Karlsruhe Institute of Technology

KASTEL

IntelliSEC
Artificial Intelligence & Security

Anti-Backdoor Coreset Selection via Cumulative Entropy

Qi Zhao and Christian Wressnegger

KASTEL Security Research Labs, Karlsruhe Institute of Technology (KIT)

Training-Time Defense Against Data Poisoning Backdoors

- **Taxonomy**

Techniques	Defenses				
	ABL Li et al. (2021)	DBD Huang et al. (2022)	CBD Zhang et al. (2023)	V&B Zhu et al. (2023)	Harvey Zhao et al. (2025)
Dataset Splitting	✓	✓	✓	✓	✓
Backdoor Unlearning	✓		✓		✓
Data Augementation				✓	

Training-Time Defense Against Data Poisoning Backdoors

- **Taxonomy**

Techniques	Defenses				
	ABL Li et al. (2021)	DBD Huang et al. (2022)	CBD Zhang et al. (2023)	V&B Zhu et al. (2023)	Harvey Zhao et al. (2025)
Dataset Splitting	✓	✓	✓	✓	✓
Backdoor Unlearning	✓		✓		✓
Data Augementation				✓	

Defense Performance for CIFAR-10

Attack	ABL	DBD	CBD	V&B	Harvey
	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR
BadNets	/ ●	/ ●	/	● / ●	● / ●
Blend	/	/ ●	/	● / ●	● / ●
Clean-Label	/ ●	/ ●	/	● / ●	● / ●
IAB	/ ●	/	/ ●	● /	● / ●
WaNet	/	/ ●	/	● / ●	● / ●
ISSBA	/ ●	/	/	● / ●	/ ●
Low-Frequency	/	/	/	● /	/
Adaptive-Blend	/	/ ●	/	/	● /

●: **promising performance** ACC reduction ≤ 1%, ASR ≤ 10%

Defense Performance for CIFAR-10

Attack	ABL	DBD	CBD	V&B	Harvey
	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR
BadNets	🟢 / ⚫	🟢 / ⚫	🟢 / 🟢	⚫ / ⚫	⚫ / ⚫
Blend	/ 🟢	🟢 / ⚫	/ 🟢	⚫ / ⚫	⚫ / ⚫
Clean-Label	/ ⚫	🟢 / ⚫	/ 🟢	⚫ / ⚫	⚫ / ⚫
IAB	🟢 / ⚫	🟢 /	🟢 / ⚫	⚫ / 🟢	⚫ / ⚫
WaNet	/	🟢 / ⚫	🟢 / 🟢	⚫ / ⚫	⚫ / ⚫
ISSBA	/ ⚫	🟢 / 🟢	🟢 /	⚫ / ⚫	🟢 / ⚫
Low-Frequency	/ 🟢	🟢 /	🟢 /	⚫ /	🟢 / 🟢
Adaptive-Blend	/	🟢 / ⚫	🟢 /	🟢 /	⚫ /

⚫: **promising performance** ACC reduction ≤ 1%, ASR ≤ 10%

🟢: **mediocre performance** ACC reduction 1% ~ 5%, ASR 10% ~ 50%

Defense Performance for CIFAR-10

Attack	ABL	DBD	CBD	V&B	Harvey
	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR
BadNets	🟢 / ●	🟢 / ●	🟢 / 🟢	● / ●	● / ●
Blend	○ / 🟢	🟢 / ●	○ / 🟢	● / ●	● / ●
Clean-Label	○ / ●	🟢 / ●	○ / 🟢	● / ●	● / ●
IAB	🟢 / ●	🟢 / ○	🟢 / ●	● / 🟢	● / ●
WaNet	○ / ○	🟢 / ●	🟢 / 🟢	● / ●	● / ●
ISSBA	○ / ●	🟢 / 🟢	🟢 / ○	● / ●	🟢 / ●
Low-Frequency	○ / 🟢	🟢 / ○	🟢 / ○	● / ○	🟢 / 🟢
Adaptive-Blend	○ / ○	🟢 / ●	🟢 / ○	🟢 / ○	● / ○

●: **promising performance** ACC reduction ≤ 1%, ASR ≤ 10%

🟢: **mediocre performance** ACC reduction 1% ~ 5%, ASR 10% ~ 50%

○: **failed performance** 😭 ACC reduction ≥ 5%, ASR ≥ 50%

Problem Statement

- **State-of-the-art defenses**
  - Insufficient robustness
  - Trade-off between ACC and ASR
- **Rethinking of the dataset splitting strategy** 🤔
  - A **non-precise** splitting is **more robust**
  - Focus on data **quality (informativeness)** rather quantity

