Deep Learning Lab Final Report: Out of distribution Detection method for a MNIST classification

Euijin Hong[1]
[1]Department of Electrical and Electronic Engineering, Yonsei University.

1 Introduction

This study proposes an algorithm called ***EuiNet*** for performing out of distribution (OOD) detection in MNIST classification that successfully incorporates the discriminator from GAN (Generative Adversarial Networks) and ODIN (Out-of-Distribution detector for Neural networks).

Previous researches have been independently investigating OOD detection tasks using models including ODIN [5] and generative-based anomaly detection approaches [6]. Both types of the algorithm show decent performance, while some inherent limitations remaining.

ODIN, known as one of the traditionally top-performing models for OOD detection, utilizes a pre-trained model on the existing training dataset and filters out OOD data with significantly higher accuracy compared to contemporary baseline models by employing input processing, temperature scaling, and softmax score thresholds. However, it has limitations in terms of flexibility in detection due to the use of a fixed threshold. Performing OOD detection solely based on a fixed threshold has its limits when the existing classifier predicts the OOD data with high-confidence softmax scores. Furthermore, ODIN has clear limitation that it solely depends on the performance of the classifier, where the model is optimized to classify the existing dataset, which is not for detecting the OOD data.

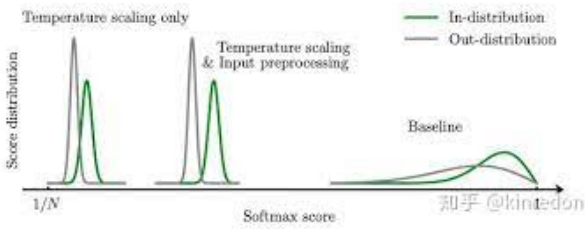

Figure 1: Softmax score distribution of the 'vanilla-ODIN' network. The distribution of the scores shows severe difference between those of in-disribution sets and out-distribution sets.

On the other hand, generative-based OOD detection has seen various proposed approaches, including attempts to utilize the discriminative ability of a discriminator for OOD detection. However, these attempts were in less interest and have generally shown low performance, which can be attributed to the following reasons.

First, the discriminator fails to exhibit the desired high discriminative performance in OOD detection tasks when it is trained in an adversarial manner. In other words, this indicates that the adversarial learning process of a GAN does not successfully train the discriminator. The reason lies in the basic fact that the training principle of GANs involves both the generator and discriminator sharing a single loss function and engaging in a min-max game. Consequently, this loss function does not converge well under such competitive dynamics, and even if it does converge, it results in *sub-optimal* values. As a result, the major issue becomes the inability to achieve improvements in both the generator and discriminator beyond each other's performance.

Additionally, when applying the output of the discriminator to OOD detection, an additional consideration arises: how the discriminator's prediction results should affect the inference results of the existing classifier. This depends on the specific task but ultimately entails finding an effective method to scale the discriminator's output.

This study proposes a new algorithm called ***EuiNet*** that combines these two distinct OOD detection approaches in a complementary manner. The

This is an extended abstract. The full paper is available at the Conditional Generative Adversarial Nets, Image-to-Image Translation with Conditional Adversarial Networks.

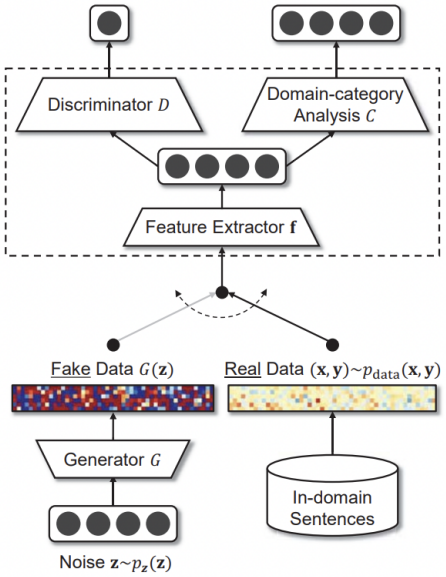

Figure 2: Generative adversarial network for out–of–domain sentence detection. Discriminator is used as detecting the authenticity of the provided data.

EuiNet model operates conceptually by complementing the limitations of each model. In terms of performance, the EuiNet model demonstrates exceptional performance in both classifying the existing MNIST data and performing OOD detection with high accuracy. The model is given an MNIST OOD detection task, where the model had been trained with handwritten digits from 1 to 9 and has to detect the OOD data - digit 0, while testing. The overall task is explained in the following figure.

