# NeuralFP: Out-of-distribution Detection using Fingerprints of Neural Networks

**Wei-Han Lee**, Steve Millman,
Nirmit Desai, Mudhakar Srivatsa, Changchang Liu

IBM T. J. Watson Research Center

# Outline

- Background: OOD Detection in Edge Devices
- Our method: NeuralFP
    - ➢Motivating Example
    - ➢Design Details
- Experimental Analysis
    - ➢Effectiveness in Detecting OOD Data
    - ➢Usefulness of One-Out Integration Strategy
    - ➢Advantageous over Previous works
- Discussion

# OOD Detection in Edge Devices

Edge devices use neural network models learnt on cloud to predict labels of its data records
- ➢ may lead to incorrect predictions especially for OOD data
- ➢ may not properly detect the drifting data

However, recent efforts in OOD detection
- ➢ require the retraining of the model
- ➢ assume the existence of a certain amount of OOD records

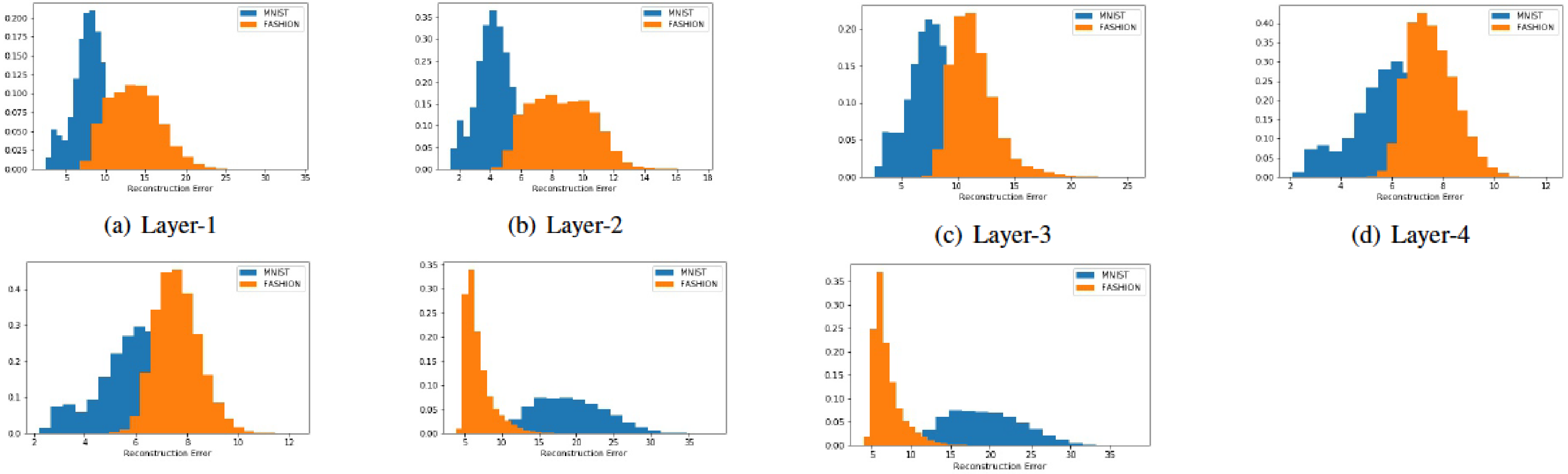
The state-of-the-art method [4] utilizes the Mahalanobis distance to explore the internal features of the convolutional neural networks, which is restricted by the adopted *linear transformation* and the limited usage of *the last hidden layer*

Motivations:

1) Is it possible to leverage nonlinear transformations to extract representative fingerprints of the training set in practice?
2) What is the optimal strategy to integrate fingerprinting information across multiple layers (instead of only using the last hidden layer) for maximizing detection accuracy?