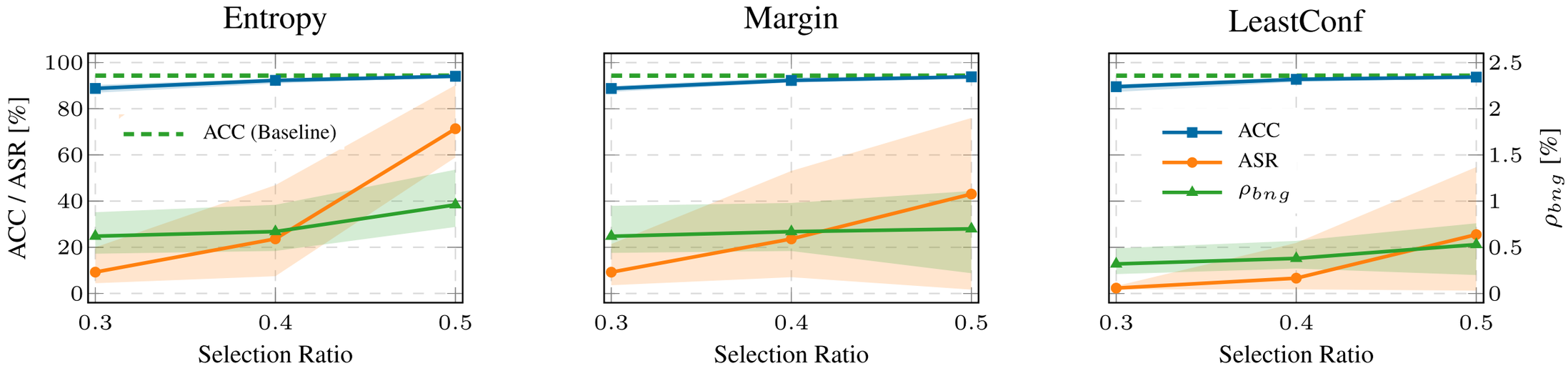
Problem Statement

- **State-of-the-art defenses**
  - Insufficient robustness
  - Trade-off between ACC and ASR
- **Rethinking of the dataset splitting strategy** 🤔
  - A **non-precise** splitting is **more robust**
  - Focus on data **quality (informativeness)** rather quantity

} **Dataset Coreset Selection**

Coreset Selection of Poisoned Datasets (1)

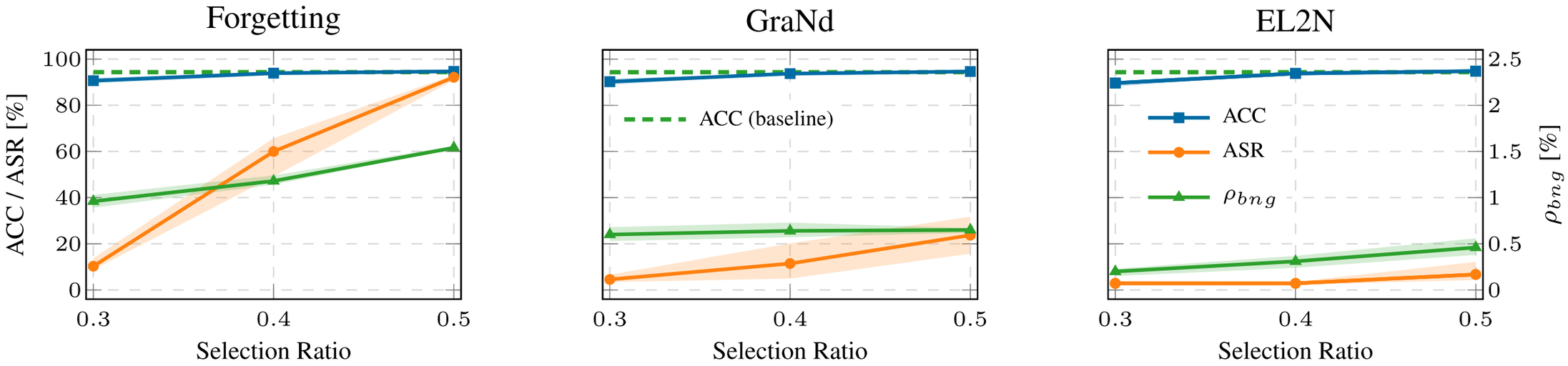
- **Investigation with uncertainty criteria** Coleman et al. (2020)
  - Default poisoning ratio: 5%



