Android Mobile Malware Detection using Machine Learning Techniques

R. Mariammal[1], Bommanaboina Haribabu[2], Konka Venkatesh[3], Korsipati Rahul Reddy[4]
Assistant Professor, Dhanalakshmi College of Engineering, Chennai, India[1]
Students, Dhanalakshmi College of Engineering, Chennai, India[2,3,4]

Abstract: *With the growth of malware and improvements in cyberattacks, malware detection is crucial to maintaining cyber security. These attacks frequently employ previously undetectable malware that is not the focus of security firms, and it is inevitable that solutions will be found to learn from unlabelled sample data. This paper introduces SHERLOCK, a deep learning model for malware detection based on self-monitoring and the Vision Transformer (ViT) architecture. Using binary representation based on images, SHERLOCK is a novel malware CA detection technology that learns distinctive traits to separate malware from the benign programmes they employ. Self-supervised learning can achieve accuracy 97% for binary malware classification, which is higher than current state-of-the-art algorithms, according to experimental results employing 1.2 million Android apps hierarchy of 47 categories and 696 families. With macro-F1 scores of 0.497 and 0.491, respectively, the suggested model can also outperform cutting-edge methods for multi-class malware classification types and families.*

Keywords: Supervised Learning, Deep Learning, Malware Detection, Android Security

REFERENCES

[1]. A. M. Al-Saffar, H. Tao, and M. A. Talab, ''Review of deep convolution neural network in image classification,'' in Proc. Int. Conf. Radar, Antenna, Microw., Electron., Telecommun. (ICRAMET), Oct. 2017, pp. 26–31.

[2]. Subramanya, V. Pillai, and H. Pirsiavash, ''Fooling network interpretation in image classification,'' in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 2020–2029.

[3]. Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu, ''Benchmarking adversarial robustness on image classification,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 321–331.

[4]. S. Qiu, Q. Liu, S. Zhou, and C. Wu, ''Review of artificial intelligence adversarial attack and defense technologies,'' Appl. Sci., vol. 9, no. 5, p. 909, 2019.

[5]. Z. Zhu and T. Dumitraş, ''FeatureSmith: Automatically engineering features for malware detection by mining the security literature,'' in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2016, pp. 767–778.

[6]. S.-J. Lee, H.-Y. Shim, Y.-R. Lee, T.-R. Park, S.-H. Park, and I.-G. Lee, ''Study on systematic ransomware detection techniques,'' in Proc. 23rd Int. Conf. Adv. Commun. Technol. (ICACT), Feb. 2021, pp. 297–301.

[7]. M. A. Omer, S. R. M. Zeebaree, M. A. M. Sadeeq, B. W. Salim, S. X. Mohsin, Z. N. Rashid, and L. M. Haji, ''Efficiency of malware detection in Android system: A survey,'' Asian J. Res. Comput. Sci., vol. 7, no. 4, pp. 59–69, Apr. 2021.

[8]. X. Deng and J. Mirkovic, ''Polymorphic malware behavior through network trace analysis,'' in Proc. 14th Int. Conf. Commun. Syst. Netw. (COMSNETS), Jan. 2022, pp. 138–146.

[9]. P. Faruki, A. Bharmal, V. Laxmi, V. Ganmoor, M. S. Gaur, M. Conti, and M. Rajarajan, ''Android security: A survey of issues, malware penetration, and defenses,'' IEEE Commun. Surveys Tuts., vol. 17, no. 2, pp. 998–1022, 2nd Quart., 2015.

[10]. E.Rezende,G.Ruppert,T.Carvalho,F.Ramos,andP.deGeus,''Malicious software classification using transfer learning of ResNet-50 deep neural network,'' in Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2017, pp. 1011–1014.

[11]. K. Allix, T. F. Bissyandé, J. Klein, and Y. L. Traon, ''Androzoo: Collecting millions of Android apps for the research community,'' in Proc. IEEE/ACM 13th Work. Conf. Mining Softw. Repositories (MSR), May 2016, pp. 468–471.

