APPENDIX

```
DeepInfer:{
    "condition": ">=",
    "prediction_interval": 0.95
}
SelfChecker:{
    "var_threshold": 1e-5,
    "only_activation_layers": true,
    "batch_size": 128
}
```

Listing 1: Configuration of each tool in TrustDNN, for the replication experiments.

```
Prophecy: {
    "only_activation_layers": true,
    "only_dense_layers": true,
    "random_state": 42,
    "skip_rules": true
}
DeepInfer:{
    "condition": ">=",
    "prediction_interval": 0.95
}
SelfChecker:{
    "var_threshold": 1e-5,
    "only_activation_layers": true,
    "only_dense_layers": true,
    "batch_size": 128
}
```

Listing 2: Configuration of each tool in TrustDNN, for comparison experiments (using both only_dense_layers and only_activation_layers flags selects dense layers coupled with activation functions.)

Table 5: Effectiveness results by model for Prophecy.

Model	Notifications			Confusion Matrix				Metrics				
	#C	*#I*	*#U*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*Prec.*	*F1*	*MCC*
BM1	944	114	0	72	42	784	160	31.03%	5.08%	63.16%	41.62%	0.346
BM2	940	118	0	75	43	788	152	33.04%	5.17%	63.56%	43.48%	0.363
BM3	707	60	291	253	98	597	110	69.7%	14.1%	72.08%	70.87%	0.561
BM4	847	92	119	151	60	695	152	49.83%	7.95%	71.56%	58.75%	0.474
BM5	936	122	0	81	41	775	161	33.47%	5.02%	66.39%	44.51%	0.374
BM6	921	137	0	85	52	769	152	35.86%	6.33%	62.04%	45.45%	0.367
BM7	923	135	0	88	47	773	150	36.97%	5.73%	65.19%	47.18%	0.391
BM8	897	93	68	94	67	774	123	43.32%	7.97%	58.39%	49.74%	0.397
BM9	925	133	0	78	55	779	146	34.82%	6.59%	58.65%	43.7%	0.348
BM10	383	675	0	221	454	299	84	72.46%	60.29%	32.74%	45.1%	0.115
BM11	380	678	0	229	449	291	89	72.01%	60.68%	33.78%	45.98%	0.108
BM12	863	82	113	130	65	739	124	51.18%	8.08%	66.67%	57.91%	0.475
CIFAR10	9723	101	176	91	186	7954	1769	4.89%	2.29%	32.85%	8.52%	0.062
GC1	43	9	48	35	22	32	11	76.09%	40.74%	61.4%	67.96%	0.356
GC2	51	11	38	27	22	40	11	71.05%	35.48%	55.1%	62.07%	0.345
GC3	63	9	28	23	14	44	19	54.76%	24.14%	62.16%	58.23%	0.313
GC4	74	26	0	14	12	53	21	40.0%	18.46%	53.85%	45.9%	0.234
GC5	65	23	12	17	18	50	15	53.12%	26.47%	48.57%	50.75%	0.261
GC6	78	22	0	10	12	57	21	32.26%	17.39%	45.45%	37.74%	0.166
GC7	75	25	0	13	12	54	21	38.24%	18.18%	52.0%	44.07%	0.219
GC8	85	15	0	9	6	58	27	25.0%	9.38%	60.0%	35.29%	0.21
GC9	64	11	25	20	16	47	17	54.05%	25.4%	55.56%	54.79%	0.288
HP1	134	12	0	6	6	118	16	27.27%	4.84%	50.0%	35.29%	0.292
HP2	133	13	0	7	6	115	18	28.0%	4.96%	53.85%	36.84%	0.305
HP3	65	81	0	19	62	55	10	65.52%	52.99%	23.46%	34.55%	0.101
HP4	123	8	15	16	7	112	11	59.26%	5.88%	69.57%	64.0%	0.569
PD1	64	13	0	11	2	50	14	44.0%	3.85%	84.62%	57.89%	0.502
PD2	63	14	0	5	9	40	23	17.86%	18.37%	35.71%	23.81%	-0.006
PD3	67	10	0	2	8	44	23	8.0%	15.38%	20.0%	11.43%	-0.103
PD4	58	9	10	15	4	47	11	57.69%	7.84%	78.95%	66.67%	0.547

Table 6: Effectiveness of Prophecy on CIFAR10 with balanced train data.