  - Proper selection ratios yield a low poisoning ratio ρ_{bng} in the coreset 👍

Coreset Selection of Poisoned Datasets (2)

- **Investigation with loss-based criteria**
  - Default poisoning ratio: 5%



Forgetting Events Toneva et al. (2019)

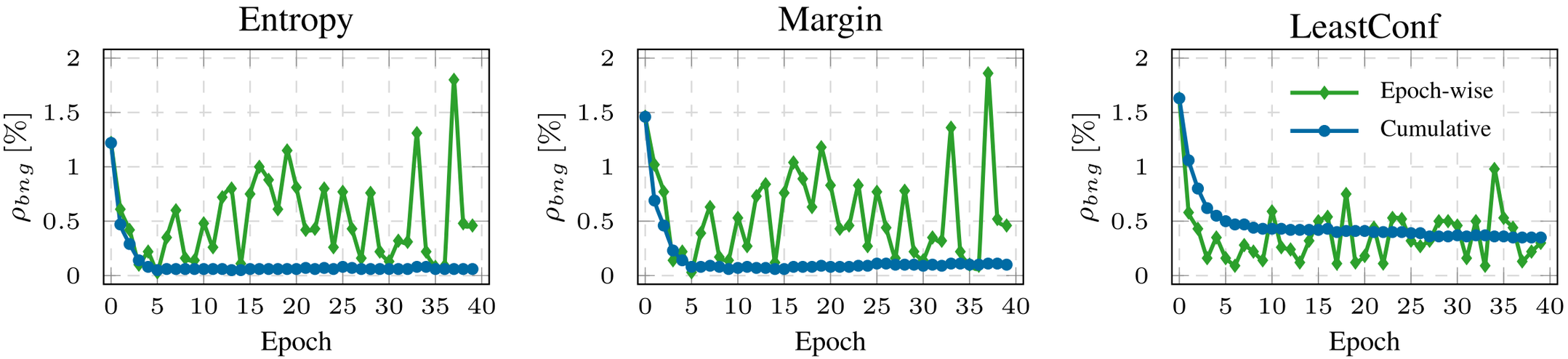
Gradient Accumulation Paul et al. (2021)

L2-Loss Accumuation Paul et al. (2021)

- The **accumulation** over training yields **high anti-backdoor stability** 👍

Coreset Selection by Cumulative Uncertainty (1)

