2024 - AI - 77C-880 - Securing Deep Neural Networks on Edge from Membership Inference Attacks Using Trusted Execution Environments - gramshan@purdue.edu - Gowri Ramshankar

CERIAS

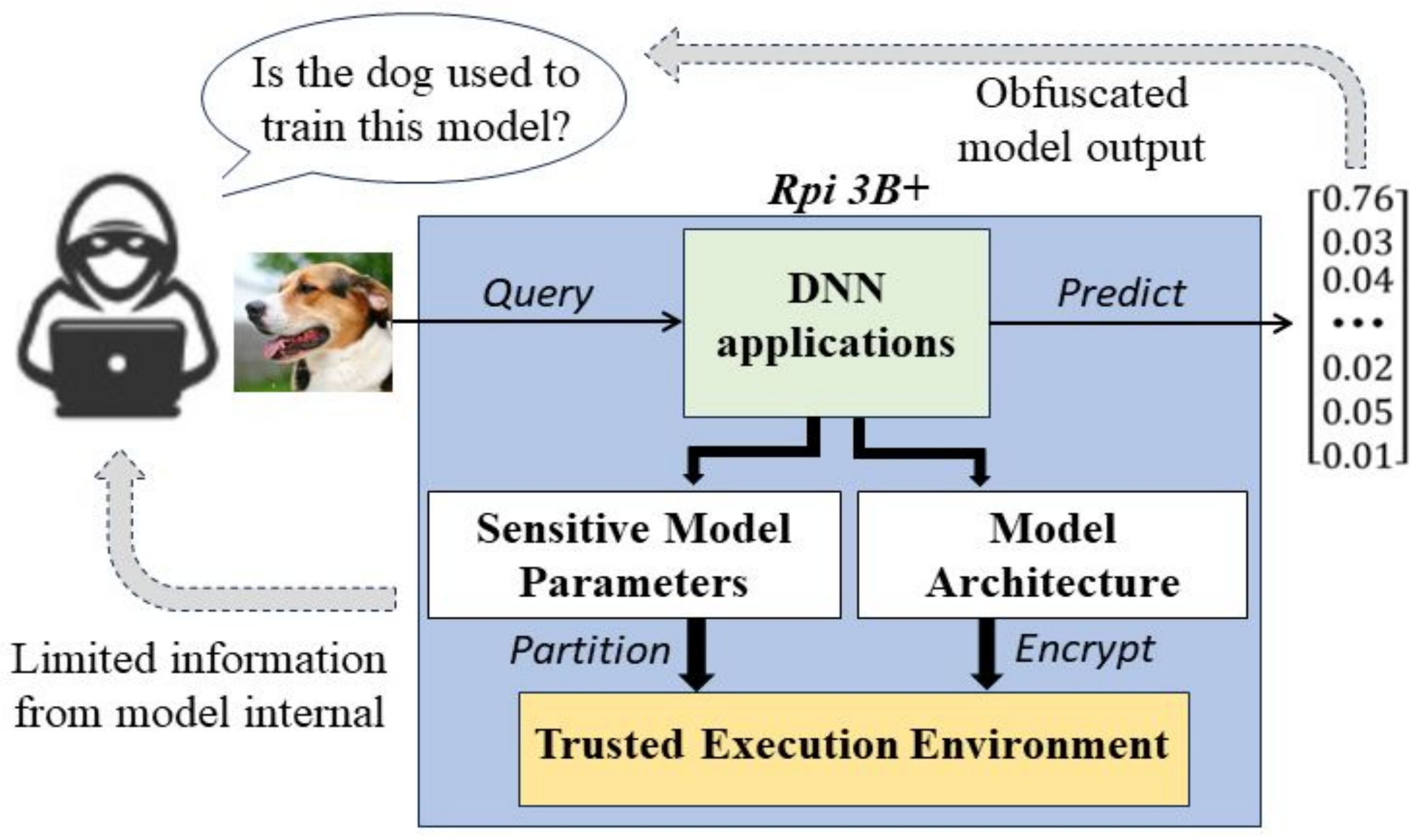
The Center for Education and Research in Information Assurance and Security