Model	Confusion Matrix				Metrics				
	TP	*FP*	*TN*	*FN*	*TPR*	*FPR*	*Precision*	*F1*	*MCC*
CIFAR10	1511	1423	6534	532	73.96%	17.88%	51.5%	60.72%	0.497

Table 7: Effectiveness results by model for DeepInfer.

Model	Notifications			Confusion Matrix				Metrics				
	#C	*#I*	*#U*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*Prec.*	*F1*	*MCC*
BM1	310	748	0	597	151	259	51	92.13%	36.83%	79.81%	85.53%	0.592
BM2	753	305	0	246	59	617	136	64.4%	8.73%	80.66%	71.62%	0.59
BM3	360	698	0	560	138	290	70	88.89%	32.24%	80.23%	84.34%	0.587
BM4	543	515	0	422	93	424	119	78.0%	17.99%	81.94%	79.92%	0.6
BM5	506	552	0	427	125	429	77	84.72%	22.56%	77.36%	80.87%	0.621
BM6	301	757	0	611	146	243	58	91.33%	37.53%	80.71%	85.69%	0.575
BM7	699	359	0	288	71	573	126	69.57%	11.02%	80.22%	74.51%	0.603
BM8	540	518	0	428	90	440	100	81.06%	16.98%	82.63%	81.84%	0.641
BM9	1058	0	0	0	0	857	201	0.0%	0.0%	0.0%	0.0%	0.0
BM10	549	509	0	260	249	260	289	47.36%	48.92%	51.08%	49.15%	-0.016
BM11	445	613	0	321	292	199	246	56.61%	59.47%	52.37%	54.41%	-0.029
BM12	535	523	0	424	99	445	90	82.49%	18.2%	81.07%	81.77%	0.643
CIFAR10	4570	5430	0	3938	1492	4107	463	89.48%	26.65%	72.52%	80.11%	0.626
GC1	52	48	0	30	18	37	15	66.67%	32.73%	62.5%	64.52%	0.338
GC2	21	79	0	53	26	14	7	88.33%	65.0%	67.09%	76.26%	0.281
GC3	74	26	0	16	10	51	23	41.03%	16.39%	61.54%	49.23%	0.274
GC4	56	44	0	38	6	29	27	58.46%	17.14%	86.36%	69.72%	0.397
GC5	68	32	0	23	9	44	24	48.94%	16.98%	71.88%	58.23%	0.342
GC6	52	48	0	40	8	27	25	61.54%	22.86%	83.33%	70.8%	0.369
GC7	29	71	0	50	21	17	12	80.65%	55.26%	70.42%	75.19%	0.272
GC8	64	36	0	33	3	34	30	52.38%	8.11%	91.67%	66.67%	0.445
GC9	41	59	0	45	14	22	19	70.31%	38.89%	76.27%	73.17%	0.307
HP1	77	68	1	64	5	60	17	79.01%	7.69%	92.75%	85.33%	0.71
HP2	82	64	0	56	8	66	16	77.78%	10.81%	87.5%	82.35%	0.675
HP3	146	0	0	0	0	74	72	0.0%	0.0%	0.0%	0.0%	0.0
HP4	78	68	0	60	8	68	10	85.71%	10.53%	88.24%	86.96%	0.753
PD1	53	22	2	21	3	40	13	61.76%	6.98%	87.5%	72.41%	0.587
PD2	77	0	0	0	0	45	32	0.0%	0.0%	0.0%	0.0%	0.0
PD3	41	36	0	19	17	27	14	57.58%	38.64%	52.78%	55.07%	0.188
PD4	20	57	0	45	12	17	3	93.75%	41.38%	78.95%	85.71%	0.579

Table 8: Effectiveness results by model for SelfChecker.

