Giulia Sellitto and Filomena Ferrucci, University of Salerno (Italy)

The Impact of Release-based Training on Software Vulnerability Prediction Models

Software vulnerability prediction models can give us the ability to predict which portions of code are more prone to contain vulnerabilities and focus testing effort, potentially increasing code quality and reducing security threats.

Most of the proposed models, as pointed out by Jimenez et al. [1], have been evaluated by researchers using cross-validation, but in a real-case scenario, one is interested in training the model using information related to prior releases of software and obtaining predictions on the current version to be released. So there is a gap between the performance observed in research studies and those that would be obtained in a real environment.

With this work we aim to start bridging this gap, by performing a preliminary study on:

What is the performance of vulnerability prediction models trained using a release-based approach when compared to models trained using cross-validation, and which modelling approach is more sensitive to the use of a different validation method?

We replicate the study performed by Walden et al. [2] considering the suggestions made by Jimenez et al. [1]. Initial findings demonstrate that models' performance drop drastically when considering a release-based training and validation method.

The table summarizes Random Forest classifier performance on PHPMyAdmin with undersampling, as measured by Matthews Correlation Coefficient

	Cross-validation	Release-based
Software Metrics	0.72	0.15
Text Mining	0.82	0.36

[1] Matthieu Jimenez, Renaud Rwemalika, Mike Papadakis, Federica Sarro, Yves Le Traon, and Mark Harman. 2019. **The Importance of Accounting for Real-World Labelling When Predicting Software Vulnerabilities**. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*(Tallinn, Estonia)(ESEC/FSE 2019). Association for Computing Machinery, New York, NY, USA,695–705. https://doi.org/10.1145/3338906.3338941

[2] James Walden, Jeff Stuckman, and Riccardo Scandariato. 2014. **Predicting Vulnerable Components: Software Metrics vs Text Mining**. In *2014 IEEE 25th International Symposium on Software Reliability Engineering*. 23–33. https://doi.org/10.1109/ISSRE.2014.32