# The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches

*Hazim Hanif[ab], Mohd Hairul Nizam Md Nasir[b], Mohd Faizal Ab Razak[c], Ahmad Firdaus[c], Nor Badrul Anuar[d]*

[a]Department of Computing, Faculty of Engineering, Imperial College London, London, SW7 2AZ, United Kingdom
[b]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603, Malaysia
[c]Faculty of Computing, Universiti Malaysia Pahang, Pekan, Pahang, 26600, Malaysia
[d]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603, Malaysia

## ABSTRACT

The detection of software vulnerability requires critical attention during the development phase to make it secure and less vulnerable. Vulnerable software always invites hackers to perform malicious activities and disrupt the operation of the software, which leads to millions in financial losses to software companies. In order to reduce the losses, there are many reliable and effective vulnerability detection systems introduced by security communities aiming to detect the software vulnerabilities as early as in the development or testing phases. To summarise the software vulnerability detection system, existing surveys discussed the conventional and data mining approaches. These approaches are widely used and mostly consist of traditional detection techniques. However, they lack discussion on the newly trending machine learning approaches, such as supervised learning and deep learning techniques. Furthermore, existing studies fail to discuss the growing research interest in the software vulnerability detection community throughout the years. With more discussion on this, we can predict and focus on what are the research problems in software vulnerability detection that need to be urgently addressed. Aiming to reduce these gaps, this paper presents the research interests' taxonomy in software vulnerability detection, such as methods, detection, features, code and dataset. The research interest categories exhibit current trends in software vulnerability detection. The analysis shows that there is considerable interest in addressing methods and detection problems, while only a few are interested in code and dataset problems. This indicates that there is still much work to be done in terms of code and dataset problems in the future. Furthermore, this paper extends the machine learning approaches taxonomy, which is used to detect the software vulnerabilities, like supervised learning, semi-supervised learning, ensemble learning and deep learning. Based on the analysis, supervised learning and deep learning approaches are trending in the software vulnerability detection community as these techniques are able to detect vulnerabilities such as buffer overflow, SQL injection and cross-site scripting effectively with a significant detection performance, up to 95% of F1 score. Finally, this paper concludes with several discussions on potential future work in software vulnerability detection in terms of datasets, multi-vulnerabilities detection, transfer learning and real-world applications.

**KEYWORDS:** Software vulnerability detection, Software security, Computer security, Machine learning, Deep learning

**DOI:** https://doi.org/10.1016/j.jnca.2021.103009

## ACKNOWLEDGEMENT

The work of the authors was supported by the University of Malaya Research Grant Programme (UMRG) under grant RF008D-2018 and partly by the Ministry of Higher Education, Malaysia under grant RACER/1/2019/ICT03/UMP/2.

## REFERENCES

Afifi, F., Anuar, N.B., Shamshirband, S., Choo, K.-K.R., 2016. DyHAP: dynamic hybrid ANFIS-PSO approach for predicting mobile malware. PloS One 11 (9), e0162627. https://doi.org/10.1371/journal.pone.0162627.

Alves, H., Fonseca, B., Antunes, N., 2016. Experimenting machine learning techniques to predict vulnerabilities. In: Paper Presented at the 2016 Seventh Latin-American Symposium on Dependable Computing (LADC).

Ban, X., Liu, S., Chen, C., Chua, C., 2019. A performance evaluation of deep-learnt features for software vulnerability detection. Concurrency Comput. Pract. Ex. 31 (19), e5103 https://doi.org/10.1002/cpe.5103.

Bissell, K., Cin, R.M.L.P.D., 2019. Ninth Annual Cost of Cybercrime Study. Retrieved from. https://www.accenture.com/us-en/insights/security/cost-cybercrime-study.

Bosu, A., Carver, J.C., Hafiz, M., Hilley, P., Janni, D., 2014. Identifying the characteristics of vulnerable code changes: an empirical study. In: Paper Presented at the Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering, Hong Kong, China.