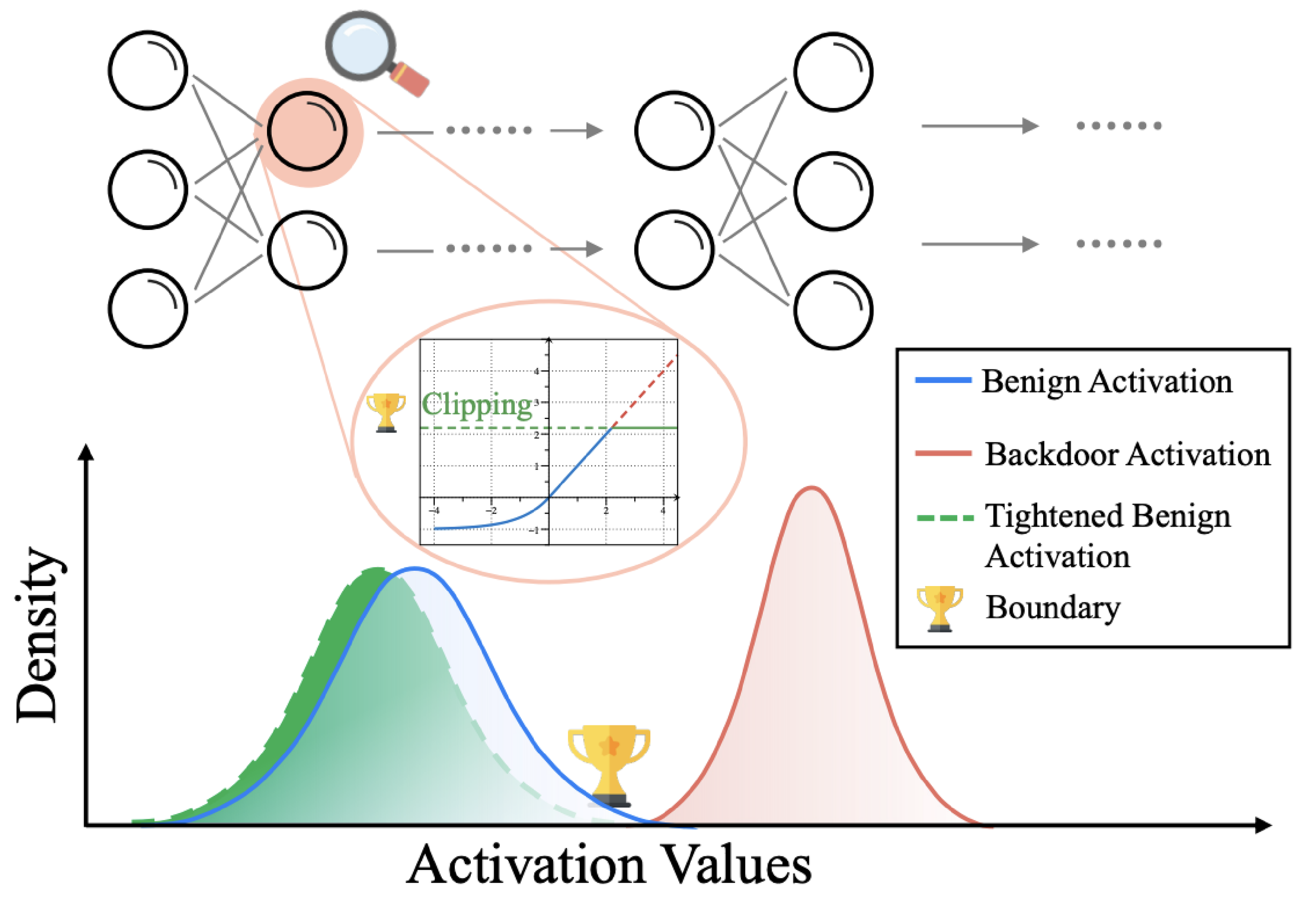
UNIT: Backdoor Mitigation via Automated Neural Distribution Tightening