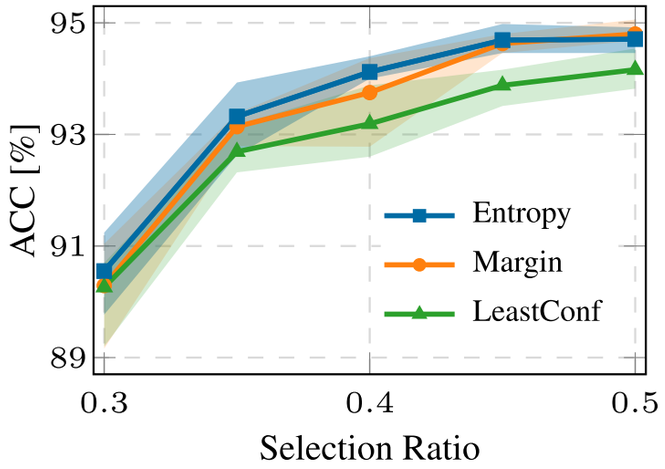
- **Learning dynamics**

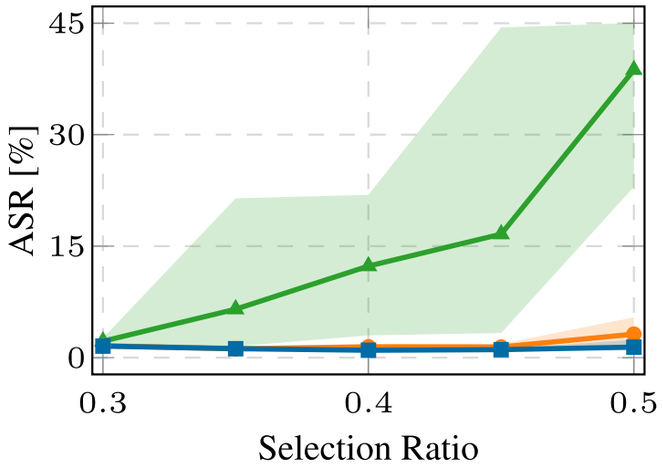


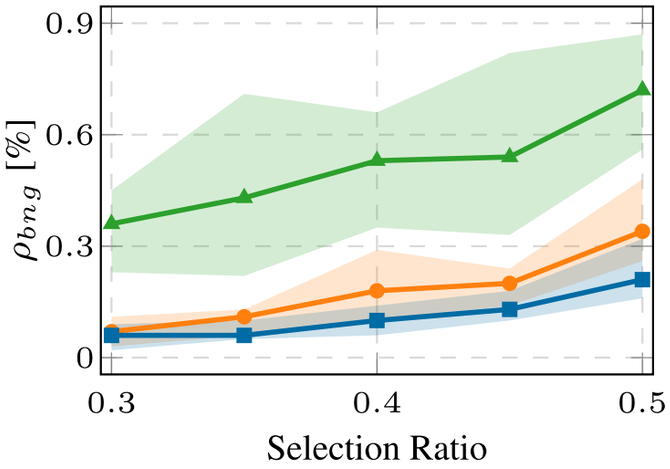
- **Accumulation over epochs** can effectively excluding poisonous samples

Coreset Selection by Cumulative Uncertainty (2)

- **Defensive performance**







- The best adaption to the accumulation strategy: **Cumulative Entropy**

ABCS: Anti-Backdoor Coreset Seletion

- ***Cumulative Entropy* (CENT) criterion**
  - Expression for individual training sample $\mathbf{x}_i$

$$\mathrm{CENT}(\mathbf{x}_i) = \frac{1}{T_{se}} \sum_{t=1}^{T_{se}} \underbrace{\mathrm{normalize}\left(\sum_{c=0}^{C-1} -p_{\theta_t}\left(c|\mathbf{x}_i\right) \cdot \log\left(p_{\theta_t}\left(c|\mathbf{x}_i\right)\right) \right)}_{\text{Normalized Shannon Entropy } H(\mathbf{x}_i)}$$

where T_{se} denotes the number of selection epochs and C is the number of classes

ABCS: Coreset Selection Procedure

- **Three-stages pipeline**
  1. Warm-up training on the poisoned dataset $\widetilde{\mathcal{D}}$ for T_{wa} epochs

ABCS: Coreset Selection Procedure

- **Three-stages pipeline**

  1. Warm-up training on the poisoned dataset $\widetilde{\mathcal{D}}$ for T_{wa} epochs
  2. Selecting the coreset over T_{se} epochs. In each epoch, we:

     Step 1: Naively train on $\widetilde{\mathcal{D}}$ for one epoch

     *Step 2: Unlearn high-CENT samples via **label-smoothing*** Di et al. (2026)

     Step 3: Compute $H(\mathbf{x}_i)$ and accumulate it to the latest $\mathrm{CENT}(\mathbf{x}_i)$

ABCS: Coreset Selection Procedure

- **Three-stages pipeline**
  1. Warm-up training on the poisoned dataset $\widetilde{\mathcal{D}}$ for T_{wa} epochs
  2. Selecting the coreset over T_{se} epochs. In each epoch, we:

