Supplements of Experimental Data and Results

1 The Datasets

Security scanning is a time-consuming task and in order to train the machine learning models, a large number of training examples are required to effectively learn the signatures of security vulnerabilities directly from the source code. So we used 17 application which source code was previously tagged as vulnerable or not vulnerable. We used 17 applications in this work from 3 different pre-labelled datasets. The used applications are the OWASP Benchmark Project, the Juliet Java test suit and 16 android applications from the Android Study dataset. In the next sections we enumerate and describe the different datasets.

1.1 OWASP Benchmark Project

The *Open Web Application Security Project* (OWASP) Benchmark Project [6] is a free test suite designed to evaluate the automated software vulnerability detection tools. It is a fully runnable java application that contains different test cases (vulnerable and not vulnerable), which are mapped to specific CWEs. This Project contains 2 740 test cases, 1415 vulnerable files (52%) and 1325 not vulnerable files (48%).

1.2 Juliet test suite in Java

The *Juliet test suite* in Java [8] is a collection of small test program in Java. These test cases exhibit examples of 112 different vulnerabilities and errors such as buffer overflow, OS injection, hard-coded password, absolute path traversal, NULL pointer dereference, uncaught exception, deadlock, and missing release of resource. This test suite is used to evaluate the capability of vulnerability detection tools in identifying vulnerabilities and flows. This application contains 217 vulnerable files (58%) and 297 not vulnerable files (42%).

1.3 Android Study

The *Android Study* is a public dataset that contains the label of 20 Java applications files that cover a variety of domains. This dataset was created to be used in a previous work of R. Scandariato, J. Walden[7].

To generate the dataset, multiple mobile applications were scanned with a security-oriented scanning tool, Fortify Source Code Analyzer[2], and each file was labelled as vulnerable or not vulnerable.

The dataset does not provide us with the exact type of vulnerability. However, according to [7] we know that it contains 22 776 unique vulnerabilities in all the code files such as cross-site scripting, SQL injection, header manipulation, privacy violation and command injection. The dataset provides the name of each application, the version and the paths of the vulnerable files. However, it did not contain the source code. Using the the project names and the versions, we retrieved 16 of the 20 android applications which we used in this work.

2 Experimental Results

2.1 Machine Learning Results

The tables 1, 2 and 3 shows the results obtained from the 408 experiments. We experiment with two tokenization techniques, and three types of embedding methods (namely bag-of-words[10]; word2vec[5], and fastText[4]) across three kinds of machine learning algorithms. The three machine learning algorithms include a weak learner-based model (random forest[9]), a kernel vector-based model(support vector machines[1]), and a neural network model(residual neural network[3]). We evaluate their combination effects on the accuracy of vulnerability detection over 17 java projects.

Comparing the results of table 2 and 1 show that comments and symbols do not affect the learning of software vulnerabilities from source code. Also, comparing the use of the different embedding method for each experiment, bag-of-words shows better results than the other two models in the leaning process of vulnerabilities signatures.

And in the two tables 2 and 1 the model random forest has shown to be more suitable for learning vulnerability detection compared to SVM and Resnet.

The results in table 3 show that combining the tokens and the metrics didn't improve the results and using the tokens only showed better performance than using the metrics only for vulnerability detection.

Table 1: Results from using all the tokens of the source code with bag-of-words, word2vec and fastText