Siyuan Cheng*, Guangyu Shen*, Kaiyuan Zhang, Guanhong Tao,
Shengwei An, Hanxi Guo, Shiqing Ma‡, Xiangyu Zhang

* denotes equal contribution

UNIT Overview



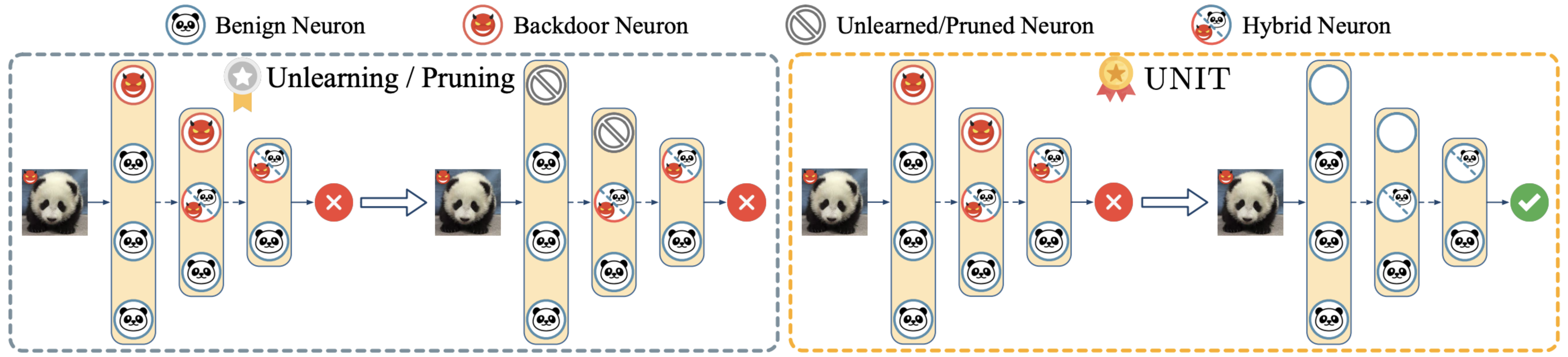
Detailed Algorithm

Algorithm 1 Automated Neural Distribution Tightening

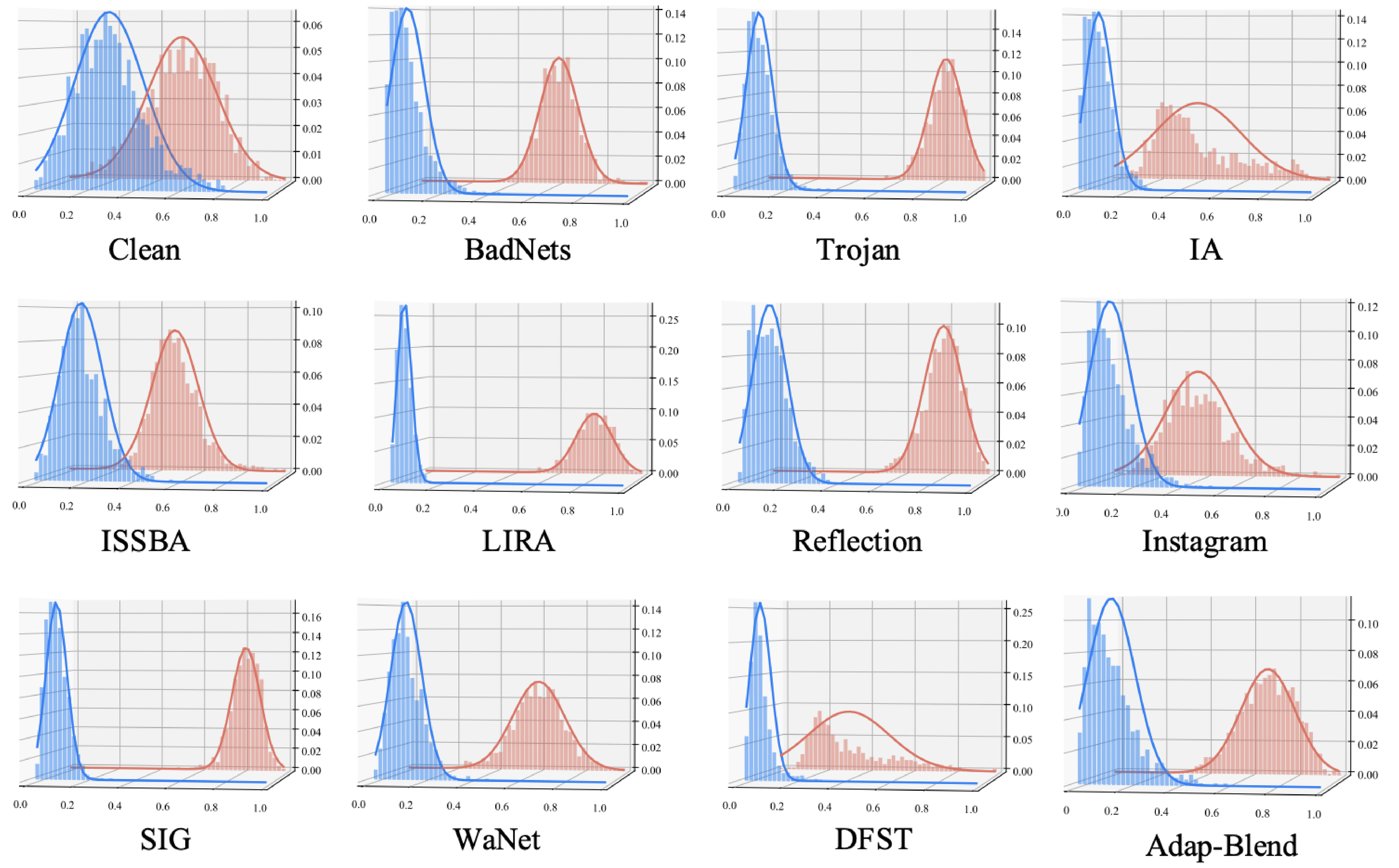
Input: Subject model M, Accuracy drop expectation ϵ, Training data $\{(x_i^t, y_i^t)\}_{i=1}^{n_t}$, Validation data $\{(x_i^v, y_i^v)\}_{i=1}^{n_v}$, Initial benign distribution boundary σ_0, Initial trade-off coefficient α_0, Optimization steps S, and Learning rate η.
Initialize: $\sigma = \sigma^\star = \sigma_0$, $\alpha = \alpha_0$
 ▷ *Calculate original accuracy on validation samples*
$P_0 = \frac{1}{n_v}\sum_{i=1}^{n_v} \mathbb{1}(M(x_i^v) = y_i^v)$
for $s = 1$ **to** S **do**
  ▷ *Cross-entropy loss plus boundary penalty*
  $\mathcal{L} = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}_{CE}(M_\sigma(x_i^t), y_i^t) + \alpha \cdot ||\sigma||_1$
  $\sigma = \sigma - \eta \cdot \frac{\partial \mathcal{L}}{\partial \sigma}$
  ▷ *Calculate accuracy when applying current bound*
  $P' = \frac{1}{n_v}\sum_{i=1}^{n_v} \mathbb{1}(M_\sigma(x_i^v) = y_i^v)$
  if $P_0 - P' > \epsilon$ **then**
    $\alpha = \alpha/2$
  else
    $\alpha = \alpha \cdot 2$
  end if
  ▷ *Update the best boundary value*
  if $P' \geqslant P_0 - \epsilon$ and $||\sigma||_1 < ||\sigma^\star||_1$ **then**
    $\sigma^\star = \sigma$
  end if
end for
Return: $\sigma^\star$

Limitation of Existing Backdoor Mitigation Methods

Existing methods either retrain the entire model without precise guidance for reducing backdoor effects or directly prune some specific neurons.
Advanced backdoor attacks tend to hide backdoor behavior within benign neurons that primarily process normal features.