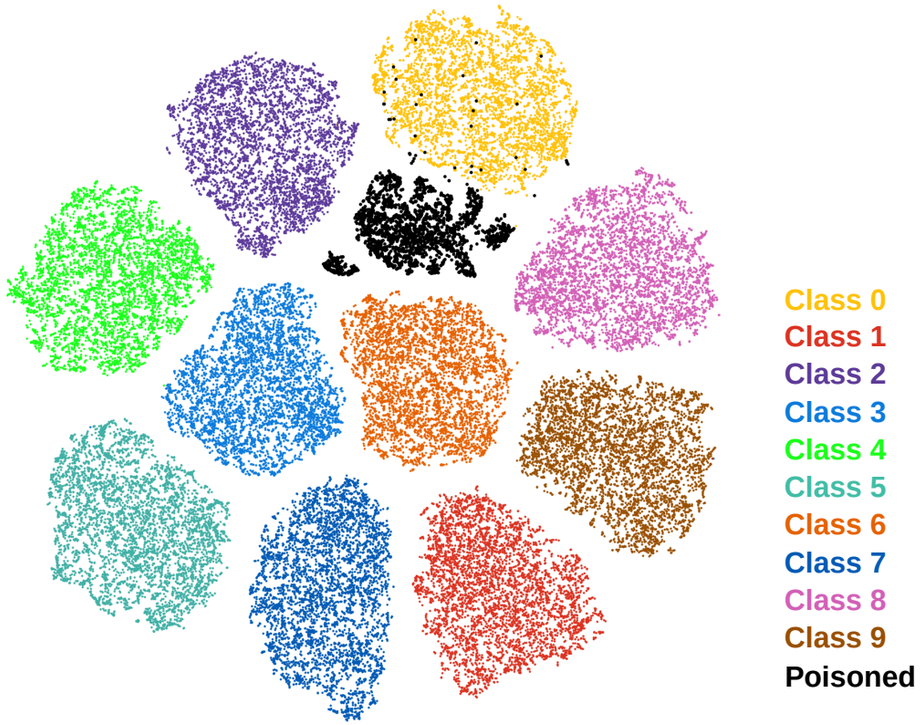
     Step 1: Naively train on $\widetilde{\mathcal{D}}$ for one epoch

     Step 2: Unlearn high-CENT samples via ***label-smoothing*** Di et al. (2026)

     Step 3: Compute $H(\mathbf{x}_i)$ and accumulate it to the latest $\mathrm{CENT}(\mathbf{x}_i)$
  3. Sampling samples with highest CENT values by
     - Mode 1: a fixed size (e.g., 0.4, 0.5, ...)
     - Mode 2: an adaptive cut-off threshold τ for coreset size calculation

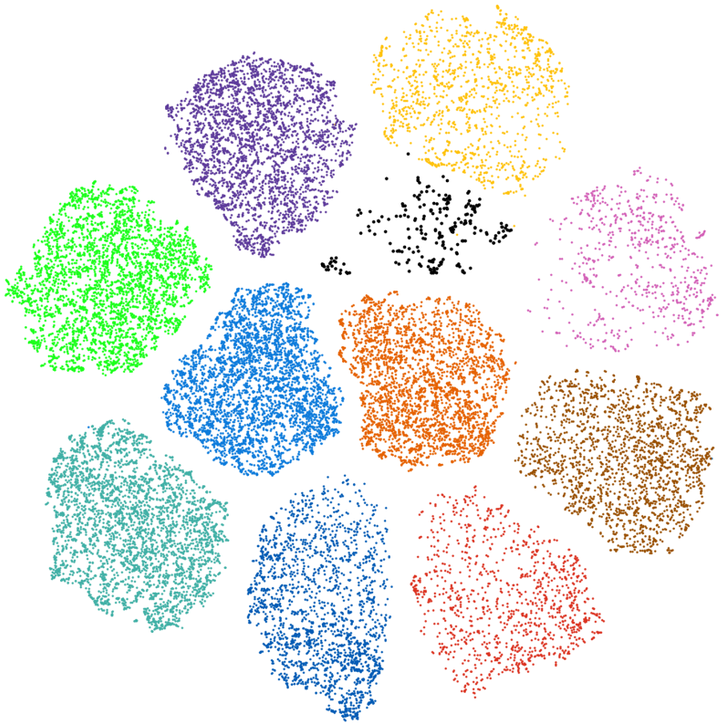
ABCS: Visualization of The Selection Stage