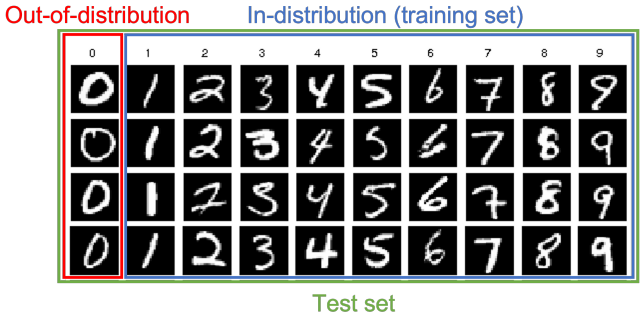

Figure 3: The settings of MNIST OOD detection task. MNIST handwritten digits of from 1 to 9 are given as a training set, which are regarded as an in-distribution set, and digit 0 serves as an out-of-distribution data that is additionally applied on the test set. Thus, the test set comprises of digits from 0 to 9.

2 Related Works

2.1 ODIN (Out-of-Distribution detector for Neural networks)

ODIN has significant advantages in its ability to effectively identify out-of-distribution samples with a simple neural network that has already completed training, without requiring any additional training. The key ideas behind ODIN are temperature scaling and input preprocessing.

Temperature Scaling is a method proposed by Professor Hinton in Knowledge Distillation [3], which is also used to calibrate prediction confidence in classification tasks. During training, the temperature scaling parameter, T, is set to 1, and the training proceeds accordingly. In the testing phase, to obtain the predicted probabilities of a given input, the softmax scores are calculated by dividing each logit by T. This approach can be implemented in OOD detection to increase the distance between the softmax scores of in-distribution and out-of-distribution samples, making it easier to distinguish the out-of-distribution samples. Here, the temperature scaling parameter, T, is a hyper-parameter, and selecting an appropriate value is crucial.

Input Preprocessing is derived from the Fast Gradient Sign Method (FGSM) proposed in the seminal paper that initiated adversarial attacks [1]. It utilizes the concept of backpropagation to minimize the loss but in the opposite direction. By calculating the gradient in the direction that increases the loss, a minimal perturbation can be obtained. Adding this perturbation to the input lowers the softmax score for the true label. In the ODIN paper, on the other hand, reverses this process by *subtracting* a minimal perturbation from the input. This aims to increase the softmax score for the given input and strengthen the predictions for in-distribution samples, thereby better separating them from out-of-distribution samples. The perturbation magnitude parameter, epsilon, is a hyper-parameter, and selecting an appropriate value is important.

Through these two processes, the maximum softmax score is obtained, and simple thresholding is used to determine whether a sample is in distribution or out-of-distribution. The threshold value used in this process is also a hyper-parameter. Thus, a total of three hyper-parameters contribute to the overall performance.

In the ODIN paper, the authors selected the hyper-parameters at the point where the True Positive Rate (TPR) for in-distribution samples reached 95%. Through experimentation, they discovered the optimal values of $T = 1000$ and $epsilon = 0.0012$ and fixed them to evaluate the performance.

2.2 Out-of-domain Detection based on Generative Adversarial Network

The proposal in this paper involves utilizing a GAN for OOD detection, which consists of a generator (G) and a discriminator (D). G generates artificial data to deceive D, which aims to distinguish real data from the generated artificial data. Since GAN learning doesn't require labels, it is considered as an unsupervised algorithm.

$$min_G max_D V(D,G) = \mathrm{E}_{x\sim p_{data}}[logD(x)] + \mathrm{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \quad (1)$$

The GAN framework can be seen as a minimax two-player game, where G minimizes V(D, G) and D maximizes V(D, G). Researchers suggest using GANs to create a one-class classifier for OOD detection. By training G to generate data similar to in-distribution data and training D to classify both real in-distribution data and fake data generated by G, D is expected to reject OOD data. Hence, researchers consider the low confidence score assigned by D to an input data as evidence of it being an OOD data.

Additionally, GANs often suffer from mode collapse, where G generates samples with low variance. To address this issue, researchers removed biases in the generator and emphasize the use of weights in generating data.