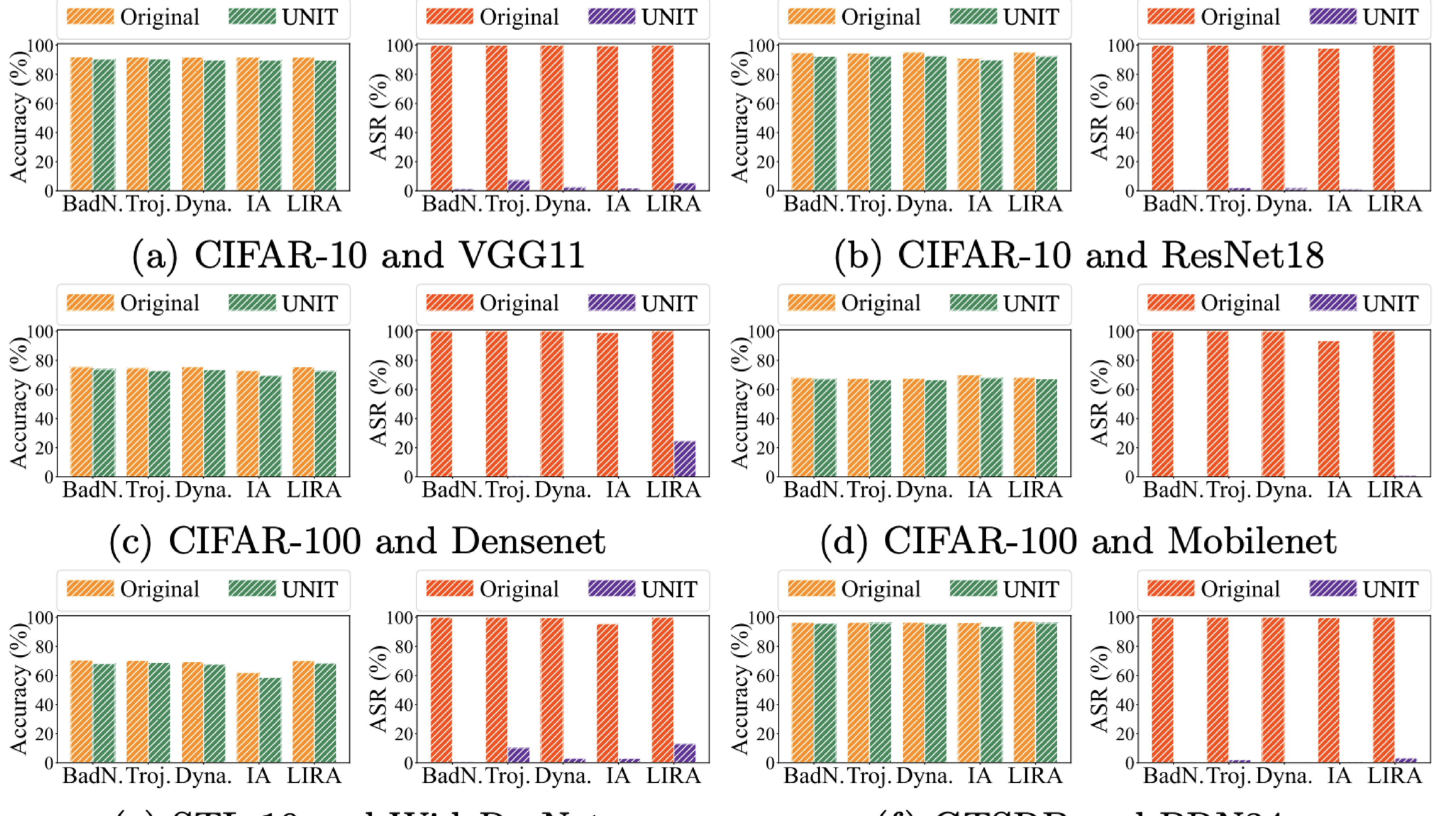
Neural Activation for Benign and Poisoned Samples



Evaluation Results

Table 1: Comparison of UNIT with 7 backdoor mitigation baselines against 14 backdoor attacks. Results are measured in percentages (%). All methods have access to 5% of the clean training data. The best results are highlighted in bold.

