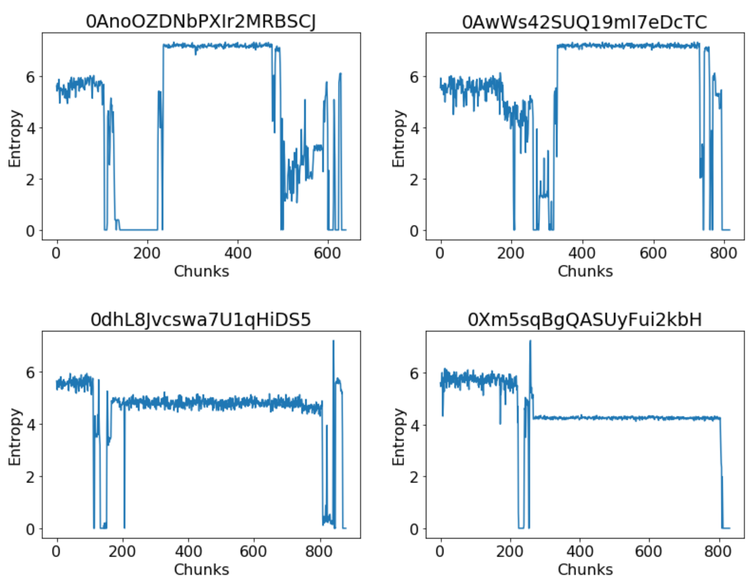
Malware Classification System Overview



Grayscale Image



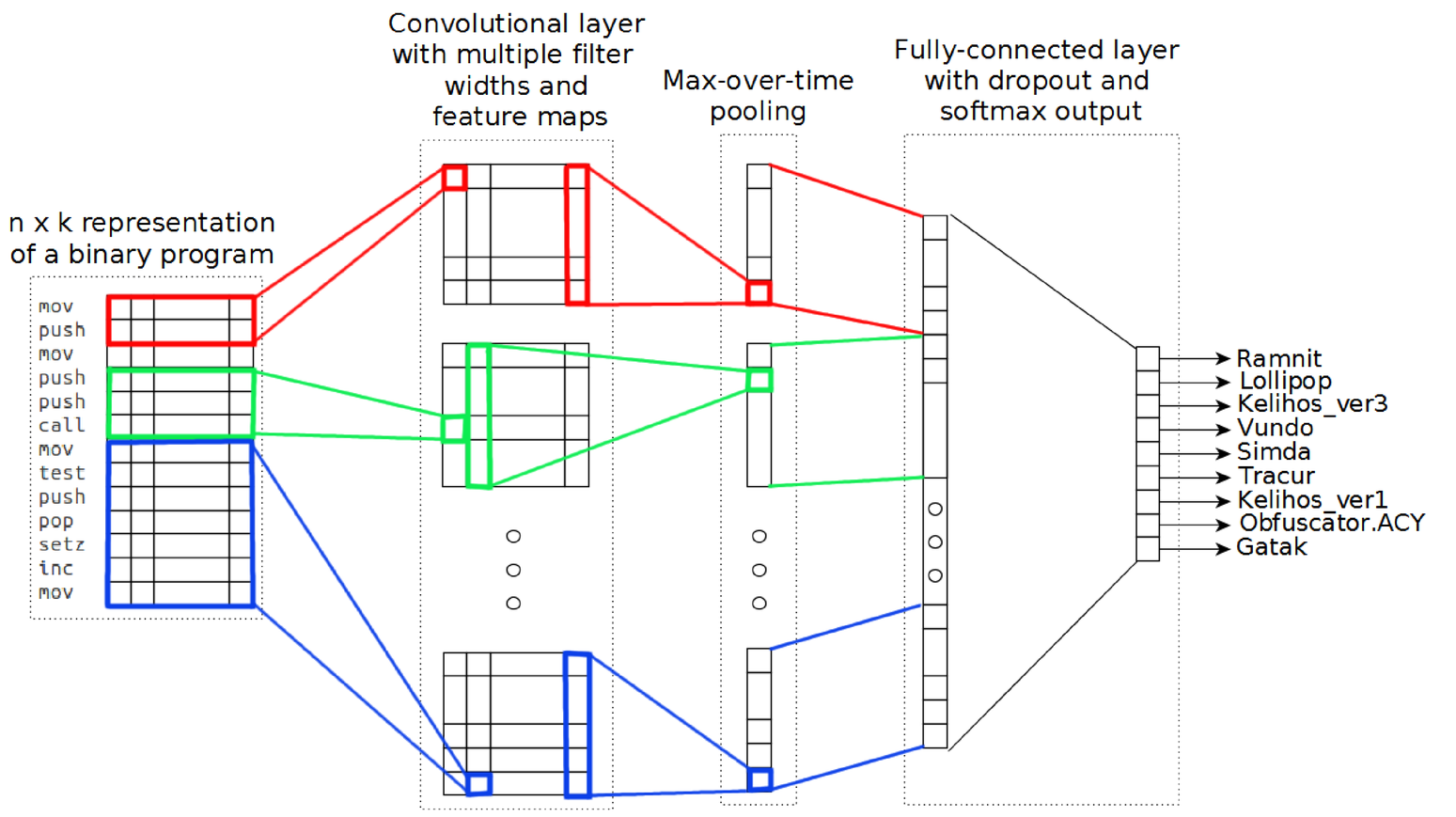
Structural Entropy



Contributions

- Hybrid approach to classify malware that combines feature engineering and deep learning.
- Feature level or early fusion mechanism to combine different types of features.
- SOTA performance on the Microsoft Malware Classification Challenge.

CNN-based N-Gram Extractor



SOTA Comparison

Table 1: Comparison with the state-of-the-art methods on the Microsoft Malware Classification Challenge benchmark.

		10-Fold Cross Validation		Test
Method	Input	Accuracy	Logloss	Logloss
Kalash et al. (2018)**	Grayscale images	0.9852	-	0.0571
Yuan et al. (2020)	Markov images	0.9926	0.0518	-
Jiang et al. (2019)	RGB images	0.9973	-	0.0220
Xiao et al. (2020)	Structural entropy	0.9972	-	0.0314
Yan et al. (2019)	Control flow graph	0.9925	0.0543	-
Hu et al. (2016)	Opcode 4-grams	0.9930	-	0.0546
Gibert et al. (2017)	Opcode sequence	0.9917	-	0.0244
Raff et al. (2018)	Byte sequence	0.9641	-	0.3071
Le et al. (2018)	Compressed byte sequence	0.9820	-	0.0774
Gibert et al.. (2020)	API calls, Opcode and byte sequences	0.9975	-	-
Gao et al. (2020)	Hand-crafted features	0.9969	-	-
Ahmadi et al. (2016)*	Hand-crafted features	0.9977	0.0096	0.0063
Zhang et al. (2016)	Hand-crafted features	0.9976	-	0.0042
Proposed system	**Hand-crafted and deep features**	**0.9981**	**0.0070**	**0.0040**

Acknowledgements

This project has received funding from Enterprise Ireland, the Spanish Science and Innovation Ministry funded project PID2019-111544GB-C22, the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska- Curie grant agreement No 847402.