Another technique employed in this paper is feature matching[7]. Rather than directly maximizing the discriminator's output, the new objective requires the generator to generate data that matches the statistics of real data, using the discriminator to specify the relevant statistics. Consequently, G is trained to generate high-variance datas $(G(z))$ using an additional objective function.

3 The Model

In order to solve the existing limitations of ODIN and generative-based OOD detection models, an originally formed architecture which will be denoted as ***EuiNet*** is proposed. The overall structure of EuiNet model and its surrounding network models are illustrated in Figure 4 and Figure 5, where each of the figures correspond to the flow while training and testing, respectively.

The key idea of EuiNet can be listed as follows.

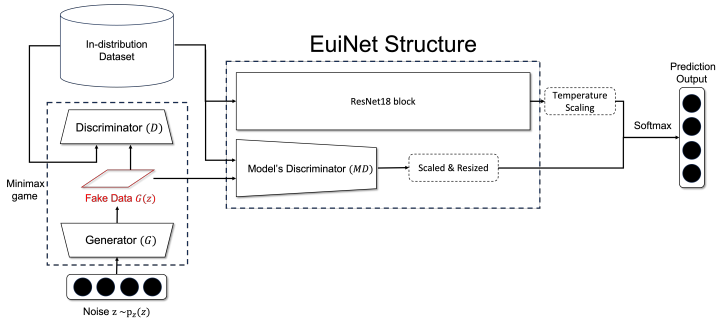

Figure 4: Structure of EuiNet and its surrounding networks with the corresponding flow of data, while training.

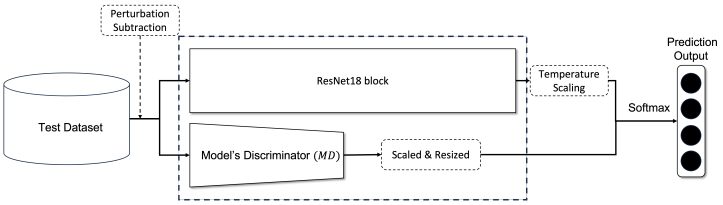

Figure 5: Structure of EuiNet and its surrounding networks with the corresponding flow of data, while testing.

1. EuiNet Structure: A unified model structure comprised with a model discriminator and a Residual Net-based classifier [2] trained by a single objective function and a single optimizer. The objective function for the model discriminator guarantees the convergence to the optimal value.
2. Introduction of two separate discriminators: Inspired by off-policy control in reinforcement learning (such as Q-learning [4]), two separate discriminators are used: the model discriminator & GAN discriminator. This is a solution for sub-optimal convergence of the existing objective function of GAN. Each of the discriminators have their respective objectives such as:
   - model discriminator - determining the authenticy of the input and detect the OOD
   - GAN discriminator - adversarial training with the generator by pursuing the minimax game objective function.
3. Proper scaling method for the model discriminator's output to be used as an input to the identical softmax function with the classifier's output. This serves as an *active threshold* for the classifier's softmax scores, which thus can be an improvement of the fixed threshold used in ODIN.
4. Temperature scaling is used during training and testing, and perturbation subtraction is applied while testing the model, which result a strengthened softmax prediction scores of the classifier. These two methods can be considered as almost identical to those used in ODIN.

The unified model of EuiNet structure allows to be trained by a single objective function and a single optimizer, which is more simple and effective since it is forwarded and trained in a single sweep. The objective function of EuiNet is:

$$\begin{aligned}\mathcal{L}_{EuiNet}(\vec{x_r},\vec{x_f}) = \frac{1}{N}\sum_{i=1}^{N} & y_i \cdot log(\hat{y_i}) + (1-y_i)\cdot log(1-\hat{y_i}) \\ & + d_r \cdot log(\hat{d_r}) + (1-d_r)\cdot log(1-\hat{d_r}) \\ & + d_f \cdot log(\hat{d_f}) + (1-d_f)\cdot log(1-\hat{d_f})\end{aligned} \quad (2)$$

where y_i are one-hot encoded label, $\hat{y_i}$ are prediction outputs of the classifier for the real data, d_r and d_f are labels for real and fake data that correspond to 0 and 1, and $\hat{d_r}$ and $\hat{d_f}$ are prediction outputs of real and fake data.