Attacks	Original		FT		FP		NAD		ANP		NC		I-BAU		SEAM		UNIT	
	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR	Acc.	ASR
BadNets	94.82	100.0	90.91	9.78	89.68	3.52	92.41	4.79	91.35	3.26	**93.04**	**0.34**	91.60	3.66	91.61	1.05	92.48	0.78
Trojan	94.73	100.0	91.63	35.11	90.76	31.14	91.52	22.30	92.37	58.88	91.89	4.01	90.73	11.58	92.28	12.69	**92.38**	**2.17**
CL	94.58	98.46	90.34	58.72	87.71	3.69	88.47	4.42	89.92	18.18	90.72	1.79	88.75	5.52	92.02	23.04	**92.21**	**1.09**
Dynamic	95.08	100.0	89.11	9.29	84.93	3.23	89.26	2.34	91.99	3.09	92.09	1.78	92.48	1.63	92.61	3.22	**92.77**	**1.54**
IA	91.15	97.96	88.44	2.92	89.71	82.20	88.51	2.67	89.05	5.49	89.32	1.12	89.79	62.45	89.77	1.23	**89.93**	**1.03**
Reflection	93.29	99.33	91.38	74.77	89.68	84.51	90.99	52.97	90.66	93.28	91.38	93.31	89.94	87.85	90.54	21.37	**91.44**	**6.63**
SIG	94.97	99.80	91.29	63.94	90.88	1.03	91.69	10.46	90.80	36.79	91.70	97.88	91.51	22.11	**92.57**	**0.68**	92.48	1.74
Blend	94.62	100.0	90.68	7.30	91.47	2.01	91.62	3.32	91.04	16.79	91.90	1.53	91.43	3.61	91.38	1.80	**91.99**	**1.18**
WaNet	94.36	99.80	90.32	2.85	91.48	1.48	92.36	1.91	**91.99**	**0.61**	90.60	0.97	89.67	12.01	91.34	1.44	91.02	2.44
ISSBA	94.55	100.0	91.40	4.17	90.79	2.11	92.45	2.43	92.42	2.98	**92.52**	**0.46**	83.03	84.58	91.17	3.00	91.84	1.57
LIRA	95.11	100.0	91.42	15.09	89.58	14.76	91.64	2.06	91.98	47.91	92.11	1.17	92.18	12.65	92.18	3.02	**92.29**	**0.58**
Instagram	94.62	99.59	91.40	29.25	90.38	8.03	89.50	7.17	90.10	5.10	90.19	15.88	89.25	7.24	91.35	5.89	**91.43**	**4.98**
DFST	93.25	99.77	90.88	35.22	90.66	14.03	91.05	14.59	89.70	20.51	91.22	24.77	89.12	6.19	91.22	12.93	**91.64**	**4.02**
Adap-Bl.	94.22	82.80	90.15	48.76	87.62	31.36	90.42	49.50	90.80	69.51	90.33	18.25	90.81	19.97	89.58	24.19	**90.84**	**15.03**
Average	94.26	98.39	90.57	28.37	89.67	20.22	90.85	12.92	91.01	27.31	91.36	18.80	90.02	24.36	91.48	8.08	**91.77**	**3.20**



(a) CIFAR-10 and VGG11
(b) CIFAR-10 and ResNet18
(c) CIFAR-100 and Densenet
(d) CIFAR-100 and Mobilenet
(e) STL-10 and WideResNet
(f) GTSRB and PRN34