Table 1: Overview of the studied datasets.

Domain	Dataset	#Files	#Features	Defective (%)	EPV	Domain	Dataset	#Files	#Features	Defective (%)	EPV
NASA	CM1	327	37	12.8%	1.135	Kim	Apache	194	26	50.5%	3.769
	MC1	1,988	38	2.3%	1.21		Eclipse34_debug ∗ †	1,065	17	24.7%	15.47
	MC2	125	39	35.2%	1.128		Eclipse34_swt ∗ †	1,485	17	44%	38.41
	MW1	253	37	10.7%	0.729		Safe	56	26	39.2%	0.846
	JM1 ∗ †	7,782	21	21.5%	79.62		Zxing	399	26	29.6%	4.538
	PC1	705	37	8.7%	1.648	Ambros Zimmermann	Equinox	324	15	39.8%	8.600
	PC2	745	36	2.1%	0.444		Jdt ∗ †	997	15	20.7%	13.73
	PC3	1,077	37	12.4%	3.621		Lucene	691	15	9.3%	4.266
	PC4	1,287	37	13.8%	4.784		Mylyn ∗ †	1,862	15	13.2%	16.33
	PC5 ∗ †	1,711	38	27.5%	12.39		Pde ∗ †	1,497	15	14%	13.93
	a1	121	27	7.4%	0.331	PROMISE	Eclipse-2.0 ∗ †	6,729	32	14.5%	30.47
	a3	63	29	12.6%	0.273		Eclipse-2.1 ∗ †	7,888	32	10.8%	26.68
	a4	107	29	18.7%	0.689		Eclipse-3.0 ∗ †	10,593	32	14.8%	49.00
	a5	36	29	22.2%	0.275		Pdftranslator	33	20	45.5%	0.750
	a6	101	29	14.9%	0.518		Poi-1.5	237	20	59.5%	7.050
	kc2	522	21	20.5%	5.095		Poi-2.0	314	20	11.8%	1.850
	kc3	194	39	18.6%	0.923		Poi-2.5 †	385	20	64.4%	12.40
PROMISE	Ant-1.3	126	20	16%	1.000		Poi-3.0 †	442	20	63.6%	14.05
	Ant-1.4	178	20	22.5%	2.000		Prop-1 ∗ †	18,471	20	14.8%	136.9
	Ant-1.5	293	20	10.9%	1.600		Prop-2 ∗ †	23,014	20	10.6%	121.6
	Ant-1.6	351	20	26.2%	4.600		Prop-3 ∗ †	10,274	20	11.5%	59.00
	Ant-1.7	745	20	22.2%	8.300		Prop-4 ∗ †	8,718	20	9.6%	42.00
	Arc	234	20	11.5%	1.350		Prop-5 ∗ †	8,516	20	15.3%	64.95
	Berek	43	20	37.2%	0.800		Prop-6	660	20	9.1%	3.300
	Camel-1.0	339	20	3.8%	0.650		Redaktor	176	20	15.3%	1.350
	Camel-1.2 ∗ †	608	20	35.5%	10.80		Serapion	45	20	20%	0.450
	Camel-1.4	872	20	16.6%	7.250		Skarbonka	45	20	20%	0.450
	Camel-1.6	965	20	19.5%	9.400		Sklebagd	20	20	60%	0.600
	Ckjm	10	20	50%	0.250		Synapse-1.0	157	20	10.2%	0.800
	E-learning	64	20	7.8%	0.250		Synapse-1.1	222	20	27%	3.000
	Forrest-0.6	6	20	16.7%	0.050		Synapse-1.2	256	20	33.6%	4.300
	Forrest-0.7	29	20	17.2%	0.250		Systemdata	65	20	13.8%	0.450
	Forrest-0.8	32	20	6.3%	0.100		Szybkafucha	25	20	56%	0.700
	Intercafe	27	20	14.8%	0.200		Termoproject	42	20	30.1%	0.650
	Ivy-1.1	111	20	56.8%	3.150		Tomcat	858	20	9%	3.850
	Ivy-1.4	241	20	6.6%	0.800		Velocity-1.4	194	20	75%	7.350
	Ivy-2.0	352	20	11.4%	2.000		Velocity-1.5	214	20	66.4%	7.100
	Jedit-3.2	272	20	33.1%	4.500		Velocity-1.6	229	20	34.1%	3.900
	Jedit-4.0	306	20	24.5%	3.750		Workflow	39	20	51.3%	1.000
	Jedit-4.1	312	20	25.3%	3.950		Wspomaganiepi	18	20	66.7%	0.600
	Jedit-4.2	367	20	13.1%	2.400		Xalan-2.4	723	20	15.2%	5.500
	Jedit-4.3	492	20	2.2%	0.550		Xalan-2.5 ∗ †	803	20	48.2%	19.35
	Kalkulator	27	20	22.2%	0.300		Xalan-2.6 ∗ †	885	20	46.4%	20.55
	Log4j-1.0	135	20	25.2%	1.700		Xalan-2.7 †	909	20	98.8%	44.90
	Log4j-1.1	109	20	33.9%	1.850		Xerces-1.2	440	20	16.1%	3.550
	Log4j-1.2	189	20	92.2%	9.450		Xerces-1.3	453	20	15.2%	3.450
	Lucene-2.0	195	20	46.7%	4.550		Xerces-1.4 †	588	20	74.3%	21.85
	Lucene-2.2	247	20	58.3%	7.200		Xerces-init	162	20	47.5%	3.850
	Lucene-2.4 †	340	20	59.7%	10.15		Zuzel	29	20	44.8%	0.650
	Nieruchomosci	27	20	37%	0.500		PBeans2	51	20	19.6%	0.500
	PBeans1	26	20	76.9%	1.000						

- In total, there are 101 datasets included in our study.
- [∗] Marks the 18 datasets stuided by prior research Tantithamthavorn et al [2016,2018].
- [†] Marks the 23 `High EPV` datasets included in our study. The remaining 78 datasets are included as `Low EPV` datasets.