Securing Deep Neural Networks on Edge from Membership Inference Attacks Using Trusted Execution Environments

Gowri Ramshankar, Cheng-Yun Yang, Yung-Hsiang Lu

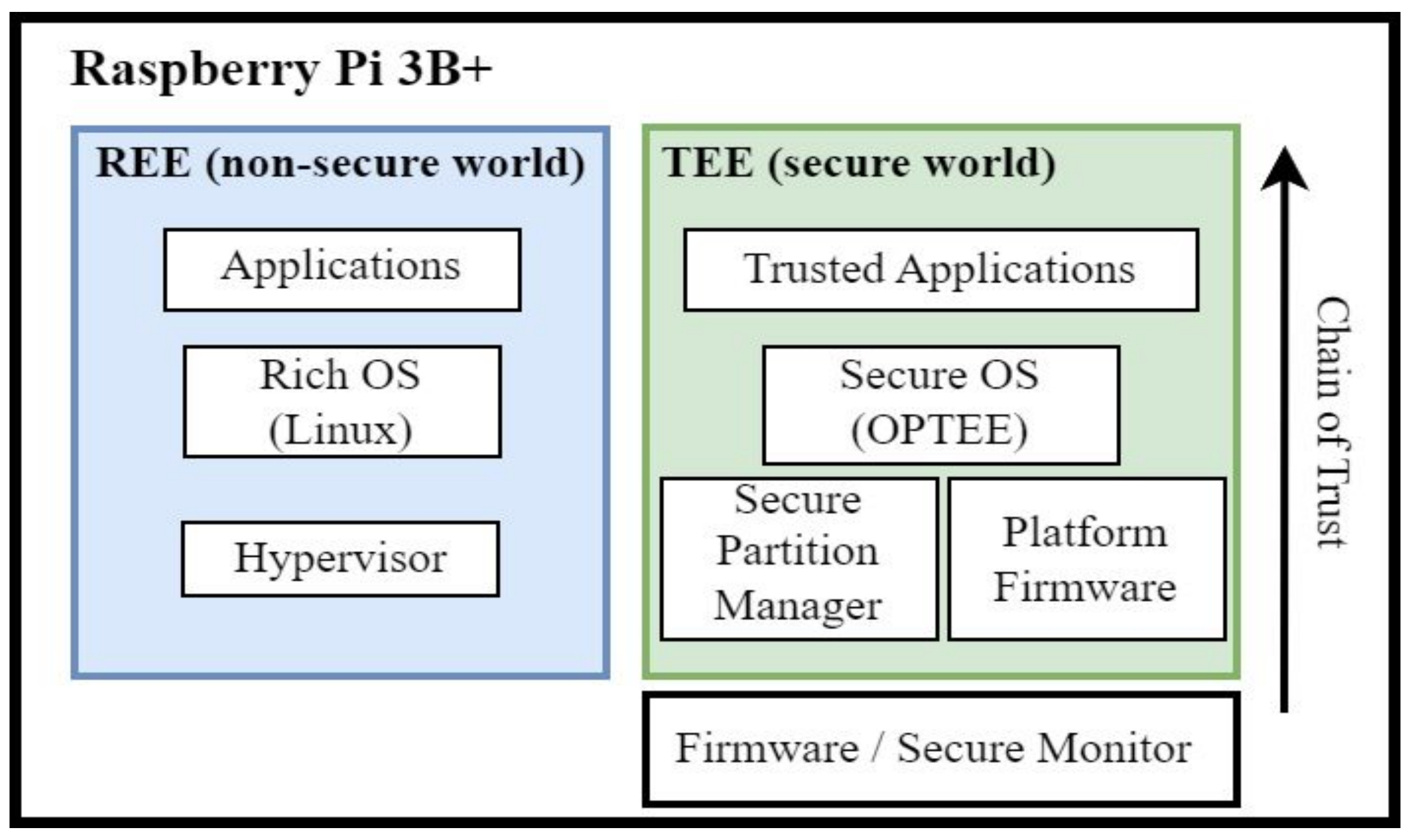
Background

- **Deep Neural Networks (DNNs) are under high security risk** when a large amount of sensitive data is used for training. For example, some business organizations apply DNNs to analyze the medical history of their customers.
- **Membership inference attacks (MIAs)** are designed to target such DNN models. They take the predictions or gradients during inference as input to determine if a specific data is part of the model's training set. We apply **Trusted Execution Environments,** which is a hardware feature in a processor to secure DNN models.
- An overview of our research problem is illustrated below:


