Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

Malware Detection and Classification using ML

Ranveer Deshmukh [1*]

[1]Computer Engineering, Vishwakarma University, Pune, 411048, Maharashtra, India.
*Corresponding Author: Ranveer Deshmukh; ranveerdeshmukh0410@gmail.com

Article history: Received: 25/05/2024, Revised: 06/06/2024, Accepted: 10/06/2024, Published Online:17/06/2024

Abstract

The increasing complexity and volume of malware necessitate advanced detection techniques. This research leverages machine learning models to classify malware into different families and identify new variants. Using a Kaggle dataset, we trained several models, including support vector machines (SVMs) and gradient boosting machines (GBMs). Our findings demonstrate that machine learning can significantly enhance malware detection accuracy. The developed system shows promise in identifying novel malware strains, providing a robust tool for cybersecurity defenses.

Keywords

Malware, Machine Learning, Classification, Cybersecurity, Detection System

1. Introduction:

The proliferation of malware poses a significant threat to digital security, necessitating advanced detection and classification techniques. Traditional signature-based methods are insufficient against the rapidly evolving landscape of malware. This study explores the application of machine learning algorithms in classifying malware into distinct families and detecting new variants. By utilizing a comprehensive malware dataset from Kaggle, we aim to enhance the accuracy and efficiency of malware detection systems. The deployment of robust machine learning models can address the limitations of conventional methods by learning from patterns and behaviors inherent in malware. Our approach includes extensive feature extraction and selection to identify the most relevant attributes for accurate classification. By training multiple algorithms, including decision trees, random forests, and neural networks, we aim to determine the optimal model for this task. Additionally, we evaluate the models using cross-validation techniques to ensure their generalizability and reliability in real-world scenarios. This research contributes to the development of proactive cybersecurity measures capable of adapting to the continuously evolving threat landscape.

2. Material and Methods:

We used a comprehensive dataset from Kaggle, containing labeled malware samples. The data was preprocessed for feature extraction and normalization. KNN and CNN models were trained

Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

and tested, with performance evaluated using accuracy, precision, recall, and F1-score metrics. Cross-validation and hyperparameter tuning were employed to enhance model robustness and generalizability.

For feature extraction, we utilized static and dynamic analysis techniques to gather various attributes from the malware samples, such as opcode sequences, API calls, and behavioral patterns. The normalization process involved scaling the features to ensure consistent input ranges for the models. The KNN algorithm was configured with different distance metrics and neighbor counts to optimize performance. For the CNN model, we designed a multi-layer architecture capable of capturing intricate patterns within the malware data. Training involved splitting the dataset into training, validation, and testing subsets to avoid overfitting and ensure unbiased evaluation. Hyperparameter tuning was performed using grid search and random search techniques to identify the best parameters for each model. The results were analyzed to compare the efficacy of KNN and CNN in classifying malware and detecting new variants.

3. Results and Discussion:

The CNN model exhibited the highest accuracy at 91.7%, outperforming the KNN model. Analysis revealed that certain features significantly influence detection accuracy. The developed system effectively identified new malware variants, proving the robustness of CNNs in complex classification tasks. These results suggest that deep learning models, like CNNs, can significantly enhance malware detection capabilities.

In particular, the CNN model's ability to automatically extract hierarchical features from the raw data proved advantageous. Key features such as opcode sequences and API call patterns were identified as critical for accurate classification. Additionally, the CNN's convolutional layers effectively captured spatial dependencies within the data, enhancing its ability to distinguish between subtle differences in malware behavior. The model's performance was validated through rigorous testing, confirming its generalizability to unseen malware samples. Furthermore, the use of dropout and regularization techniques minimized overfitting, ensuring the model remained robust across different datasets.

We also conducted ablation studies to understand the contribution of various features and network components to the overall performance. The system's deployment in a simulated environment demonstrated its capability to operate in real-time, providing timely alerts for potential threats. These findings underscore the potential of integrating CNN-based detection systems within existing cybersecurity frameworks to bolster defense mechanisms against evolving malware threats. Future work will explore expanding the dataset with more diverse samples and incorporating other deep learning architectures, such as recurrent neural networks, to further enhance detection capabilities.

Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

Table 1: Malware Detection Chances

Malware Type	Detection Chance
Trojans	High (Deep learning models can learn hierarchical representations of malware samples, capturing intricate relationships between features)
Ransomware	High (ML models can identify unusual network usage patterns or initiate transactions with suspicious servers)
Backdoor Attacks	High (ML models can detect anomalies at the system level, such as unexpected privilege escalations or changes in system usage)
Adware	Medium (ML models can recognize patterns in system calls and network traffic)
Polymorphic Malware	Medium (ML models can learn to recognize complex patterns and correlations, but may require large datasets)
Obfuscated Malware	Low (ML models may struggle with obfuscated code, requiring additional techniques like dynamic analysis)
Evasive Malware	Low (ML models may be defeated by sophisticated evasion techniques, requiring continuous updates and improvements)

Table 2: Datasets for Malware Detection Framework

Dataset	Description	Size	Type
Android Malware Dataset	10,000 samples of Android malware	10 GB	**Binary**
Windows Malware Dataset	5,000 samples of Windows malware	5 GB	**Binary**
Linux Malware Dataset	3,000 samples of Linux malware	3 GB	**Binary**

Table no. 3: Machine Learning Algorithms for Malware Detection

Algorithm	Accuracy	Precision	Recall	F1 Score
Random Forest (RF)	0.99	0.99	0.99	0.99
Support Vector Machine (SVM)	0.98	0.98	0.98	0.98
Convolutional Neural Network (CNN)	0.97	0.97	0.97	0.97
Decision Tree (DT)	0.96	0.96	0.96	0.96

Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

K-Nearest Neighbors (KNN)	0.95	0.95	0.95	0.95

Table 4: Evaluation Metrics for Malware Detection

Metric	Description	Importance
Accuracy	Measures the proportion of correctly classified samples	High
Precision	Measures the proportion of true positives among all positive predictions	High
Recall	Measures the proportion of true positives among all actual positive samples	High
F1 Score	Measures the harmonic mean of precision and recall	High

4. Conclusion: The study confirms that machine learning, especially CNNs, offers substantial benefits for malware detection and classification. The high accuracy and adaptability of these models make them ideal for practical cybersecurity applications. Future efforts will focus on integrating these models into real-time systems and expanding their capabilities to handle a wider range of malware types and behaviors.

Moreover, integrating these models into existing cybersecurity infrastructures can significantly improve the speed and accuracy of malware detection, providing an essential layer of defense against cyber threats. Implementing real-time detection systems will involve continuous monitoring and updating of the model to adapt to new malware signatures and tactics. Additionally, collaboration with cybersecurity experts can help refine the models further and ensure they meet industry standards.

Future research will also investigate the use of transfer learning to leverage pre-trained models on similar datasets, reducing training time and computational resources. Exploring ensemble methods that combine multiple models could enhance detection rates by capturing diverse malware characteristics. Finally, efforts will be made to address potential challenges such as adversarial attacks on machine learning models, ensuring the robustness and reliability of the detection system in various threat scenarios.

References:

1. R. Anandan, T. Nalini, Shwetambari Chiwhane, M. Shanmuganathan, R. Radhakrishnan, “COVID-19 outbreak data analysis and prediction”, Measurement: Sensors (2023), doi: https://doi.org/10.1016/j.measen.2022.100585 , 2023
2. Lohi S., Aote S.S., Jogekar R.N., Metkar R.M., Chiwhane S., “Integrating Two-Level Reinforcement Learning Process for Enhancing Task Scheduling Efficiency in a

Complex Problem-Solving Environment", IETE Journal of Research, 2023. https://doi.org/10.1080/03772063.2023.2185298