[12]. G. Conti, E. Dean, M. Sinda, and B. Sangster, ''Visual reverse engineering of binary and data files,'' in Proc. Int. Workshop Vis. Comput. Secur. Berlin, Germany: Springer-Verlag, 2008, pp. 1–17.

[13]. L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, ''Malware images: Visualization and automatic classification,'' in Proc. 8th Int. Symp. Vis. Cyber Secur. (VizSec). New York, NY, USA: Association for Computing Machinery, 2011, pp. 1–7, doi: 10.1145/2016904.2016908.

[14]. S. Freitas, R. Duggal, and D. H. Chau, ''MalNet: A large-scale cybersecurity image database of malicious software,'' CoRR, vol. abs/2102.01072, pp. 1–7, Jan. 2021.

[15]. J. Gennissen, L. Cavallaro, V. Moonsamy, and L. Batina, ''Gamut: Sifting through images to detect Android malware,'' Bachelor thesis, Inst. Comput. Inf. Sci., Roy. Holloway Univ., London, U.K., 2017.

[16]. A. Mohaisen, A. G. West, A. Mankin, and O. Alrawi, ''Chatter: Classifying malware families using system event ordering,'' in Proc. IEEE Conf. Commun. Netw. Secur., Oct. 2014, pp. 283–291.

[17]. D. Votipka, S. Rabin, K. Micinski, J. S. Foster, and M. L. Mazurek, ''An observational investigation of reverse engineers' processes,'' in Proc. 29th USENIX Secur. Symp. (USENIX Security), 2020, pp. 1875–1892.

[18]. A. Kantchelian, M. C. Tschantz, S. Afroz, B. Miller, V. Shankar, R. Bachwani, A. D. Joseph, and J. D. Tygar, ''Better malware ground truth: Techniques for weighting anti-virus vendor labels,'' in Proc. 8th ACM Workshop Artif. Intell. Secur., Oct. 2015, pp. 45–56.

[19]. M. Noroozi and P. Favaro, ''Unsupervised learning of visual representations by solving jigsaw puzzles,'' in Computer Vision—ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 69–84.

[20]. Doersch, A. Gupta, and A. A. Efros, ''Unsupervised visual representation learning by context prediction,'' in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1422–1430.

[21]. R. Zhang, P. Isola, and A. A. Efros, ''Colorful image colorization,'' in Computer Vision—ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 649–666.

[22]. X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, ''S4L: Self-supervised semi-supervised learning,'' in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1476–1485.

[23]. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, ''Using selfsupervised learning can improve model robustness and uncertainty,'' in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–12.

[24]. Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, ''Unsupervised data augmentation for consistency training,'' 2019, arXiv:1904.12848.

[25]. A. K. Bhunia, P. N. Chowdhury, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, ''Vectorization and rasterization: Self-supervised learning for sketch and handwriting,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 5672–5681.

[26]. S.Seneviratne,K.A.Nice,J.S.Wijnands,M.Stevenson,andJ. Thompson, ''Self-supervision. Remote sensing and abstraction: Representation learning across 3 million locations,'' in Proc. Digit. Image Comput., Techn. Appl. (DICTA), Nov. 2021, pp. 01–08.

[27]. S. Seneviratne, ''Contrastive representation learning for natural world imagery: Habitat prediction for 30,000 species,'' in Proc. CLEF Work. Notes, 2021, pp. 1639–1648.

[28]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16 × 16 words: Transformers for image recognition at scale,'' 2020, arXiv:2010.11929.

[29]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang.

Technol., vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[30]. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, ''Masked autoencoders are scalable vision learners,'' 2021, arXiv:2111.06377.

[31]. G. Lev Lafayette, L. Vu, and B. Meade, ''Spartan performance and flexibility: An HPC-cloud chimera,'' in Proc. OpenStack Summit, vol. 10, Barcelona, Spain, 2016, p. 49.

[32]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[33]. A. Abusitta, M. Q. Li, and B. C. M. Fung, ''Malware classification and composition analysis: A survey of recent developments,'' J. Inf. Secur. Appl., vol. 59, Jun. 2021, Art. no. 102828.