Is the dog used to train this model?
Obfuscated model output
Rpi 3B+
Query
DNN applications
Predict
0.76
0.03
0.04
...
0.02
0.05
0.01
Sensitive Model Parameters
Model Architecture
Partition
Encrypt
Trusted Execution Environment
Limited information from model internal

Trusted Execution Environment (TEE)

- TEE is an area on the main processor of a device that is separated from the main operating system or the Rich Execution Environment (REE). We apply **OP-TEE (Secure operating System)** on Raspberry Pi 3B+ with support of **TrustZone-A** technology as our implementation of the TEE.

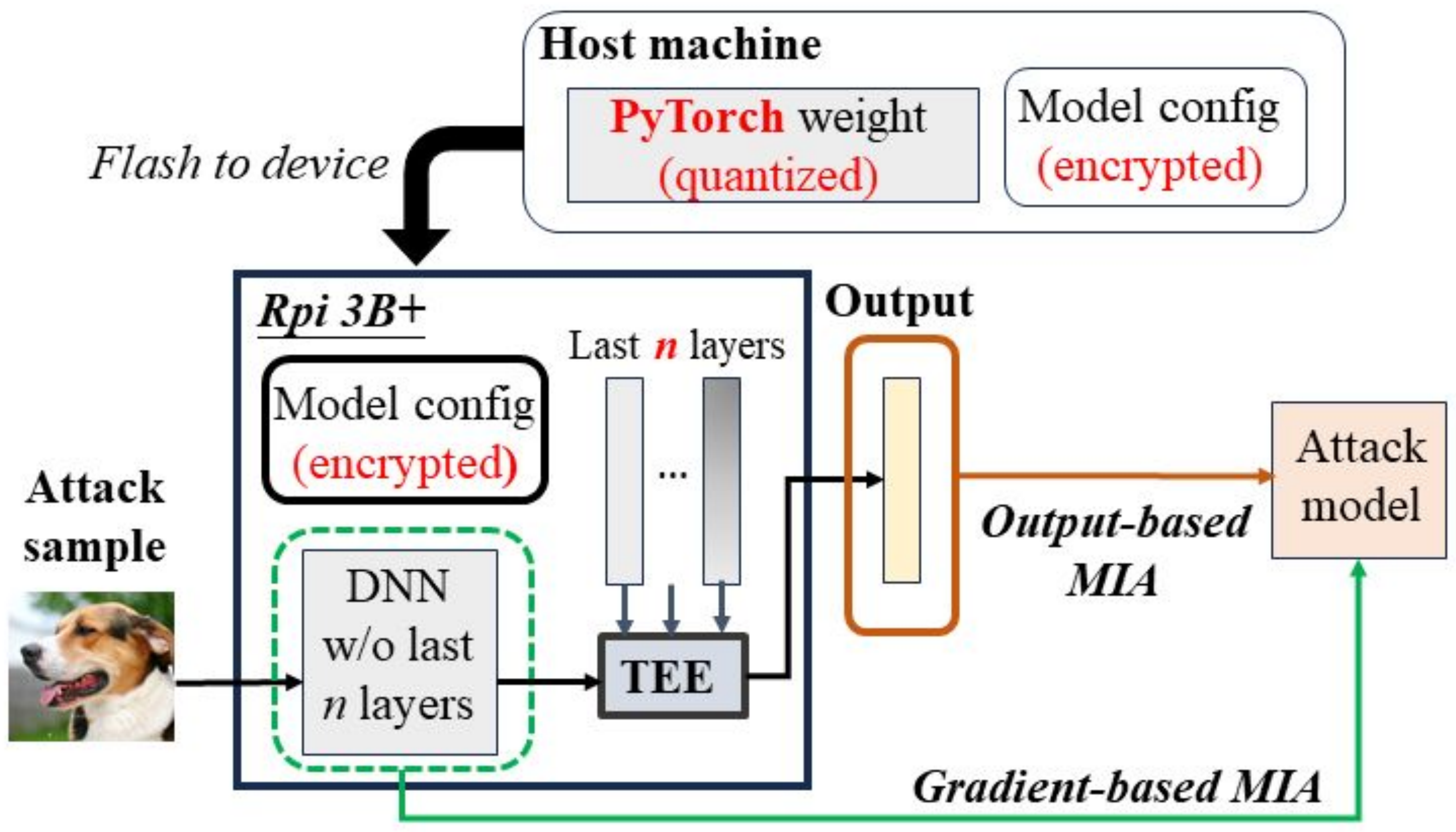


Problems of using TEE to protect data privacy

- The protected hardware resources in a TEE are constrained. It is impossible to deploy a DNN model with a large number of parameters into TEE. The computational cost in TEE is relatively high compared with that in REE because the popular OS of TEE only enables CPU with very limited memory.
- Our approach is to split the inference process of a model between the REE and the TEE. It also employs depth-wise partitioning of layers and model quantization to make deployment of DNNs in TEEs feasible.

Secure DNN Applications on Edge

- An illustration of our secure system on edge is given below. Our system can automate the process of securing **PyTorch**-based DNN models inside TEEs.



- Experiments on **Raspberry Pi 3B+** show that our method can reduce the accuracy of gradient-based MIAs and output-based MIAs when compared with DarknetZ and vanilla models as shown below:

Dataset	Model	MIA Accuracy		
		Vanilla	Mo et al. [8]	Ours
CIFAR-100	AlexNet	88.0%	65.1%	59.2%