Thus, a generator is required for generating a fake data from a random noise latent vector, and a discriminator which serves as a pair of this generator for playing a minimax game that leads to adversarial training and increased quality of the generator's output. The training follows the objective function as:

$$min_G max_D V(D,G) = \mathrm{E}_{x\sim p_{data}}[logD(x)] + \mathrm{E}_{z\sim p_z(z)}[log(1-D(G(z)))] \quad (3)$$

We do not use the discriminator trained in this process since the objective function of GAN does not guarantee a conversion to the optimal value for the discriminator. Inspired by the philosophy of off-policy learning of reinforcement learning, in which the agent's behavior and learning is based on separate policies, two discriminators are utilized separately where one competes with the generator while other learns and converges to the optimal solution of the OOD-discriminating task.

As we can see in equation (2), real labels and fake labels are set to 0 and 1, respectively. This enables the model discriminator to result 0 as an output for in-distribution data and 1 for out-of-distribution data, where the function works similar to binary classification of each labels from 1 to 9 of the classifier. In other words, the score of the discriminator and the classifier can thus be incorporated through proper scaling and resizing. The proposed scaling process follows:

$$h_a = \alpha \cdot h + \gamma \quad (4)$$

where α and γ are the hyper-parameters that determine the minimum value and maximum value of the range of the scaled output of the model discriminator h_a. The model discriminator's original output is h, which has the range from 0 to 1. The values of hyper-parameters α and γ depend on the range and distribution of the classifier's scores which are dependent of the number of classes and the structure of the classifier model.

By scaling the discriminator's output score, the output remains in the same range as that of the classifier's output before entering the softmax function. Thus, we can concatenate the two scores and let them enter the same softmax function, so that we can obtain the prediction result of MNIST labels 0 to 9.

Temperature scaling is applied on both training and inference stage, and image processing based on perturbation subtraction is applied only in the inference stage. Both methods plays a crucial role for increasing the gap of the predicted softmax scores of each label. Temperature value T and perturbation subtraction coefficient ε are also hyper-parameters.

4 Performance

4.1 Hyperparameter Setting

While training and testing the EuiNet model, different types of hyperparameters are used. During the training stage, the hyperparameters of the EuiNet model are set as follows:

- Learning rate / β parameters for Adam Optimizer:
  - EuiNet Model - learning rate: 2×10^{-4}, β_1: 0.5, β_2: 0.599
  - Generator - learning rate: 10^{-4}, β_1: 0.35, β_2: 0.699
  - Discriminator - learning rate: 10^{-4}, β_1: 0.35, β_2: 0.699
- Temperature scaling factor T: 10^{-3}

During the inference stage, the hyperparameters of the EuiNet model are set as follows:

- Model discriminator scaling factor - α: 0.004, γ: 0.008
- Temperature scaling factor T: 10^{-3}
- Perturbation subtraction coefficient - ε: 0.3

4.2 Result Analysis

Fake data created by the generator as in Figure 6 were used for training the model discriminator and adversarial training with the discriminator. Generator outputs show random and indistinguishable properties compared to real MNIST digits.

From Figure 8 to Figure 17 shows the prediction results of EuiNet in terms of softmax scores distributions. Scores of each label show that their maximum values are be clustered on their corresponding label, which indicates that the model predicts the input data to be that label. This indicates that the prediction result of the model discriminator is well-integrated with the classifier's output, and makes up a synchronized softmax score output. Furthermore, the softmax score shows high reliability for the model's principle, where viewing that the softmax score of label 0 by the model discriminator can be considered as an *active threshold* of the classifier's softmax score.

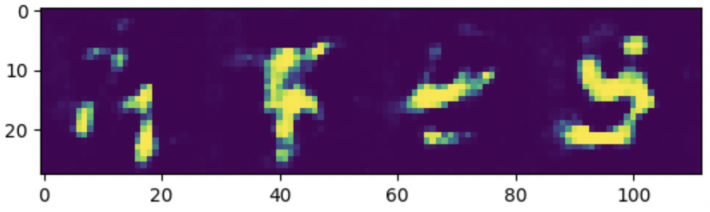

Figure 6: Fake MNIST data outputs of the generator.