[34]. Damodaran, F. D. Troia, C. A. Visaggio, T. H. Austin, and M. Stamp, ''A comparison of static, dynamic, and hybrid analysis for malware detection,'' J. Comput. Virol. Hacking Techn., vol. 13, no. 1, pp. 1–12, 2017.

[35]. W. Huang and J. W. Stokes, ''MtNet: A multi-task neural network for dynamic malware classification,'' in Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment. Berlin, Germany: Springer-Verlag, 2016, pp. 399–418.

[36]. G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu, ''Large-scale malware classification using random projections and neural networks,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 3422–3426.

[37]. Y.-T. Lee, T. Ban, T.-L. Wan, S.-M. Cheng, R. Isawa, T. Takahashi, and D. Inoue, ''Cross platform IoT-malware family classification based on printable strings,'' in Proc. IEEE 19th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom), Dec. 2020, pp. 775–784.

[38]. Anderson, C. Storlie, and T. Lane, ''Improving malware classification: Bridging the static/dynamic gap,'' in Proc. 5th ACM Workshop Secur. Artif. Intell., 2012, pp. 3–14.

[39]. J. Saxe and K. Berlin, ''Deep neural network based malware detection using two dimensional binary program features,'' in Proc. 10th Int. Conf. Malicious Unwanted Softw. (MALWARE), Oct. 2015, pp. 11–20.

[40]. M. Amin, T. A. Tanveer, M. Tehseen, M. Khan, F. A. Khan, and S. Anwar, ''Static malware detection and attribution in Android byte-code through an end-to-end deep system,'' Future Gener. Comput. Syst., vol. 102, pp. 112–126, Jan. 2020.

[41]. N. Daoudi, J. Samhi, A. K. Kabore, K. Allix, T. F. Bissyandé, and J. Klein, ''DEXRAY: A simple, yet effective deep learning approach to Android malware detection based on image representation of bytecode,'' in Proc. Int. Workshop Deployable Mach. Learn. Secur. Defense. Cham, Switzerland: Springer, 2021, pp. 81–106.

[42]. Anderson, D. Quist, J. Neil, C. Storlie, and T. Lane, ''Graph-based malware detection using dynamic analysis,'' J. Comput. Virol., vol. 7, no. 4, pp. 247–258, 2011.

[43]. J. Dai, R. K. Guha, and J. Lee, ''Efficient virus detection using dynamic instruction sequences,'' J. Comput., vol. 4, no. 5, pp. 405–414, 2009.

[44]. M. K. Shankarapani, S. Ramamoorthy, R. S. Movva, and S. Mukkamala, ''Malware detection using assembly and API call sequences,'' J. Comput. Virol., vol. 7, no. 2, pp. 107–119, 2011.

[45]. A. Sami, B. Yadegari, N. Peiravian, S. Hashemi, and A. Hamze, ''Malware detection based on mining API calls,'' in Proc. ACM Symp. Appl. Comput., 2010, pp. 1020–1025.

[46]. Z. Salehi, A. Sami, and M. Ghiasi, ''Using feature generation from API calls for malware detection,'' Comput. Fraud Secur., vol. 2014, no. 9, pp. 9–18, 2014.

[47]. Z. Ma, H. Ge, Y. Liu, M. Zhao, and J. Ma, ''A combination method for Android malware detection based on control flow graphs and machine learning algorithms,'' IEEE Access, vol. 7, pp. 21235–21245, 2019.

[48]. J. Yan, G. Yan, and D. Jin, ''Classifying malware represented as control flow graphs using deep graph convolutional neural network,'' in Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN), Jun. 2019, pp. 52–63.

[49]. Bruschi, L. Martignoni, and M. Monga, ''Detecting self-mutating malware using control-flow graph matching,'' in Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment. Berlin, Germany: SpringerVerlag, 2006, pp. 129–143.

[50]. M. Cai, Y. Jiang, C. Gao, H. Li, and W. Yuan, ''Learning features from enhanced function call graphs for Android malware detection,'' Neurocomputing, vol. 423, pp. 301–307, Jan. 2021.