	VGG-16	77.9%	72.4%	66.9%
	ResNet-20	85.0%	60.6%	49.7%
Tiny ImageNet	AlexNet	94.1%	70.7%	59.1%
	VGG-16	75.5%	62.4%	60.0%
	ResNet-20	65.1%	61.0%	57.0%

Victim model	Shadow Model			
	AlexNet	VGG-11	VGG-16	ResNet-20
AlexNet	87.7%	82.3%	86.7%	78.1%
AlexNet(Q)	69.9%	68.5%	69.2%	61.8%
VGG-11	88.9%	87.9%	88.4%	79.2%
VGG-11(Q)	60.1%	77.2%	75.9%	64.0%
VGG-16	78.5%	68.2%	87.4%	77.6%
VGG-16(Q)	74.0%	63.1%	84.5%	71.5%
ResNet-20	71.7%	82.8%	63.4%	94.4%
ResNet-20(Q)	64.8%	62.5%	62.9%	88.9%

➔ The accuracy of gradient-based MIAs on AlexNet, VGG-16, and ResNet-20 evaluated on the CIFAR-100 dataset is reduced by 28.8%, 11%, and 35.3%, respectively.

➔ The accuracy of output-based MIAs on AlexNet, VGG-16, and ResNet-20 is also reduced by 18.5%, 13.4%, and 29.6%, respectively.

- The table below compares the memory and execution time per image on Cifar-100 dataset with and without the proposed method:

Victim model	Memory size (MB)	Execution Time (s)	
		Vanilla	Ours
AlexNet	9.76	4.65	13.06
AlexNet(Q)	2.44	1.88	6.62
VGG-11	36.23	3.97	10.89
VGG-11(Q)	9.06	1.61	5.53
VGG-16	57.69	4.65	11.41
VGG-16(Q)	14.42	1.88	5.80
ResNet-20	1.14	4.19	5.30
ResNet-20(Q)	0.28	1.69	2.27

➔ The memory sizes of all DNN models are reduced by four times. The execution time of ResNet-20 protected by our method is even less than the vanilla one.

References

- OP-TEE: https://optee.readthedocs.io/en/latest/index.html
- Mo, Fan, et al. "Darknetz: towards model privacy at the edge using trusted execution environments." MobiSys, 2020.
- Shokri, Reza, et al. "Membership inference attacks against machine learning models." IEEE symposium on security and privacy, 2017.

PURDUE UNIVERSITY

CERIAS