		Bag-of-words							Word2vec							FastText						
Project	Classifier	P	R	F1	FPR	ROC AUC	PR AUC	z	P	R	F3	FPR	ROC AUC	PR AUC	z	P	R	F2	FPR	ROC AUC	PR AUC	z
1 OWASP	RF	1.00	1.00	1.00	0.21	1.00	1.00	0.95	0.68	0.76	0.72	0.21	0.70	0.63	0.60	1.00	1.00	1.00	0.21	1.00	1.00	0.95
	Resnet	0.99	0.92	0.95	0.10	0.96	0.95	0.92	0.95	0.93	0.94	0.04	0.94	0.92	0.92	0.76	0.96	0.84	0.04	0.85	0.87	0.82
	SVM	0.99	0.99	0.99	0.24	0.99	0.99	0.93	0.91	0.93	0.92	0.19	0.93	0.89	0.86	0.82	0.90	0.86	0.08	0.93	0.92	0.85
2 Juliet	RF	1.00	1.00	1.00	0.08	1.00	1.00	0.98	0.07	0.05	0.05	0.06	0.29	0.41	0.02	0.12	0.09	0.10	0.04	0.30	0.40	0.05
	Resnet	1.00	0.73	0.84	0.13	0.86	0.84	0.80	0.21	0.18	0.20	0.02	0.34	0.38	0.13	0.38	0.68	0.49	0.11	0.44	0.64	0.42
	SVM	1.00	1.00	1.00	0.07	1.00	1.00	0.98	0.17	0.05	0.07	0.05	0.50	0.42	0.10	0.42	0.23	0.29	0.84	0.50	0.42	0.07
3 Anki-Android	RF	0.80	0.89	0.84	0.15	0.84	0.76	0.76	0.85	0.97	0.91	0.10	0.89	0.84	0.85	0.85	0.97	0.91	0.10	0.89	0.84	0.85
	Resnet	0.79	0.85	0.82	0.21	0.82	0.75	0.72	0.88	0.50	0.64	0.08	0.71	0.71	0.62	0.80	0.13	0.23	0.06	0.55	0.57	0.35
	SVM	0.80	0.89	0.84	0.13	0.86	0.78	0.78	0.80	0.93	0.86	0.34	0.83	0.78	0.73	0.82	0.93	0.87	0.63	0.85	0.80	0.69
4 Browser	RF	0.97	0.93	0.95	0.08	1.00	1.00	0.95	0.97	0.94	0.95	0.06	0.96	0.93	0.92	0.89	1.00	0.94	0.06	0.97	0.92	0.92
	Resnet	0.94	0.88	0.91	0.09	0.92	0.87	0.87	0.38	0.97	0.55	0.10	0.56	0.38	0.47	0.83	0.91	0.87	0.02	0.90	0.88	0.85
	SVM	0.91	0.94	0.93	0.10	0.94	0.88	0.88	0.82	0.90	0.86	0.17	0.90	0.78	0.79	0.88	0.72	0.79	0.46	0.90	0.85	0.69
5 Calendar	RF	0.87	0.87	0.89	0.00	0.92	0.82	0.85	0.89	0.86	0.88	0.00	0.88	0.84	0.85	1.00	0.86	0.93	0.00	0.92	0.95	0.92
	Resnet	0.85	1.00	0.92	0.00	0.95	0.85	0.90	0.58	0.97	0.73	0.00	0.67	0.58	0.66	0.88	0.48	0.62	0.00	0.71	0.80	0.65
	SVM	0.88	0.95	0.91	0.00	0.94	0.85	0.89	0.86	0.86	0.86	0.00	0.87	0.81	0.83	0.84	0.72	0.78	0.00	0.88	0.91	0.80
6 Camera	RF	0.94	0.91	0.93	0.08	0.94	0.89	0.89	0.89	0.83	0.86	0.08	0.89	0.80	0.81	0.82	0.75	0.78	0.11	0.95	0.84	0.77
	Resnet	0.91	0.91	0.91	0.10	0.93	0.87	0.87	0.55	1.00	0.71	0.05	0.77	0.55	0.65	0.92	0.46	0.61	0.10	0.72	0.76	0.62
	SVM	0.91	0.94	0.93	0.04	0.94	0.88	0.90	0.82	0.77	0.79	0.06	0.80	0.67	0.72	0.72	0.75	0.73	0.37	0.90	0.82	0.66
7 Connectbot	RF	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.80	0.89	0.04	0.90	0.90	0.87	1.00	0.80	0.89	0.00	0.90	0.90	0.88
	Resnet	1.00	1.00	1.00	0.00	1.00	0.90	0.98	1.00	0.07	0.13	0.12	0.53	0.55	0.33	1.00	0.80	0.89	0.12	0.90	0.90	0.85
	SVM	1.00	1.00	1.00	0.02	1.00	1.00	0.99	1.00	0.80	0.89	0.06	0.90	0.90	0.87	1.00	0.80	0.89	0.12	0.90	0.90	0.85
8 Contacts	RF	0.90	0.96	0.93	0.11	0.96	0.88	0.89	0.83	0.86	0.85	0.06	0.89	0.76	0.80	0.89	1.00	0.94	0.06	0.99	0.97	0.94
	Resnet	0.78	0.90	0.83	0.08	0.89	0.73	0.78	0.56	1.00	0.72	0.15	0.81	0.56	0.65	0.79	0.67	0.73	1.00	0.79	0.79	0.48
	SVM	0.90	0.96	0.93	0.21	0.96	0.88	0.86	0.80	0.78	0.79	0.36	0.87	0.77	0.69	0.85	0.80	0.82	1.00	0.93	0.82	0.58
9 CoolReader	RF	1.00	0.98	0.99	0.09	1.00	1.00	0.97	1.00	1.00	1.00	0.03	1.00	1.00	0.99	1.00	1.00	1.00	0.05	1.00	1.00	0.99