[4] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” NeuralPS 2018

# NeuralFP: Motivating Example



(a) Layer-1 (b) Layer-2 (c) Layer-3 (d) Layer-4

(e) Layer-5 (f) Layer-6 (g) Layer-7

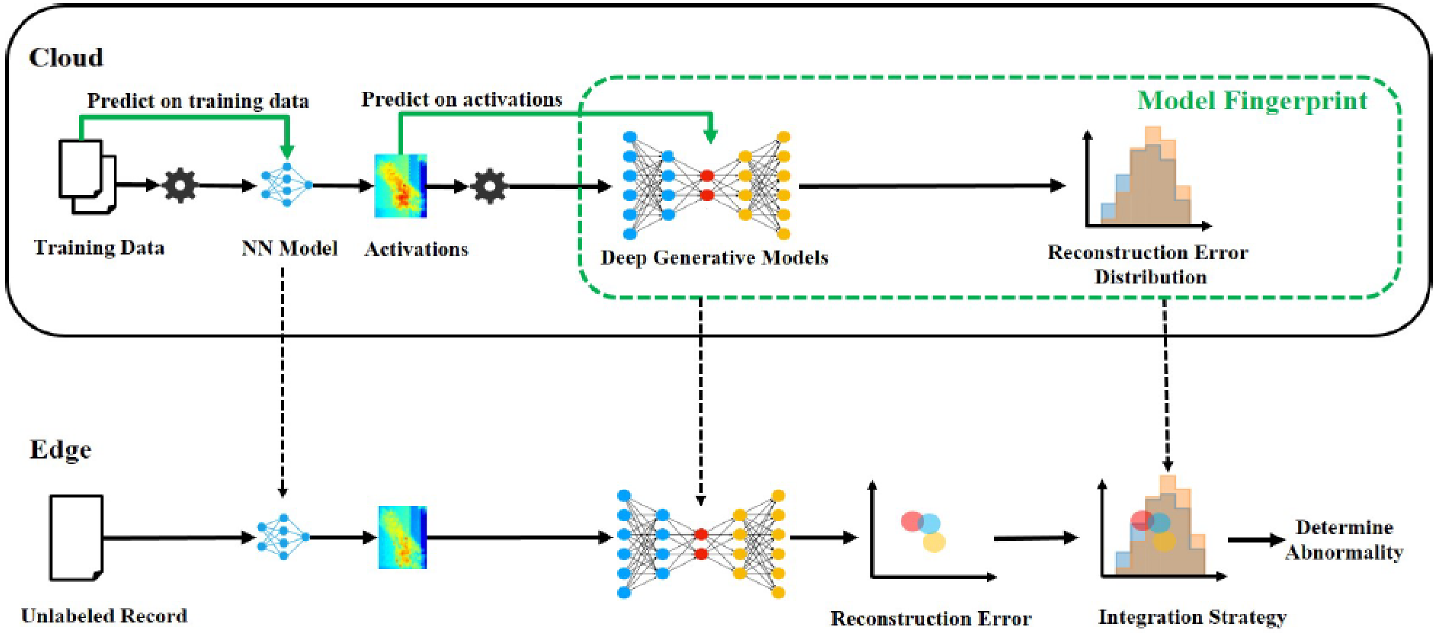
Distribution of reconstruction errors for MNIST testing data (blue) and Fashion-MNIST testing data (orange) applied to the model fingerprints of MNIST training data.

1) the reconstruction errors are strong signals for distinguishing OOD records from the in-distribution data
2) The reconstruction error of OOD records may not always be larger than that of the in-distribution data (Layer-6 vs Layer-7)
3) different layers have various capability in distinguishing OOD records (Layer-1 vs Layer-3)

**Key Intuition:**

*Neural network responds differently to in-distribution data and the OOD data*

# NeuralFP: Design Details



Framework of NeuralFP

**Fingerprinting on the Cloud**

1) obtain activations for each layer by passing training data through neural network model
2) construct deep generative models (such as autoencoders) of each layer based on activations
3) compute the reconstruction errors of these autoencoders applied on training data

**OOD Detection on the Edge**

pass a given data through the neural network model to compare with the stored model fingerprints for determining abnormality.