- **Distribution of a poisoned CIFAR-10**



Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Poisoned

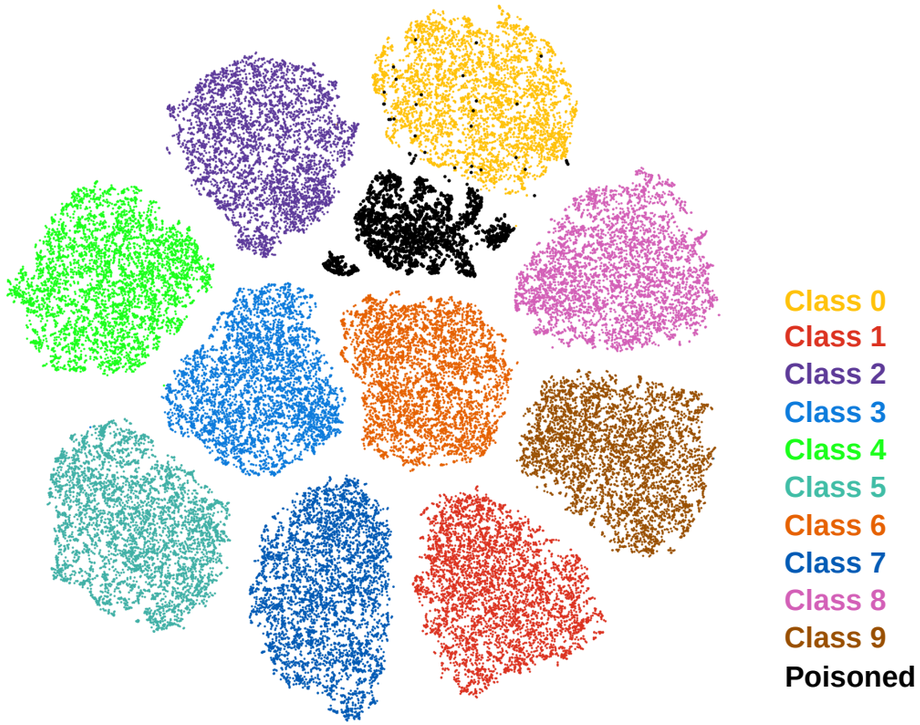
Coreset Selection



Epoch 1

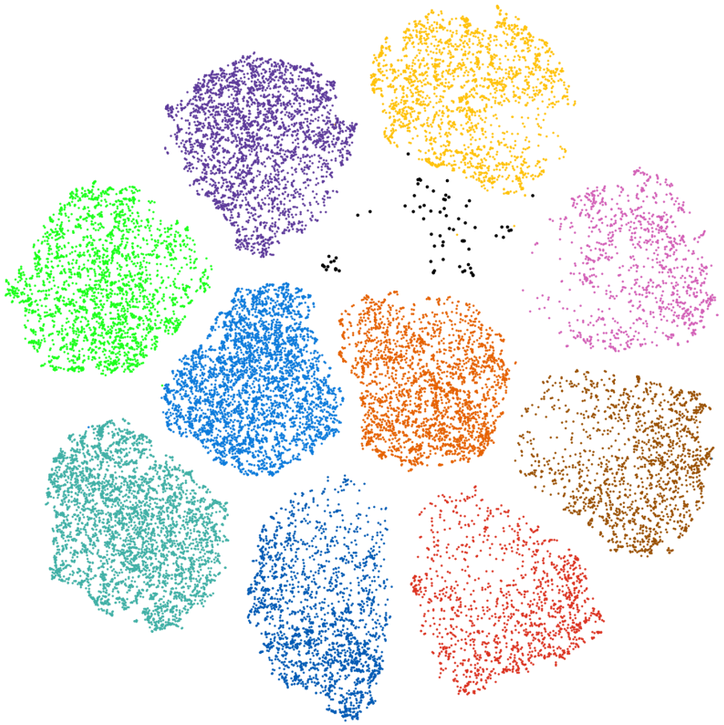
ABCS: Visualization of The Selection Stage