Model	Notifications			Confusion Matrix				Metrics				
	#C	*#I*	*#U*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*	*Prec.*	*F1*	*MCC*
BM1	-	-	-	538	444	76	0	100%	85.38%	54.79%	70.79%	0.283
BM2	-	-	-	536	457	63	2	99.63%	87.88%	53.98%	70.02%	0.244
BM3	-	-	-	534	397	123	4	99.26%	76.35%	57.36%	72.7%	0.352
BM4	-	-	-	535	424	96	3	99.44%	81.54%	55.79%	71.48%	0.307
BM5	-	-	-	537	454	66	1	99.81%	87.31%	54.19%	70.24%	0.257
BM6	-	-	-	535	450	70	3	99.44%	86.54%	54.31%	70.26%	0.255
BM7	-	-	-	531	372	148	7	98.7%	71.54%	58.8%	73.7%	0.384
BM8	-	-	-	535	420	100	3	99.44%	80.77%	56.02%	71.67%	0.315
BM9	-	-	-	534	373	147	4	99.26%	71.73%	58.88%	73.91%	0.393
BM10	-	-	-	248	28	492	290	46.1%	5.38%	89.86%	60.93%	0.464
BM11	-	-	-	538	520	0	0	100.0%	100.0%	50.85%	67.42%	0.0
BM12	-	-	-	534	397	123	4	99.26%	76.35%	57.36%	72.7%	0.352
CIFAR10	-	-	-	1244	207	7838	711	63.63%	2.57%	85.73%	73.05%	0.688
GC1	-	-	-	0	10	57	33	0.0%	14.93%	0.0%	0.0%	-0.234
GC2	-	-	-	19	17	50	14	57.58%	25.37%	52.78%	55.07%	0.315
GC3	-	-	-	22	34	33	11	66.67%	50.75%	39.29%	49.44%	0.151
GC4	-	-	-	17	31	36	16	51.52%	46.27%	35.42%	41.98%	0.049
GC5	-	-	-	20	17	50	13	60.61%	25.37%	54.05%	57.14%	0.343
GC6	-	-	-	12	12	55	21	36.36%	17.91%	50.0%	42.11%	0.203
GC7	-	-	-	13	14	53	20	39.39%	20.9%	48.15%	43.33%	0.196
GC8	-	-	-	13	11	56	20	39.39%	16.42%	54.17%	45.61%	0.253
GC9	-	-	-	17	12	55	16	51.52%	17.91%	58.62%	54.84%	0.348
HP1	-	-	-	72	66	8	0	100.0%	89.19%	52.17%	68.57%	0.237
HP2	-	-	-	60	9	65	12	83.33%	12.16%	86.96%	85.11%	0.713
HP3	-	-	-	58	8	66	14	80.56%	10.81%	87.88%	84.06%	0.701
HP4	-	-	-	62	10	64	10	86.11%	13.51%	86.11%	86.11%	0.726
PD1	-	-	-	26	8	37	6	81.25%	17.78%	76.47%	78.79%	0.63
PD2	-	-	-	11	5	40	21	34.38%	11.11%	68.75%	45.83%	0.283
PD3	-	-	-	29	28	17	3	90.62%	62.22%	50.88%	65.17%	0.319
PD4	-	-	-	24	6	39	8	75.0%	13.33%	80.0%	77.42%	0.623

Table 9: Time (average duration in seconds) and Memory (peak memory usage in Mebibytes) efficiency for each approach by dataset.

Tool	Phase	Dataset	Duration	Memory
DeepInfer	analyze	BM	3.45	687.58
	analyze	CIFAR10	55.72	29 150.96
	analyze	GC	4.89	661.32
	analyze	HP	2.82	643.71
	analyze	PD	2.67	659.31
	infer	BM	2.63	646.03
	infer	CIFAR10	29.01	29 146.96
	infer	GC	2.64	642.5
	infer	HP	2.62	646.69
	infer	PD	2.67	648.37
Prophecy	analyze	BM	3.94	730.33
	analyze	CIFAR10	118.42	4 177.04
	analyze	GC	3.33	711.58
	analyze	HP	3.22	709.97
	analyze	PD	3.27	710.97
	infer	BM	3.49	696.4
	infer	CIFAR10	25.97	2 012.66
	infer	GC	3.04	696.16
	infer	HP	3.07	699.33
	infer	PD	3.02	699.5
SelfChecker	analyze	BM	9.05	4 063.45
	analyze	CIFAR10	1 930.52	40 871.41
	analyze	GC	3.51	3 959.27
	analyze	HP	3.47	4 069.1
	analyze	PD	3.52	3 902.47
	infer	BM	4.44	3 773.3
	infer	CIFAR10	1 352.59	7 363.5
	infer	GC	3.22	3 796.63
	infer	HP	3.17	3 833.67
	infer	PD	3.17	3 758.48