	Resnet	1.00	0.93	0.96	0.12	0.96	0.98	0.93	0.80	0.73	0.76	0.08	0.80	0.69	0.70	0.83	0.91	0.87	0.30	0.89	0.79	0.76
	SVM	0.97	0.97	0.97	0.03	0.96	0.93	0.95	1.00	1.00	1.00	0.11	1.00	1.00	0.97	1.00	1.00	1.00	0.39	1.00	1.00	0.91
10 DeskClock	RF	0.89	1.00	0.94	0.02	0.97	0.89	0.92	0.86	1.00	0.92	0.02	0.93	0.86	0.89	0.90	1.00	0.95	0.02	0.99	0.98	0.95
	Resnet	0.88	0.88	0.88	0.03	0.91	0.80	0.84	0.46	1.00	0.63	0.05	0.50	0.46	0.53	0.35	1.00	0.51	0.01	0.50	0.67	0.54
	SVM	0.89	1.00	0.94	0.02	0.97	0.89	0.92	0.80	1.00	0.89	0.04	0.93	0.86	0.87	0.82	1.00	0.90	0.02	1.00	0.99	0.93
11 Email	RF	0.97	0.98	0.97	0.30	0.99	0.99	0.91	0.98	0.98	0.98	0.00	0.99	1.00	0.98	0.91	0.95	0.93	0.00	0.98	0.97	0.94
	Resnet	0.96	0.90	0.93	0.23	0.93	0.96	0.87	0.74	0.88	0.81	0.33	0.75	0.84	0.69	0.76	0.91	0.83	0.56	0.80	0.86	0.67
	SVM	0.95	0.82	0.88	0.23	0.94	0.95	0.84	0.79	0.84	0.81	0.27	0.88	0.89	0.75	0.80	0.92	0.86	0.27	0.91	0.90	0.79
12 FBReaderJ	RF	0.96	0.93	0.94	0.01	0.98	0.98	0.95	0.96	0.95	0.95	0.01	0.99	0.99	0.96	0.97	0.94	0.95	0.06	0.99	0.99	0.95
	Resnet	0.95	0.96	0.96	0.10	0.97	0.93	0.92	0.96	0.94	0.95	0.00	0.96	0.96	0.94	0.97	0.90	0.93	0.01	0.94	0.95	0.93
	SVM	0.95	0.97	0.96	0.00	0.97	0.93	0.95	0.76	0.72	0.74	0.00	0.91	0.83	0.75	0.84	0.91	0.87	0.05	0.95	0.92	0.87
13 K9Mail	RF	0.97	0.99	0.98	0.00	1.00	1.00	0.99	0.99	1.00	0.99	0.00	1.00	1.00	0.99	0.99	1.00	1.00	0.01	1.00	1.00	0.99
	Resnet	0.94	1.00	0.97	0.00	0.97	0.97	0.96	0.94	0.85	0.89	0.01	0.90	0.94	0.88	0.99	1.00	0.99	0.01	0.99	1.00	0.99
	SVM	0.99	1.00	0.99	0.01	0.99	0.99	0.99	0.83	0.88	0.85	0.00	0.92	0.92	0.86	0.98	0.98	0.98	0.01	0.99	0.99	0.98
14 KeePassAndroid	RF	0.99	1.00	1.00	0.01	1.00	1.00	1.00	1.00	0.99	1.00	0.01	1.00	1.00	0.99	0.98	0.99	0.98	0.01	1.00	1.00	0.99
	Resnet	0.99	0.99	0.99	0.05	0.99	0.98	0.97	0.97	0.84	0.90	0.03	0.91	0.94	0.89	0.99	1.00	0.99	0.08	1.00	0.99	0.97
	SVM	0.99	0.99	0.99	0.02	0.99	0.98	0.98	0.90	0.82	0.86	0.07	0.95	0.92	0.85	0.98	1.00	0.99	0.42	1.00	0.99	0.89
15 MMS	RF	0.98	0.97	0.98	0.22	1.00	0.99	0.93	0.98	0.97	0.98	0.01	0.98	0.96	0.97	0.94	1.00	0.97	0.01	0.98	0.93	0.96
	Resnet	0.98	0.93	0.96	0.44	0.96	0.94	0.84	0.57	0.98	0.72	0.17	0.78	0.56	0.64	0.86	0.95	0.90	0.32	0.94	0.91	0.82
	SVM	0.98	0.97	0.97	0.12	0.97	0.95	0.93	0.98	0.95	0.97	0.08	0.97	0.95	0.94	0.89	0.93	0.91	0.07	0.98	0.95	0.90
16 Xwords	RF	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.97	0.99	0.00	0.99	0.99	0.98	0.99	1.00	0.99	0.00	1.00	0.99	0.99
	Resnet	1.00	1.00	1.00	0.01	1.00	1.00	1.00	0.94	0.91	0.92	0.11	0.93	0.89	0.88	0.88	0.89	0.88	0.09	0.90	0.82	0.83
	SVM	1.00	0.99	0.99	0.15	0.99	0.99	0.96	0.97	0.92	0.95	0.00	0.97	0.95	0.95	0.95	1.00	0.97	0.05	0.98	0.95	0.95
17 QuickSearchBox	RF	0.95	0.87	0.91	0.01	0.93	0.85	0.96	0.77	0.89	0.83	0.07	0.91	0.71	0.85	0.94	0.94	0.94	0.03	0.96	0.90	1.00
	Resnet	1.00	0.78	0.88	0.00	0.89	0.82	0.93	0.58	0.96	0.72	0.18	0.89	0.56	0.72	0.85	0.79	0.82	0.06	0.87	0.74	0.84
	SVM	0.95	0.87	0.91	0.01	0.93	0.85	0.96	0.74	0.63	0.68	0.06	0.79	0.54	0.66	0.86	0.86	0.90	0.06	0.94	0.84	0.92