- **Distribution of a poisoned CIFAR-10**



Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Poisoned

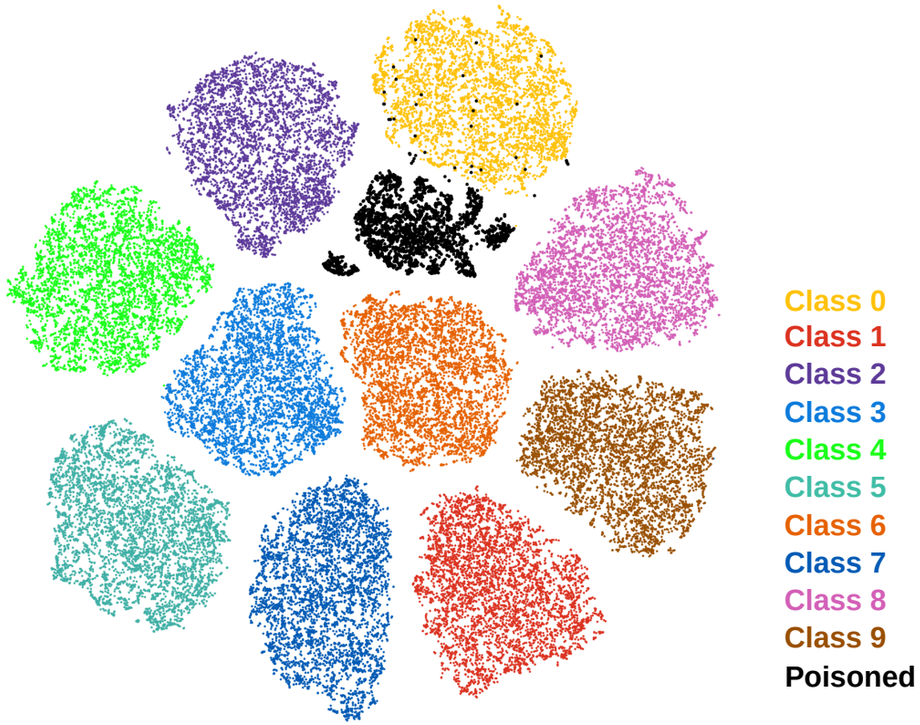
Coreset Selection



Epoch 10

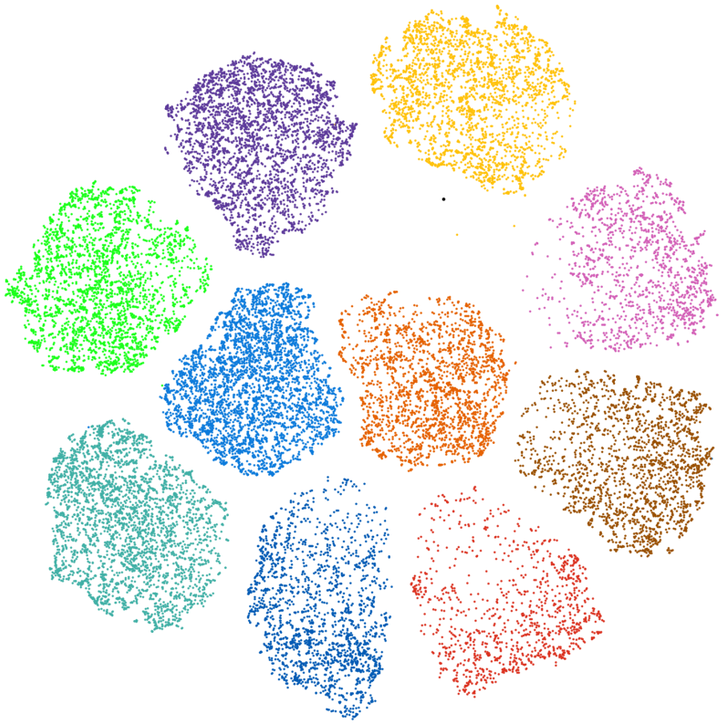
ABCS: Visualization of The Selection Stage

- **Distribution of a poisoned CIFAR-10**



Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Poisoned

Coreset Selection



Epoch 20

Evaluation on CIFAR-10

Attack	ABL	DBD	CBD	V&B	Harvey	ABCS
	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR	ACC / ASR
BadNets	● / ●	● / ●	● / ●	● / ●	● / ●	● / ●
Blend	○ / ●	● / ●	○ / ●	● / ●	● / ●	● / ●
Clean-Label	○ / ●	● / ●	○ / ●	● / ●	● / ●	● / ●
IAB	● / ●	● / ○	● / ●	● / ●	● / ●	● / ●
WaNet	○ / ○	● / ●	● / ●	● / ●	● / ●	● / ●
ISSBA	○ / ●	● / ●	● / ○	● / ●	● / ●	● / ●
Low-Frequency	○ / ●	● / ○	● / ○	● / ○	● / ●	● / ●
Adaptive-Blend	○ / ○	● / ●	● / ○	● / ○	● / ○	● / ●