Specifically, a data record $x^*$ would be classified as Outlier if its reconstruction error $e_l^* = \mathcal{L}(a_l(x^*), \hat{a}_l(x^*))$ is

$$\exists\ l, \quad e_l^* > \tau_l \quad \text{or} \quad e_l^* < \mu_l$$

**One-out integration strategy**

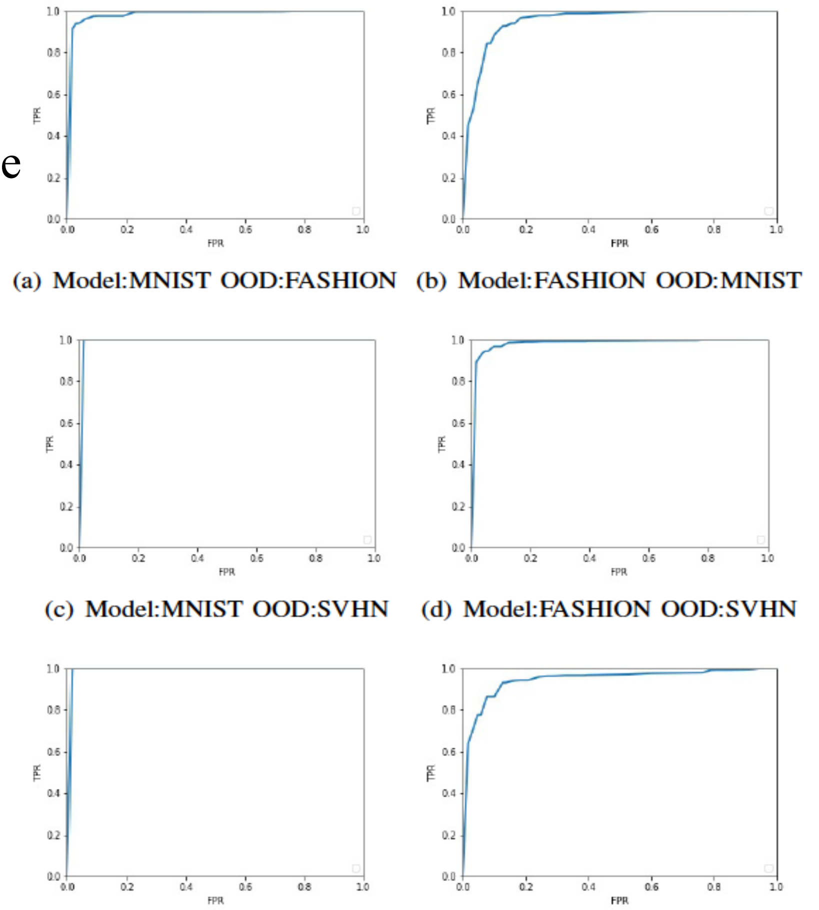
# NeuralFP: Experimental Results

## Effectiveness in Detecting OOD Data

1) NeuralFP achieves good performance in detecting outliers since the true positive rate can quickly increase to 1 under small values of false positive rate for most scenarios.
2) The high AUC scores (over 0.94 for all cases) demonstrate the robustness of NeuralFP under various settings of parameters.
3) ROC curves obtained by NeuralFP can be leveraged by practitioners to select appropriate values of the threshold parameters so that a user-specified tradeoff between true positive rate and false positive rate can be achieved.

AUCs OF DETECTING VARIOUS OOD RECORDS

| OOD<br>Training | MNIST | FASHION | CIFAR10 | SVHN |
|---|---|---|---|---|
| MNIST | N.A. | 0.9826 | 0.9921 | 0.9921 |
| FASHION | 0.9498 | N.A. | 0.9442 | 0.9809 |



(a) Model:MNIST OOD:FASHION (b) Model:FASHION OOD:MNIST

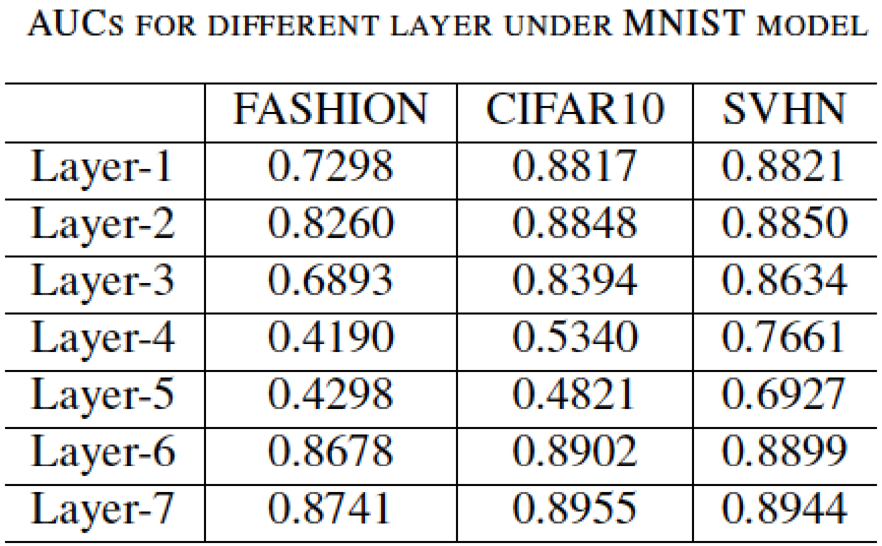
(c) Model:MNIST OOD:SVHN (d) Model:FASHION OOD:SVHN