Table 2: Results from using all source code without the symbols and comment with bag-of-words, word2vec and fastText

		Bag-of-Words							Word2vec							FastText						
Project	Classifier	P	R	F1	FPR	ROC AUC	PR AUC	z	P	R	F3	FPR	ROC AUC	PR AUC	z	P	R	F2	FPR	ROC AUC	PR AUC	z
1 OWASP	RF	0.99	1.00	0.99	0.21	0.99	0.99	0.94	0.71	0.73	0.72	0.21	0.72	0.65	0.60	0.75	0.83	0.79	0.21	0.89	0.89	0.75
	Resnet	0.82	1.00	0.90	0.17	0.89	0.82	0.83	0.79	0.93	0.86	0.19	0.85	0.77	0.77	0.57	0.58	0.58	0.19	0.57	0.68	0.48
	SVM	0.99	0.99	0.99	0.24	0.99	0.99	0.93	0.88	0.90	0.89	0.31	0.89	0.84	0.78	0.60	0.79	0.68	0.19	0.68	0.67	0.58
2 Juliet	RF	0.03	0.02	0.03	0.04	0.26	0.42	0.00	0.12	0.09	0.10	0.04	0.30	0.40	0.05	0.23	0.18	0.20	0.02	0.52	0.36	0.17
	Resnet	0.33	0.05	0.08	0.04	0.35	0.42	0.11	0.22	0.09	0.13	0.03	0.43	0.40	0.12	0.47	0.34	0.37	0.07	0.54	0.54	0.34
	SVM	0.41	0.02	0.03	0.04	0.68	0.54	0.21	0.20	0.09	0.13	0.03	0.47	0.42	0.13	0.42	0.16	0.23	0.02	0.62	0.46	0.27
3 Anki-Android	RF	0.81	0.96	0.88	0.08	0.88	0.80	0.83	0.78	0.93	0.85	0.08	0.84	0.76	0.78	0.84	0.96	0.90	0.08	0.90	0.83	0.85
	Resnet	0.82	1.00	0.90	0.11	0.90	0.82	0.84	0.78	0.93	0.85	0.05	0.84	0.76	0.79	0.82	0.52	0.64	0.00	0.71	0.66	0.61
	SVM	0.81	0.96	0.88	0.16	0.90	0.82	0.82	0.80	0.89	0.84	0.13	0.84	0.76	0.77	0.77	0.85	0.81	0.09	0.65	0.57	0.66
4 Browser	RF	0.94	0.91	0.93	0.04	0.94	0.89	0.90	0.94	0.94	0.94	0.04	0.95	0.90	0.91	0.93	0.84	0.89	0.04	0.96	0.87	0.87
	Resnet	0.88	0.85	0.87	0.03	0.89	0.81	0.83	0.88	0.91	0.90	0.03	0.92	0.84	0.86	0.81	0.91	0.86	0.07	0.89	0.77	0.81
	SVM	0.91	0.91	0.91	0.05	0.94	0.88	0.89	0.89	1.00	0.94	0.07	0.96	0.89	0.91	0.90	0.82	0.86	0.06	0.88	0.80	0.81
5 Calendar	RF	0.88	0.95	0.91	0.00	0.94	0.85	0.89	0.73	0.70	0.71	0.00	0.77	0.62	0.65	0.81	0.74	0.77	0.00	0.82	0.70	0.73
	Resnet	0.78	0.95	0.86	0.00	0.90	0.76	0.82	0.76	0.70	0.73	0.00	0.78	0.64	0.68	0.78	0.89	0.83	0.00	0.84	0.75	0.79
	SVM	0.88	0.95	0.91	0.00	0.94	0.85	0.89	0.71	0.74	0.72	0.00	0.78	0.62	0.67	0.77	0.74	0.76	0.00	0.80	0.67	0.71
6 Camera	RF	0.87	0.91	0.93	0.07	0.94	0.89	0.88	0.87	0.83	0.85	0.05	0.89	0.77	0.81	0.92	0.88	0.90	0.05	0.97	0.94	0.90
	Resnet	0.88	0.88	0.88	0.92	0.90	0.83	0.64	0.78	0.88	0.82	0.05	0.89	0.72	0.77	0.81	0.85	0.83	0.07	0.88	0.85	0.80
	SVM	0.91	0.91	0.91	0.04	0.94	0.88	0.89	0.88	0.92	0.90	0.07	0.93	0.93	0.88	0.81	0.81	0.81	0.08	0.89	0.74	0.76
7 Connectbot	RF	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.82	0.90	0.00	0.91	0.93	0.90
	SVM	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00
8 Contacts	RF	0.85	0.96	0.90	0.06	0.98	0.95	0.90	0.93	0.91	0.92	0.05	0.94	0.88	0.89	0.96	0.89	0.92	0.06	0.93	0.89	0.89
	Resnet	0.83	0.92	0.87	0.08	0.92	0.89	0.85	0.90	0.67	0.77	0.15	0.81	0.72	0.70	0.81	0.87	0.84	0.06	0.88	0.74	0.78
	SVM	0.80	0.67	0.73	0.07	0.90	0.84	0.73	0.86	0.77	0.81	0.14	0.85	0.76	0.75	0.85	0.83	0.84	0.14	0.74	0.62	0.70
9 CoolReader	RF	1.00	0.97	0.99	0.04	0.99	0.99	0.97	1.00	1.00	1.00	0.01	1.00	1.00	1.00	1.00	0.98	0.99	0.04	0.99	0.99	0.98