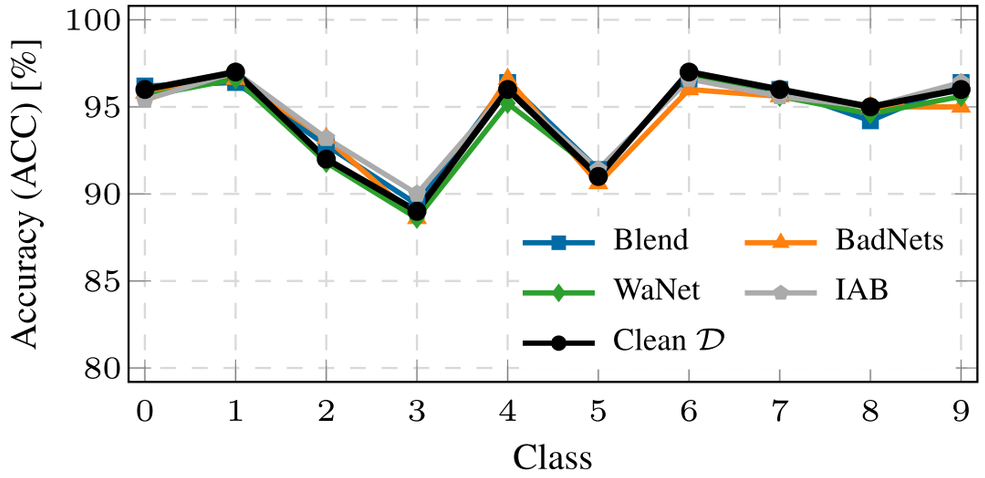
●: **promising performance** ACC reduction ≤ 1%, ASR ≤ 10%

●: **mediocre performance** ACC reduction 1% ~ 5%, ASR 10% ~ 50%

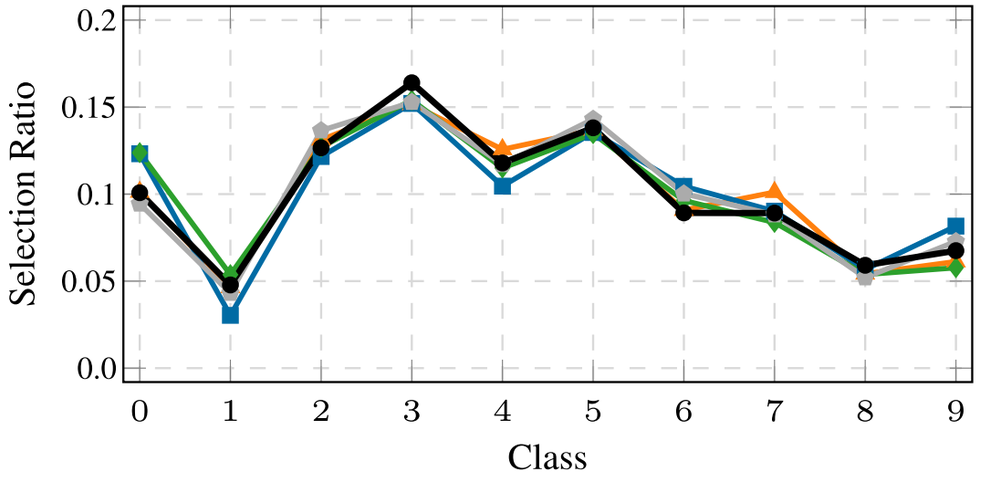
○: **failed performance** ACC reduction ≥ 5%, ASR ≥ 50%

Class-wise Analysis

- **Accuracy is preserved**



- **Class-wise distribution is preserved**



Conclusion

- **Anti-backdoor coreset selection (ABCS)**
  - Allows a non-precise dataset splitting → improving the defense robustness
  - Focuses on the selection of clean and informative samples
- **The robust selection criterion, Cumulative Entropy (CENT), that**
  - Substantially improves the stability of coreset selection
  - Ensures selecting a coreset with a significant backdoor elimination

Thank You!

Artificial Intelligence & Security
Karlsruhe Institute of Technology

https://intellisec.de/team/qi

https://intellisec.de/research/abcs

References

Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J., and Zaharia, M. Selection via proxy: Efficient data selection for deep learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2020.

Di, Z., Zhu, Z., Jia, J., Liu, J., Takhirov, Z., Jiang, B., Yao, Y., Liu, S., and Liu, Y. Label smoothing improves machine unlearning. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2026.

Huang, K., Li, Y., Wu, B., Qin, Z., and Ren, K. Backdoor defense via decoupling the training process. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS),* 2021.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: finding important examples early in training. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

Zhang, Z., Liu, Q., Wang, Z., Lu, Z., and Hu, Q. Backdoor defense via deconfounded representation learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhao, Q. and Wressnegger, C. Two sides of the same coin: Learning the backdoor to remove the backdoor. In *Proc. of the Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Zhu, Z., Wang, R., Zou, C., and Jing, L. The victim and the beneficiary: Exploiting a poisoned model to train a clean model on poisoned data. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.