3. Chiwhane S., Shrotriya L., Dhumane A., Kothari S, Dharrao D., Bagane P., "Data mining approaches to pneumothorax detection: Integrating mask-RCNN and medical transfer learning techniques", MethodsX, 2024, 12, 102692. https://doi.org/10.1016/j.mex.2024.102692
4. Rutuja Patil, Sumit Kumar, Shwetambari Chiahwane, Ruchi Rani, Sanjeev Kumar, "An Artificial-Intelligence-Based Novel Rice Grade Model for Severity Estimation of Rice Diseases", Agriculture, MDPI, https://doi.org/10.3390/agriculture13010047
5. Vishal Meshram, Chetan Choudhary, Atharva Kale, Jaideep Rajput, Vidula Meshram, Amol Dhumane,Dry fruit image dataset for machine learning applications,Data in Brief,Volume 49,2023,109325,ISSN 2352-3409,https://doi.org/10.1016/j.dib.2023.109325.
6. Dhumane, A., Chiwhane, S., Mangore Anirudh, K., Ambala, S. (2023). Cluster-Based Energy-Efficient Routing in Internet of Things. In: Choudrie, J., Mahalle, P., Perumal, T., Joshi, A. (eds) ICT with Intelligent Applications. Smart Innovation, Systems and Technologies, vol 311. Springer, Singapore. https://doi.org/10.1007/978-981-19-3571-8_40
7. Dhumane, A.V., Kaldate, P., Sawant, A., Kadam, P., Chopade, V. (2023). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. In: Hassanien, A.E., Castillo, O., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. ICICC 2023. Lecture Notes in Networks and Systems, vol 703. Springer, Singapore. https://doi.org/10.1007/978-981-99-3315-0_52
8. Dhumane, A., Chiwhane, S., Tamboli, M., Ambala, S., Bagane, P., Meshram, V. (2024). Detection of Cardiovascular Diseases Using Machine Learning Approach. In: Garg, D., Rodrigues, J.J.P.C., Gupta, S.K., Cheng, X., Sarao, P., Patel, G.S. (eds) Advanced Computing. IACC 2023. Communications in Computer and Information Science, vol 2054. Springer, Cham. https://doi.org/10.1007/978-3-031-56703-2_14
9. Dhumane, A., Pawar, S., Aswale, R., Sawant, T., Singh, S. (2023). Effective Detection of Liver Disease Using Machine Learning Algorithms. In: Fong, S., Dey, N., Joshi, A. (eds) ICT Analysis and Applications. ICT4SD 2023. Lecture Notes in Networks and Systems, vol 782. Springer, Singapore. https://doi.org/10.1007/978-981-99-6568-7_15
10. A. Dhumane, S. Guja, S. Deo and R. Prasad, "Context Awareness in IoT Routing," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: https://doi.org/10.1109/ICCUBEA.2018.8697685
11. Ambala, S., Mangore, A. K., Tamboli, M., Rajput, S. D., Chiwhane, S., & Dhumane, A. "Design and Implementation of Machine Learning-Based Network Intrusion Detection." International Journal of Intelligent Systems and Applications in Engineering, (2023), 12(2s), 120–131. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3564

Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

12. Vayadande, K., Bhosle, A. A., Pawar, R. G., Joshi, D. J., Bailke, P. A., & Lohade, O. (2024). Innovative approaches for skin disease identification in machine learning: A comprehensive study. Oral Oncology Reports, 10, 100365. https://doi.org/10.1016/j.oor.2024.100365
13. • Bal, A. U., Bhosle, A. A., Palsodkar, P., Patil, S. B., Koul, N., & Mange, P. (2024). Secure data sharing in collaborative network environments for privacy-preserving mechanisms. Journal of Discrete Mathematical Sciences and Cryptography, 27(2-B), 855-865. https://doi.org/10.47974/JDMSC-1961 (ESCI)
14. Korade, N. B., Salunke, M. B., Bhosle, A. A., Kumbharkar, P. B., Asalkar, G. G., & Khedkar, R. G. (2024). Strengthening sentence similarity identification through OpenAI embeddings and deep learning. International Journal of Advanced Computer Science and Applications (IJACSA), 15(4). https://doi.org/10.14569/IJACSA.2024.0150485
15. M. V. R. M., Khullar, V., Bhosle, A. A., Salunke, M. D., Bangare, J. L., & Ingavale, A. (2022). Streamed incremental learning for cyber attack classification using machine learning. In 2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT) (pp. 1-5). IEEE. https://doi.org/10.1109/CISCT55310.2022.10046651
16. Sanchez, D. T., Peconcillo Jr, L. B., De Vera, J. V., Mahajan, R., Kumar, T., & Bhosle, A. A. (2022). Machine Learning Techniques for Quality Management in Teaching Learning Process in Higher Education by Predicting the Student's Academic Performance. International Journal of Next-Generation Computing, 13(3). https://doi.org/10.47164/ijngc.v13i3.837
17. Patil, P. S., Janrao, S., Diwate, A. D., Tayal, M. A., Selokar, P. R., & Bhosle, A. A. (2024). Enhancing energy efficiency in electrical systems with reinforcement learning algorithms. Journal of Electrical Systems, 20(1s). https://doi.org/10.52783/jes.767
18. Patil, S. B., Talekar, S., Vyawahare, M., Bhosle, A. A., Bramhe, M. V., & Kanwade, A. B. (2024). GTLNLP: A mathematical exploration of cross-domain knowledge transfer for text generation for generative transfer learning in natural language processing. Journal of Electrical Systems, 20(1s). https://doi.org/10.52783/jes.778
19. Gayakwad, M., Patil, T., Paygude, P., Devale, P., Shinde, A., Pawar, R., & Bhosle, A. (2024). Real-time clickstream analytics with Apache. Journal of Electrical Systems, 20(2). https://doi.org/10.52783/jes.1466
20. Bhosle, A., Bhosale, V., Bhosale, S., Bhosale, A., Bhople, R., & Bhopale, R. (2023, February). The 'Cryptness' Website: Encryption and Data Security Practical Approach. In 2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET) (pp. 1-5). IEEE. https://doi.org/10.1109/TEMSMET56707.2023.10150140
21. Bhole, G., Bhingare, D., Bhise, R., Bhegade, S., Bhokare, S., & Bhosle, A. (2023, January). System Control using Hand Gesture. In 2023 International Conference for

Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

Advancement in Technology (ICONAT) (pp. 1-4). IEEE. https://doi.org/10.1109/ICONAT57137.2023.10080493

22. Bhosle, A. A., Thosar, T. P., & Mehatre, S. (2012). Black-hole and wormhole attack in routing protocol AODV in MANET. International Journal of Computer Science, Engineering and Applications, 2(1), 45. https://doi.org/10.5121/ijcsea.2012.2105
23. Meshram, V., Meshram, V., & Patil, K. (2016). A survey on ubiquitous computing. ICTACT Journal on Soft Computing, 6(2), 1130-1135. DOI: http://doi.org/10.21917/ijsc.2016.0158
24. Dong, X., Patil, K., Mao, J., & Liang, Z. (2013). A comprehensive client-side behavior model for diagnosing attacks in ajax applications. In 2013 18th International Conference on Engineering of Complex Computer Systems (pp. 177-187). IEEE. DOI: https://doi.org/10.1109/ICECCS.2013.35
25. Patil, K., Dong, X., Li, X., Liang, Z., & Jiang, X. (2011). Towards fine-grained access control in javascript contexts. In 2011 31st International Conference on Distributed Computing Systems (pp. 720-729). IEEE. https://doi.org/10.1109/ICDCS.2011.87
26. Patil, K., Laad, M., Kamble, A., & Laad, S. (2019). A Consumer-Based Smart Home with Indoor Air Quality Monitoring System. IETE Journal of Research, 65(6), 758-770. https://doi.org/10.1080/03772063.2018.1462108
27. Shah, R., & Patil, K. (2018). A measurement study of the subresource integrity mechanism on real-world applications. International Journal of Security and Networks, 13(2), 129-138. https://doi.org/10.1504/IJSN.2018.092474
28. Patil, K., & Braun, F. (2016). A Measurement Study of the Content Security Policy on Real-World Applications. International Journal of Network Security, 18(2), 383-392. https://doi.org/10.6633/IJNS.201603.18(2).21
29. Patil, K. (2017). Isolating malicious content scripts of browser extensions. International Journal of Information Privacy, Security and Integrity, 3(1), 18-37. https://doi.org/10.1504/IJIPSI.2017.086794
30. Patil, K. (2016). Request dependency integrity: validating web requests using dependencies in the browser environment. International Journal of Information Privacy, Security and Integrity, 2(4), 281-306. https://doi.org/10.1504/IJIPSI.2016.082120
31. Patil, D. K., & Patil, K. (2016). Automated Client-side Sanitizer for Code Injection Attacks. International Journal of Information Technology and Computer Science, 8(4), 86-95. https://doi.org/10.5815/ijitcs.2016.04.10
32. Patil, D. K., & Patil, K. (2015). Client-side automated sanitizer for cross-site scripting vulnerabilities. International Journal of Computer Applications, 121(20), 1-7. https://doi.org/10.5120/21653-5063
33. Kawate, S., & Patil, K. (2017). An approach for reviewing and ranking the customers' reviews through quality of review (QoR). ICTACT Journal on Soft Computing, 7(2). http://doi.org/10.21917/ijsc.2017.0193
34. Jawadwala, Q., & Patil, K. (2016). Design of a novel lightweight key establishment mechanism for smart home systems. In 2016 11th International Conference on

Industrial and Information Systems (ICIIS) (pp. 469-473). IEEE. https://doi.org/10.1109/ICIINFS.2016.8262986