	Resnet	1.00	0.97	0.99	0.04	0.99	0.99	0.97	0.98	1.00	0.99	0.01	0.99	0.98	0.98	1.00	1.00	1.00	0.10	1.00	1.00	0.98
	SVM	0.97	0.97	0.97	0.09	0.96	0.93	0.94	0.98	1.00	0.99	0.00	0.99	0.98	0.98	1.00	0.98	0.99	0.03	0.99	0.99	0.98
10 DeskClock	RF	0.89	1.00	0.94	0.02	0.97	0.89	0.92	0.91	0.83	0.87	0.02	0.88	0.83	0.84	0.92	0.79	0.85	0.02	0.93	0.87	0.84
	Resnet	0.98	1.00	0.89	0.01	0.94	0.80	0.91	0.75	0.75	0.75	0.01	0.77	0.68	0.69	0.50	0.07	0.13	0.01	0.49	0.54	0.23
	SVM	0.89	1.00	0.94	0.01	0.97	0.89	0.93	0.83	0.83	0.83	0.02	0.85	0.77	0.79	0.80	0.57	0.67	0.02	0.86	0.88	0.71
11 Email	RF	0.97	0.98	0.97	0.50	1.00	1.00	0.86	0.93	0.99	0.96	0.00	0.98	0.97	0.96	0.97	0.98	0.97	0.00	0.97	0.96	0.97
	Resnet	0.93	1.00	0.96	0.41	0.95	0.97	0.86	0.97	0.75	0.85	0.47	0.86	0.87	0.72	0.91	0.99	0.95	0.45	0.93	0.91	0.82
	SVM	0.96	0.85	0.90	0.50	0.96	0.97	0.80	0.97	0.98	0.97	0.45	0.97	0.96	0.86	0.94	0.95	0.94	0.45	0.93	0.92	0.82
12 FBReaderJ	RF	0.96	0.96	0.97	0.01	0.98	0.95	0.96	0.98	0.95	0.97	0.02	0.97	0.95	0.95	0.97	0.95	0.96	0.02	1.00	0.99	0.96
	Resnet	0.96	0.96	0.96	0.01	0.97	0.93	0.95	0.95	0.90	0.93	0.00	0.94	0.89	0.91	0.92	0.82	0.87	0.01	0.90	0.90	0.86
	SVM	0.95	0.97	0.96	0.02	0.97	0.93	0.94	0.93	0.85	0.89	0.01	0.91	0.83	0.86	0.75	0.50	0.60	0.01	0.86	0.69	0.62
13 K9Mail	RF	0.99	1.00	0.99	0.00	0.99	0.99	0.99	0.98	0.99	0.98	0.00	1.00	1.00	0.99	0.99	0.99	0.99	0.00	1.00	1.00	0.99
	Resnet	0.99	0.99	0.99	0.00	0.99	0.99	0.99	1.00	0.94	0.97	0.00	0.97	0.97	0.96	0.92	0.91	0.91	0.01	0.91	0.94	0.90
	SVM	0.99	1.00	1.00	0.01	0.99	0.99	0.99	0.98	1.00	0.99	0.00	0.98	0.97	0.98	0.77	0.88	0.82	0.00	0.85	0.80	0.79
14 KeePassAndroid	RF	1.00	1.00	1.00	0.01	1.00	1.00	1.00	0.99	1.00	1.00	0.01	1.00	0.99	0.99	0.99	1.00	1.00	0.01	1.00	1.00	1.00
	Resnet	1.00	1.00	1.00	0.03	1.00	1.00	0.99	0.99	1.00	1.00	0.03	1.00	0.99	0.99	1.00	0.99	0.99	0.03	0.99	1.00	0.98
	SVM	0.98	0.97	0.97	0.03	1.00	1.00	0.97	0.99	1.00	1.00	0.03	1.00	0.99	0.99	0.83	0.86	0.84	0.01	0.92	0.84	0.83
15 MMS	RF	0.98	0.97	0.97	0.01	0.98	0.96	0.96	0.96	0.98	0.97	0.01	0.98	0.95	0.96	0.96	0.98	0.97	0.01	0.98	0.95	0.96
	Resnet	0.98	0.91	0.94	0.28	0.95	0.96	0.87	0.96	0.67	0.79	0.00	0.82	0.76	0.76	0.91	0.95	0.93	0.00	0.95	0.89	0.92
	SVM	0.98	0.97	0.97	0.12	0.98	0.96	0.94	0.96	0.98	0.97	0.04	0.97	0.94	0.95	0.96	0.98	0.97	0.09	0.98	0.95	0.94
16 Xwords	RF	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.99	1.00	0.99	0.00	0.99	0.99	0.99	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	1.00	0.99	0.99	0.00	0.99	0.99	0.99	0.86	1.00	0.92	0.00	0.92	0.86	0.90	0.95	0.25	0.40	0.01	0.62	0.62	0.49
	SVM	1.00	0.99	0.99	0.01	0.99	0.99	0.99	1.00	0.99	0.99	0.00	0.99	0.98	0.99	1.00	1.00	1.00	0.00	1.00	1.00	1.00
17 QuickSearchBox	RF	0.95	0.87	0.91	0.01	0.93	0.85	0.96	0.88	0.81	0.85	0.03	0.89	0.76	0.87	0.93	0.95	0.94	0.03	0.96	0.90	1.00
	Resnet	0.95	0.91	0.93	0.01	0.95	0.89	0.99	0.77	0.89	0.83	0.07	0.91	0.71	0.85	0.86	0.99	0.92	0.07	0.96	0.86	0.97
	SVM	0.95	0.87	0.91	0.01	0.93	0.85	0.96	0.89	0.89	0.89	0.03	0.90	0.82	0.92	0.85	0.88	0.86	0.07	0.89	0.77	0.88