(e) Model:MNIST OOD:CIFAR10 (f) Model:FASHION OOD:CIFAR10

# NeuralFP: Experimental Results

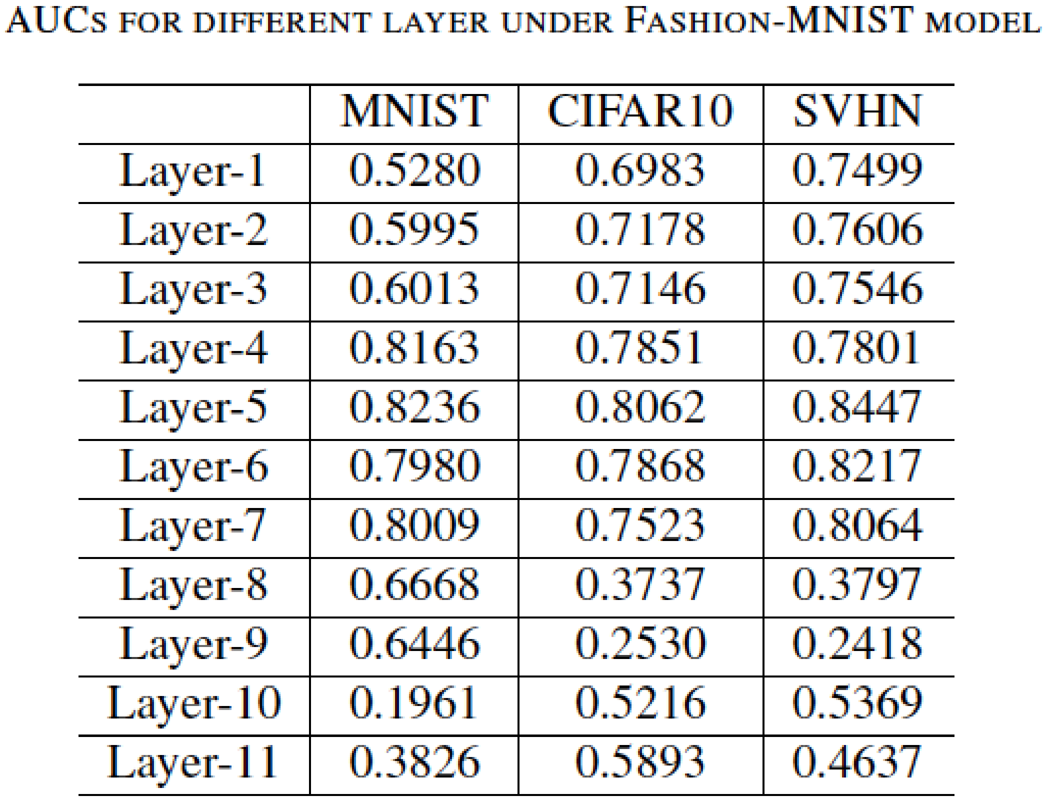
## Usefulness of One-Out Integration Strategy

1) the fingerprint information of different layers are complementary to each other thus combining fingerprints across multiple layers can make the in- and out-of distribution data more separable
2) the overall detection performance is significantly enhanced (with greater AUC scores) as compared to that of an individual layer, validating the necessity and usefulness of the one-out strategy in NeuralFP.