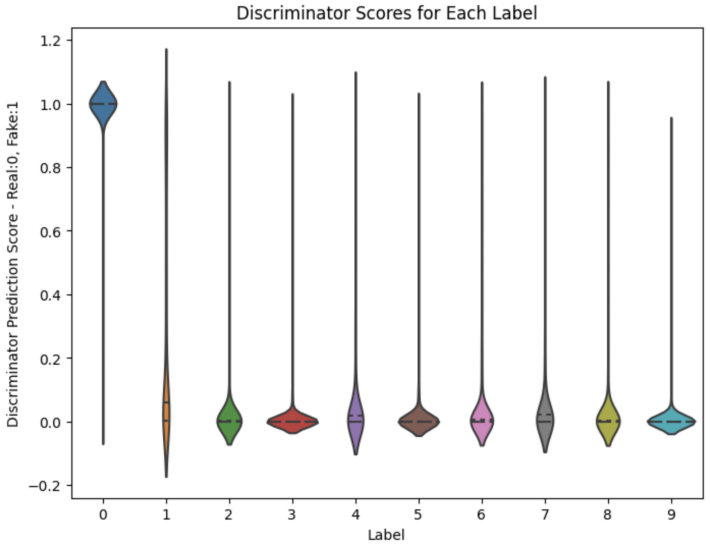

Figure 7: Model discriminator's prediction scores for each label, ranging from 0 to 1 before scaling. The prediction result of label 0 shows clustered distribution around score 1, while results of other labels show clustered distribution around score 0. This indicates that the model discriminator is well performing the OOD detection task.

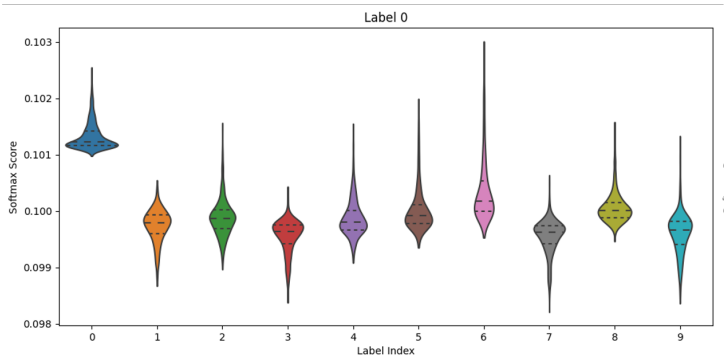

Figure 8: Softmax score output of EuiNet model on label 0.

Figure 7 and Figure 19 indicates that the model discriminator's output represents high confidence of predicting whether the data is OOD or not, where its median value shows 1.0 in label 0 but nearly zero in other labels.

Also, Figure 18 and Figure 19 shows the Max-2-Max scores (the difference between the maximum and second-maximum value) for each label. It is observed that in the softmax score outputs of data that correspond to label 0, the Max-2-Max scores tend to be relatively low, compared to those of the labels from 1 to 9. From this data, we can infer that the performance