Table 3: Architectural metrics

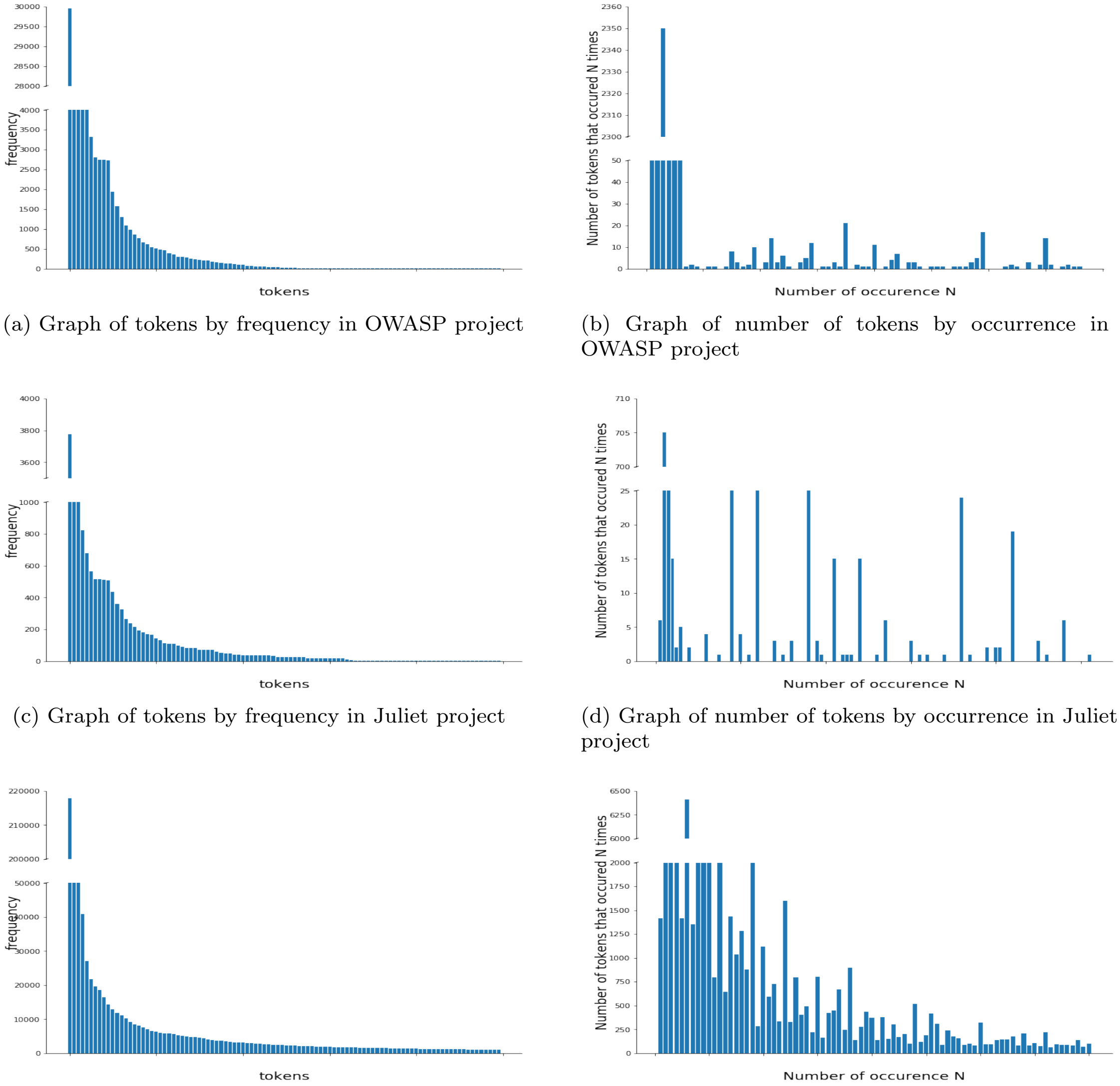
		Metrics only							Metrics + bag-of-words						
Project	Classifier	P	R	F1	FPR	ROC AUC	PR AUC	z	P	R	F1	FPR	ROC AUC	PR AUC	z
1 OWASP	RF	0.66	0.78	0.71	0.38	0.70	0.62	0.56	0.82	0.85	0.84	0.17	0.95	0.96	0.73
	Resnet	0.48	1.00	0.65	1.00	0.50	0.48	0.32	0.70	0.89	0.79	0.34	0.77	0.82	0.51
	SVM	0.57	0.93	0.70	0.66	0.64	0.56	0.48	0.67	0.74	0.70	0.74	0.82	0.85	0.30
2 Juliet	RF	0.50	0.41	0.45	0.23	0.59	0.42	0.33	1.00	0.88	0.93	0.00	0.94	0.92	0.88
	Resnet	0.35	0.97	0.52	1.00	0.48	0.35	0.21	1.00	0.81	0.90	0.00	0.91	0.88	0.82
	SVM	0.00	0.00	0.00	0.00	0.50	0.36	0.01	1.00	0.84	0.92	0.00	0.94	0.92	0.86
3 Anki-Android	RF	0.62	0.73	0.67	0.26	0.74	0.55	0.55	0.87	0.91	0.89	0.08	0.92	0.82	0.76
	Resnet	0.36	1.00	0.53	1.00	0.50	0.36	0.23	0.83	0.86	0.84	0.10	0.88	0.76	0.67
	SVM	0.71	0.23	0.34	0.05	0.50	0.36	0.32	0.88	0.95	0.91	0.08	0.94	0.85	0.81
4 Browser	RF	0.72	0.62	0.67	0.16	0.73	0.60	0.59	0.94	0.91	0.93	0.04	0.94	0.89	0.84
	Resnet	0.00	0.00	0.00	0.00	0.50	0.40	0.02	0.91	0.91	0.91	0.06	0.93	0.87	0.81
	SVM	0.00	0.00	0.00	0.00	0.50	0.40	0.02	0.89	0.94	0.93	0.06	0.94	0.88	0.83
5 Calendar	RF	1.00	0.94	0.97	0.00	0.97	0.98	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	0.90	0.53	0.67	0.08	0.72	0.75	0.66	0.75	0.88	0.81	0.42	0.73	0.85	0.50
	SVM	0.90	0.53	0.67	0.08	0.72	0.75	0.66	1.00	0.88	0.94	0.00	1.00	1.00	0.94
6 Camera	RF	0.57	0.60	0.59	0.20	0.70	0.46	0.47	0.90	0.96	0.93	0.05	0.96	0.88	0.85
	Resnet	0.39	0.56	0.46	0.39	0.59	0.35	0.28	0.84	0.88	0.86	0.07	0.90	0.77	0.71
	SVM	0.00	0.00	0.00	0.00	0.50	0.30	0.00	0.90	0.96	0.93	0.05	0.96	0.88	0.85
7 Connectbot	RF	0.80	0.76	0.78	0.15	0.80	0.71	0.71	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	0.45	0.46	0.45	0.46	0.50	0.45	0.26	1.00	0.97	0.99	0.00	0.99	0.99	0.98
	SVM	0.65	0.54	0.61	0.47	0.43	0.04	0.25	0.92	0.93	0.93	0.06	0.98	0.97	0.88
8 Contacts	RF	0.89	1.00	0.94	0.06	0.97	0.89	0.95	0.89	1.00	0.94	0.06	0.89	0.92	0.85
	Resnet	0.31	1.00	0.47	1.00	0.50	0.31	0.19	0.80	1.00	0.89	0.11	0.94	0.80	0.76
	SVM	0.54	0.88	0.67	0.33	0.77	0.51	0.55	0.89	1.00	0.94	0.06	0.89	0.92	0.85
9 CoolReader	RF	0.83	0.86	0.84	0.20	0.83	0.79	0.77	0.99	0.96	0.97	0.01	0.97	0.97	0.94