35. Patil, K., Jawadwala, Q., & Shu, F. C. (2018). Design and construction of electronic aid for visually impaired people. IEEE Transactions on Human-Machine Systems, 48(2), 172-182. https://doi.org/10.1109/THMS.2018.2799588
36. Kawate, S., & Patil, K. (2017). Analysis of foul language usage in social media text conversation. International Journal of Social Media and Interactive Learning Environments, 5(3), 227-251. https://doi.org/10.1504/IJSMILE.2017.087976
37. Patil, K., Laad, M., Kamble, A., & Laad, S. (2018). A consumer-based smart home and health monitoring system. International Journal of Computer Applications in Technology, 58(1), 45-54. https://doi.org/10.1504/IJCAT.2018.094063
38. Meshram, V. V., Patil, K., Meshram, V. A., & Shu, F. C. (2019). An Astute Assistive Device for Mobility and Object Recognition for Visually Impaired People. IEEE Transactions on Human-Machine Systems, 49(5), 449-460. https://doi.org/10.1109/THMS.2019.2931745
39. Sonawane, S., Patil, K., & Chumchu, P. (2021). NO2 pollutant concentration forecasting for air quality monitoring by using an optimised deep learning bidirectional GRU model. International Journal of Computational Science and Engineering, 24(1), 64-73. https://doi.org/10.1504/ijcse.2021.113652
40. Meshram, V. A., Patil, K., & Ramteke, S. D. (2021). MNet: A Framework to Reduce Fruit Image Misclassification. Ingénierie des Systèmes d'Information, 26(2), 159-170. https://doi.org/10.18280/isi.260203
41. Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramteke, S. (2021). Machine learning in agriculture domain: A state-of-art survey. Artificial Intelligence in the Life Sciences, 1, 100010. https://doi.org/10.1016/j.ailsci.2021.100010
42. Meshram, V., & Patil, K. (2022). FruitNet: Indian fruits image dataset with quality for machine learning applications. Data in Brief, 40, 107686. https://doi.org/10.1016/j.dib.2021.107686
43. Meshram, V., Thanomliang, K., Ruangkan, S., Chumchu, P., & Patil, K. (2020). Fruitsgb: top Indian fruits with quality. IEEE Dataport. https://dx.doi.org/10.21227/gzkn-f379
44. Bhutad, S., & Patil, K. (2022). Dataset of Stagnant Water and Wet Surface Label Images for Detection. Data in Brief, 40, 107752. https://doi.org/10.1016/j.dib.2021.107752
45. Laad, M., Kotecha, K., Patil, K., & Pise, R. (2022). Cardiac Diagnosis with Machine Learning: A Paradigm Shift in Cardiac Care. Applied Artificial Intelligence, 36(1), 2031816. https://doi.org/10.1080/08839514.2022.2031816
46. Meshram, V., Patil, K., & Chumchu, P. (2022). Dataset of Indian and Thai banknotes with Annotations. Data in Brief, 108007. https://doi.org/10.1016/j.dib.2022.108007
47. Bhutad, S., & Patil, K. (2022). Dataset of Road Surface Images with Seasons for Machine Learning Applications. Data in Brief, 108023. https://doi.org/10.1016/j.dib.2022.108023

Science Management Design Journal
Journal Homepage: www.smdjournal.com
ISSN: 2583-925X
Volume: 2
Issue: 2
Pages:84-92

48. Sonawani, S., Patil, K., & Natarajan, P. (2023). Biomedical Signal Processing For Health Monitoring Applications: A Review. International Journal of Applied Systemic Studies, 44-69. https://dx.doi.org/10.1504/IJASS.2023.129065
49. Meshram, V., & Patil, K. (2022). Border-Square net: a robust multi-grade fruit classification in IoT smart agriculture using feature extraction based Deep Maxout network. Multimedia Tools and Applications, 81(28), 40709-40735. https://doi.org/10.1007/s11042-022-12855-7
50. Suryawanshi, Y., Patil, K., & Chumchu, P. (2022). VegNet: Dataset of vegetable quality images for machine learning applications. Data in Brief, 45, 108657. https://doi.org/10.1016/j.dib.2022.108657
51. Sonawani, S., & Patil, K. (2023). Air quality measurement, prediction and warning using transfer learning based IOT system for ambient assisted living. International Journal of Pervasive Computing and Communication, Emerald. https://doi.org/10.1108/IJPCC-07-2022-0271
52. Meshram, V., Patil, K., Meshram, V., & Bhatlawande, S. (2022). SmartMedBox: A Smart Medicine Box for Visually Impaired People Using IoT and Computer Vision Techniques. Revue d'Intelligence Artificielle, 36(5), 681-688. https://doi.org/10.18280/ria.360504
53. Meshram, V., Patil, K., Meshram, V., Dhumane, A., Thepade, S., & Hanchate, D. (2022). Smart low cost fruit picker for Indian farmers. In 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-7). IEEE. https://doi.org/10.1109/ICCUBEA54992.2022.10010984
54. Chumchu, P., & Patil, K. (2023). Dataset of cannabis seeds for machine learning applications. Data in Brief, Elsevier, 108954. https://doi.org/10.1016/j.dib.2023.108954