AUCs for different layer under MNIST model

| | FASHION | CIFAR10 | SVHN |
|---|---|---|---|
| Layer-1 | 0.7298 | 0.8817 | 0.8821 |
| Layer-2 | 0.8260 | 0.8848 | 0.8850 |
| Layer-3 | 0.6893 | 0.8394 | 0.8634 |
| Layer-4 | 0.4190 | 0.5340 | 0.7661 |
| Layer-5 | 0.4298 | 0.4821 | 0.6927 |
| Layer-6 | 0.8678 | 0.8902 | 0.8899 |
| Layer-7 | 0.8741 | 0.8955 | 0.8944 |



AUCs for different layer under Fashion-MNIST model

| | MNIST | CIFAR10 | SVHN |
|---|---|---|---|
| Layer-1 | 0.5280 | 0.6983 | 0.7499 |
| Layer-2 | 0.5995 | 0.7178 | 0.7606 |
| Layer-3 | 0.6013 | 0.7146 | 0.7546 |
| Layer-4 | 0.8163 | 0.7851 | 0.7801 |
| Layer-5 | 0.8236 | 0.8062 | 0.8447 |
| Layer-6 | 0.7980 | 0.7868 | 0.8217 |
| Layer-7 | 0.8009 | 0.7523 | 0.8064 |
| Layer-8 | 0.6668 | 0.3737 | 0.3797 |
| Layer-9 | 0.6446 | 0.2530 | 0.2418 |
| Layer-10 | 0.1961 | 0.5216 | 0.5369 |
| Layer-11 | 0.3826 | 0.5893 | 0.4637 |

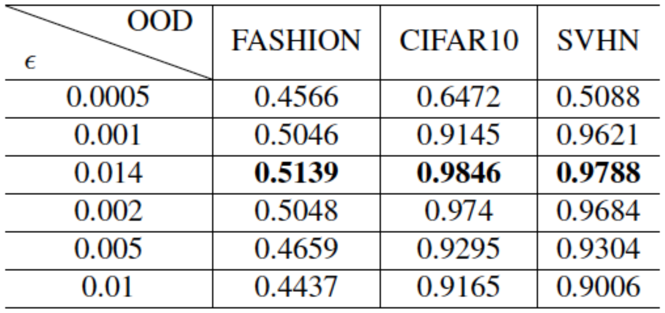
# NeuralFP: Experimental Results

## Advantageous over Previous works

NeuralFP shows significant advantages over previous methods in detecting OOD data

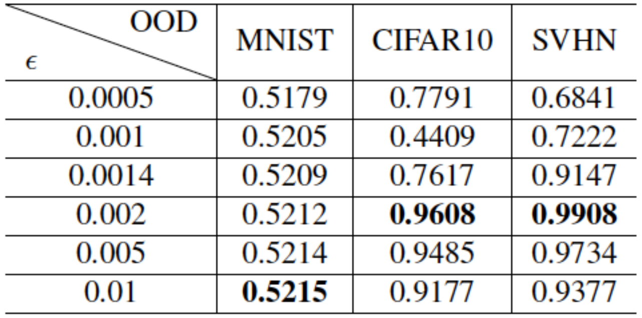
1) outperforms Mahalanobis [4] for all scenarios
2) outperforms ODIN [1] for all scenarios except for detecting SVHN data on the MNIST model

AUCs by using Mahalanobis [4] under MNIST model



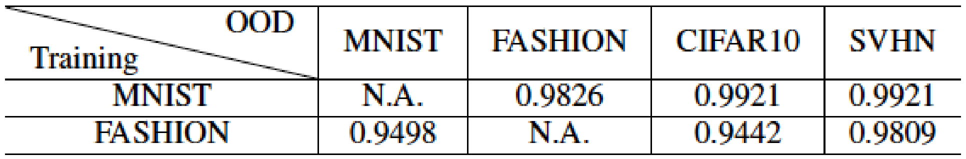
| OOD<br>$\epsilon$ | FASHION | CIFAR10 | SVHN |
|---|---|---|---|
| 0.0005 | 0.4566 | 0.6472 | 0.5088 |
| 0.001 | 0.5046 | 0.9145 | 0.9621 |
| 0.014 | **0.5139** | **0.9846** | **0.9788** |
| 0.002 | 0.5048 | 0.974 | 0.9684 |
| 0.005 | 0.4659 | 0.9295 | 0.9304 |
| 0.01 | 0.4437 | 0.9165 | 0.9006 |