	Resnet	0.61	0.84	0.71	0.62	0.61	0.60	0.48	0.98	0.96	0.97	0.03	0.97	0.96	0.93
	SVM	0.63	0.77	0.69	0.52	0.62	0.61	0.49	0.98	0.96	0.97	0.03	0.97	0.96	0.93
10 DeskClock	RF	0.83	0.68	0.75	0.06	0.81	0.67	0.71	0.98	0.96	0.97	0.01	0.97	0.95	0.94
	Resnet	0.46	0.53	0.49	0.29	0.62	0.39	0.35	0.98	0.91	0.94	0.01	0.95	0.92	0.89
	SVM	0.00	0.00	0.00	0.00	0.50	0.32	0.00	0.97	0.97	0.97	0.01	0.97	0.94	0.93
11 Email	RF	0.00	0.00	0.00	0.00	0.50	0.42	0.03	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	0.41	1.00	0.60	1.00	0.50	0.42	0.28	1.00	0.98	0.99	0.00	0.99	0.99	0.98
	SVM	0.00	0.00	0.00	0.00	0.50	0.42	0.03	1.00	1.00	1.00	0.00	1.00	1.00	1.00
12 FBReaderJ	RF	0.80	0.82	0.81	0.20	0.81	0.74	0.73	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	0.47	0.38	0.42	0.41	0.48	0.48	0.25	1.00	0.98	0.99	0.00	0.99	0.99	0.98
	SVM	0.64	0.86	0.74	0.47	0.70	0.62	0.56	0.99	1.00	0.99	0.01	0.99	0.99	0.98
13 K9Mail	RF	0.90	0.83	0.86	0.07	0.88	0.82	0.84	0.99	0.99	0.99	0.01	0.99	0.98	0.98
	Resnet	0.71	0.27	0.39	0.08	0.59	0.51	0.39	0.46	1.00	0.63	0.90	0.55	0.46	0.00
	SVM	0.00	0.00	0.00	0.00	0.50	0.43	0.03	0.99	0.99	0.99	0.01	0.99	0.98	0.98
14 KeePassAndroid	RF	0.86	0.83	0.84	0.07	0.88	0.77	0.81	0.98	0.97	0.97	0.01	0.98	0.96	0.95
	Resnet	0.43	0.71	0.54	0.45	0.63	0.40	0.36	0.95	0.95	0.96	0.01	0.97	0.95	0.92
	SVM	0.57	0.55	0.56	0.20	0.68	0.68	0.50	0.98	0.97	0.97	0.01	0.97	0.95	0.94
15 MMS	RF	0.52	0.89	0.65	0.82	0.54	0.51	0.37	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	0.51	0.49	0.50	0.46	0.51	0.50	0.31	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	SVM	0.50	1.00	0.66	1.00	0.50	0.50	0.33	0.99	0.99	0.99	0.01	0.99	0.99	0.98
16 Xwords	RF	0.86	0.87	0.87	0.16	0.86	0.82	0.82	1.00	1.00	1.00	0.00	1.00	1.00	1.00
	Resnet	0.61	0.92	0.74	0.65	0.64	0.61	0.51	0.98	0.95	0.96	0.03	0.96	0.98	0.93
	SVM	0.75	0.67	0.70	0.25	0.71	0.68	0.60	1.00	0.93	0.96	0.00	0.99	0.99	0.95
17 QuickSearchBox	RF	0.65	0.74	0.69	0.08	0.83	0.53	0.67	0.95	0.87	0.91	0.01	0.93	0.85	0.96
	Resnet	0.50	0.04	0.08	0.01	0.52	0.19	0.16	0.95	0.87	0.91	0.01	0.93	0.85	0.96
	SVM	0.00	0.00	0.00	0.00	0.50	0.18	0.00	0.95	0.87	0.91	0.01	0.93	0.85	0.96