[51]. M. Al-Asli and T. A. Ghaleb, ''Review of signature-based techniques in antivirus products,'' in Proc. Int. Conf. Comput. Inf. Sci. (ICCIS), Apr. 2019, pp. 1–6.

[52]. M. G. Schultz, E. Eskin, F. Zadok, and S. J. Stolfo, ''Data mining methods for detection of new malicious executables,'' in Proc. IEEE Symp. Secur. Privacy (SP), May 2001, pp. 38–49.

[53]. M. Christodorescu and S. Jha, ''Static analysis of executables to detect malicious patterns,'' in Proc. 12th USENIX Secur. Symp. (USENIX Security), 2003, pp. 169–186.

[54]. J. Singh and J. Singh, ''Assessment of supervised machine learning algorithms using dynamic API calls for malware detection,'' Int. J. Comput. Appl., vol. 44, no. 3, pp. 270–277, Mar. 2022.

[55]. Y. Ki, E. Kim, and H. K. Kim, ''A novel approach to detect malware based on API call sequence analysis,'' Int. J. Distrib. Sensor Netw., vol. 11, no. 6, Jun. 2015, Art. no. 659101.

[56]. A. Tang, S. Sethumadhavan, and S. J. Stolfo, ''Unsupervised anomalybased malware detection using hardware features,'' in Proc. Int. Workshop Recent Adv. Intrusion Detection. Cham, Switzerland: Springer, 2014, pp. 109–129.

[57]. Z. Liu, R. Wang, N. Japkowicz, D. Tang, W. Zhang, and J. Zhao, ''Research on unsupervised feature learning for Android malware detection based on restricted Boltzmann machines,'' Future Gener. Comput. Syst., vol. 120, pp. 91–108, Jul. 2021.

[58]. M. Fan, X. Luo, J. Liu, M. Wang, C. Nong, Q. Zheng, and T. Liu, ''Graph embedding based familial analysis of Android malware using unsupervised learning,'' in Proc. IEEE/ACM 41st Int. Conf. Softw. Eng. (ICSE), May 2019, pp. 771–782.

[59]. Santos, J. Nieves, and P. G. Bringas, ''Semi-supervised learning for unknown malware detection,'' in Proc. Int. Symp. Distrib. Comput. Artif. Intell. Springer, 2011, pp. 415–422.

[60]. A. Mahindru and A. Sangal, ''Feature-based semi-supervised learning to detect malware from Android,'' in Automated Software Engineering: A Deep Learning-Based Approach. Springer, 2020, pp. 93–118.

[61]. X.Gao,C.Hu,C.Shan,B.Liu,Z.Niu,andH.Xie,''Malwareclassification for the cloud via semi-supervised transfer learning,'' J. Inf. Secur. Appl., vol. 55, Dec. 2020, Art. no. 102661.

[62]. A. Souri and R. Hosseini, ''A state-of-the-art survey of malware detection approaches using data mining techniques,'' Hum.-Centric Comput. Inf. Sci., vol. 8, no. 1, pp. 1–22, Dec. 2018.

[63]. K. Liu, S. Xu, G. Xu, M. Zhang, D. Sun, and H. Liu, ''A review of Android malware detection approaches based on machine learning,'' IEEE Access, vol. 8, pp. 124579–124607, 2020.

[64]. P. Yadav, N. Menon, V. Ravi, S. Vishvanathan, and T. D. Pham, ''A twostage deep learning framework for image-based Android malware detection and variant classification,'' Comput. Intell., May 2022.

[65]. M. T. Ahvanooey, Q. Li, M. Rabbani, and A. R. Rajput, ''A survey on smartphones security: Software vulnerabilities, malware, and attacks,'' 2020, arXiv:2001.09406.

[66]. Z. Ren and Y. J. Lee, ''Cross-domain self-supervised multi-task feature learning using synthetic imagery,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 762–771.

[67]. Z. Chen, R. Ding, T.-W. Chin, and D. Marculescu, ''Understanding the impact of label granularity on CNN-based image classification,'' in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Nov. 2018, pp. 895–904.