AUCs by using Mahalanobis [4] under Fashion-MNIST model



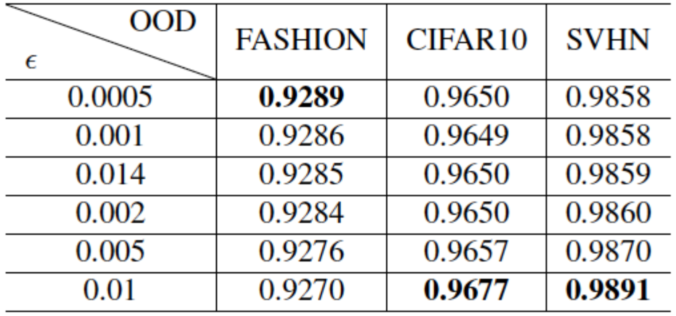
| OOD<br>$\epsilon$ | MNIST | CIFAR10 | SVHN |
|---|---|---|---|
| 0.0005 | 0.5179 | 0.7791 | 0.6841 |
| 0.001 | 0.5205 | 0.4409 | 0.7222 |
| 0.0014 | 0.5209 | 0.7617 | 0.9147 |
| 0.002 | 0.5212 | **0.9608** | **0.9908** |
| 0.005 | 0.5214 | 0.9485 | 0.9734 |
| 0.01 | **0.5215** | 0.9177 | 0.9377 |

AUCs of detecting various OOD records



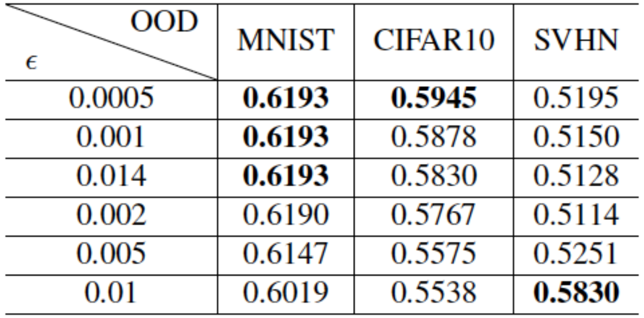
| OOD<br>Training | MNIST | FASHION | CIFAR10 | SVHN |
|---|---|---|---|---|
| MNIST | N.A. | 0.9826 | 0.9921 | 0.9921 |
| FASHION | 0.9498 | N.A. | 0.9442 | 0.9809 |

AUCs by using ODIN [1] under MNIST model



| OOD<br>$\epsilon$ | FASHION | CIFAR10 | SVHN |
|---|---|---|---|
| 0.0005 | **0.9289** | 0.9650 | 0.9858 |
| 0.001 | 0.9286 | 0.9649 | 0.9858 |
| 0.014 | 0.9285 | 0.9650 | 0.9859 |
| 0.002 | 0.9284 | 0.9650 | 0.9860 |
| 0.005 | 0.9276 | 0.9657 | 0.9870 |
| 0.01 | 0.9270 | **0.9677** | **0.9891** |

AUCs by using ODIN [1] under Fashion-MNIST model



| OOD<br>$\epsilon$ | MNIST | CIFAR10 | SVHN |
|---|---|---|---|
| 0.0005 | **0.6193** | **0.5945** | 0.5195 |
| 0.001 | **0.6193** | 0.5878 | 0.5150 |
| 0.014 | **0.6193** | 0.5830 | 0.5128 |
| 0.002 | 0.6190 | 0.5767 | 0.5114 |
| 0.005 | 0.6147 | 0.5575 | 0.5251 |
| 0.01 | 0.6019 | 0.5538 | **0.5830** |

[1] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” ICLR 2016

[4] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” NeuralPS 2018

# Discussions and Future Work

**NeuralFP can serve as a key technique of detecting outliers in practical edge computing scenarios.**

1) NeuralFP extracts representative information of the training set by constructing non-linear fingerprints of neural network models.
2) NeuralFP successfully distinguishes the differences between in-distribution data and OOD data, through carefully integrate the fingerprints across multiple layers
3) We have verified the effectiveness of NeuralFP on multiple real-world datasets, showed its advantages over existing detection methods, and provided useful guidelines for parameter selection in practice

**Future Directions**

1) Investigate the performance of NeuralFP in detecting various types of adversarial outliers
2) Integrate other deep generative models such as Generative Adversarial Networks (GAN)

# Thank You!

# Q&A