2.2 Token distributions

The graphs 1a, 1c and 1e represent the tokens by their frequencies. The graphs 1b, 1d and 1f represents the number of tokens that occurs N time in each project. 1a, 1c and 1e represent the tokens by their frequency. The graphs 1b, 1d and 1f represents the number of tokens that occurs N time in each project.

These graphs show that most tokens do not have a high frequency. Even though some of the most frequent tokens can provide some information about the vulnerabilities in the source code,most of them are neutral words such as "webservlet", "encodeforhtml". Even though these are the actual high-frequency words it is difficult to say that these words are all important words in detecting vulnerable files. So, we can assume that the signatures are learnt from the tokens that are less frequent.

- Graph 1a represents the tokens frequency in the OWASP benchmark java files shows that there is less tokens that have a high frequency than tokens that have a low one.
- Graph 1c represents the tokens frequency in the Juliet project java files shows that there is less tokens that have a high frequency than tokens that have a low one.
- Graph 1e represents the token frequency in all the android applications. This graph shows that there is less tokens that have a high frequency than tokens that have a low one.
- Graph 1b represents the token frequency in all the android applications. Like the other graphs this one shows that there is more tokens with lower number of occurrences.
- Graph 1d represents the token frequency in all the android applications. Although this graph is different than the others due to some tokens with high occurrences that occur more than the others. We can still see that the lower the number of occurrences the higher the number of tokens.
- Graph 1f represents the token frequency in all the android applications. This graph shows that there is more tokens with lower number of occurrences



(a) Graph of tokens by frequency in OWASP project

(b) Graph of number of tokens by occurrence in OWASP project

(c) Graph of tokens by frequency in Juliet project

(d) Graph of number of tokens by occurrence in Juliet project

(e) Graph of tokens by frequency in all the android projects

(f) Graph of number of tokens by occurrence in all the android projects

Figure 1: Graph representation of the tokens occurrence in the different source code projects

References

[1] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". In: *Mach. Learn.* 20.3 (Sept. 1995), pp. 273–297. ISSN: 0885-6125. DOI: 10.1023/A:1022627411411. URL: https://doi.org/10.1023/A:1022627411411.

[2] Micro Focus. *Fortify Static Code Analyzer*. 2018. URL: https://www.microfocus.com/en-us/products/static-code-analysis-sast.

[3] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[4] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *CoRR* abs/1607.01759 (2016). arXiv: 1607.01759. URL: http://arxiv.org/abs/1607.01759.

[5] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint arXiv:1301.3781* (2013).

[6] OWASP. *OWASP Benchmark Project*. 2018. URL: https://www.owasp.org/index.php/Benchmark.

[7] R. Scandariato et al. "Predicting Vulnerable Software Components via Text Mining". In: *IEEE Transactions on Software Engineering* 40.10 (Oct. 2014), pp. 993–1006. ISSN: 2326-3881. DOI: 10.1109/TSE.2014.2340398.

[8] National Institute of Standards and Technology. *Juliet Test Suite for Java v1.3*. 2017. URL: https://samate.nist.gov/SRD/testsuite.php.

[9] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. 1995, 278–282 vol.1.

[10] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: A statistical framework". In: *International Journal of Machine Learning and Cybernetics* 1 (Dec. 2010), pp. 43–52. DOI: 10.1007/s13042-010-0001-0.