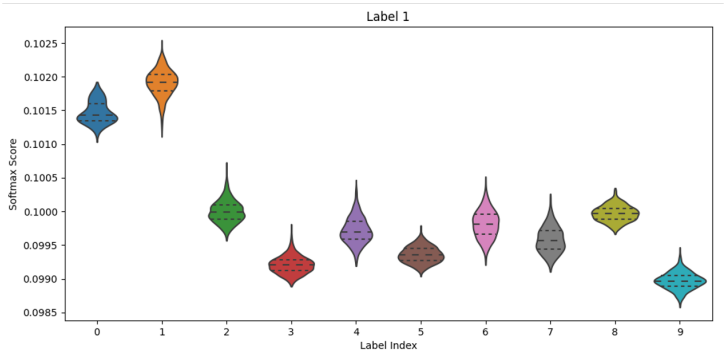

Figure 9: Softmax score output of EuiNet model on label 1

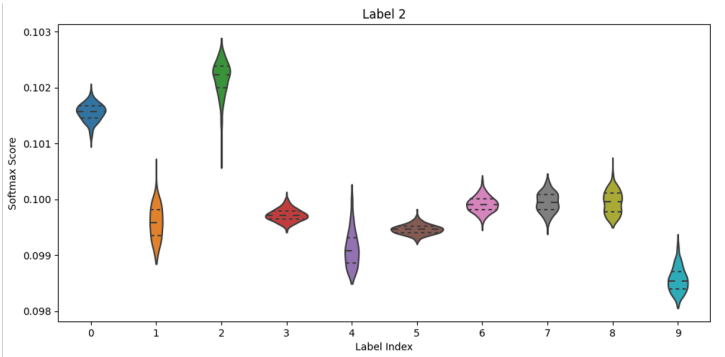

Figure 10: Softmax score output of EuiNet model on label 2

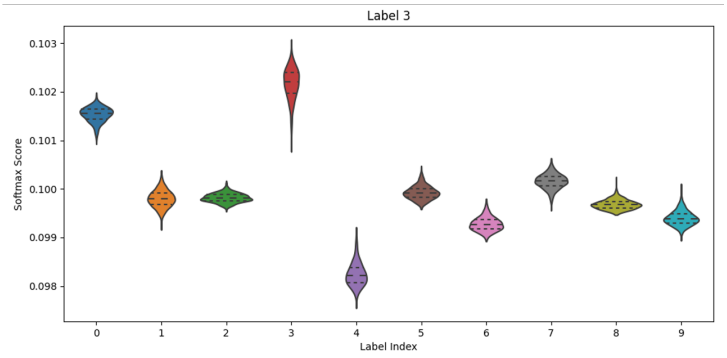

Figure 11: Softmax score output of EuiNet model on label 3

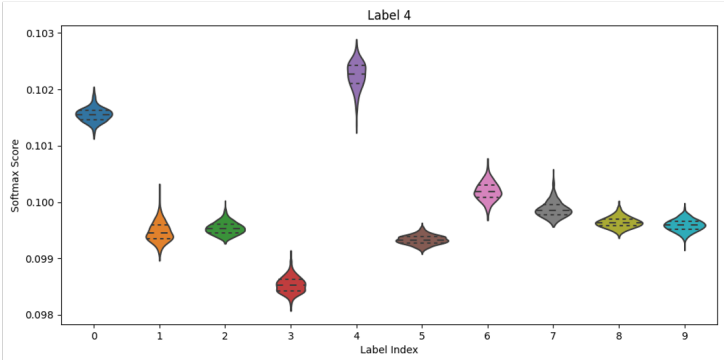

Figure 12: Softmax score output of EuiNet model on label 4

of the classifier is very decent and that it will show high accuracy in high probability.

Finally, we can see that the model achieves high accuracy for both precisely predicting the OOD and accurately classifying the labels of in-distribution MNIST digits. Confusion matrix and accuracy both shows high precision and recall score of the prediction of EuiNet model. Both results are respectively shown in Figure 20 and Figure 21. It is notable that the

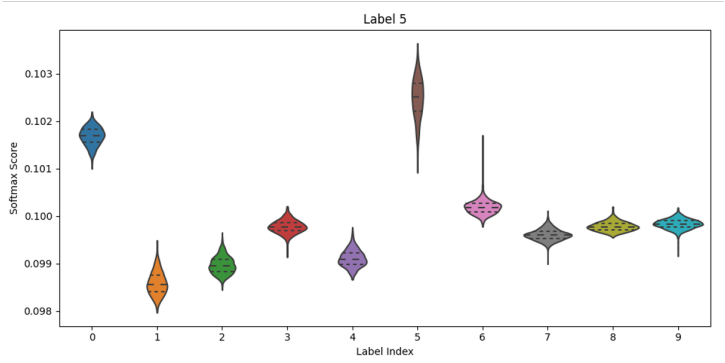

Figure 13: Softmax score output of EuiNet model on label 5

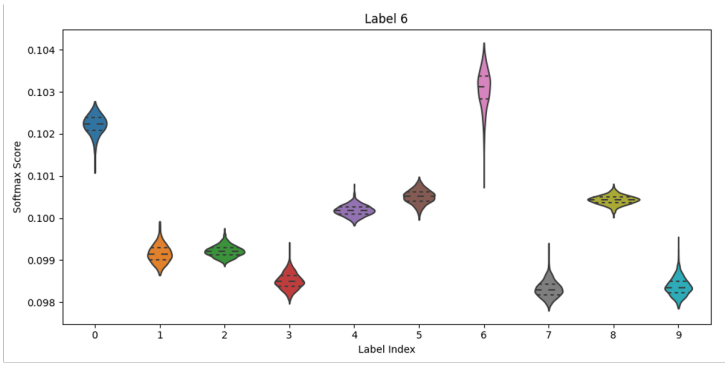

Figure 14: Softmax score output of EuiNet model on label 6

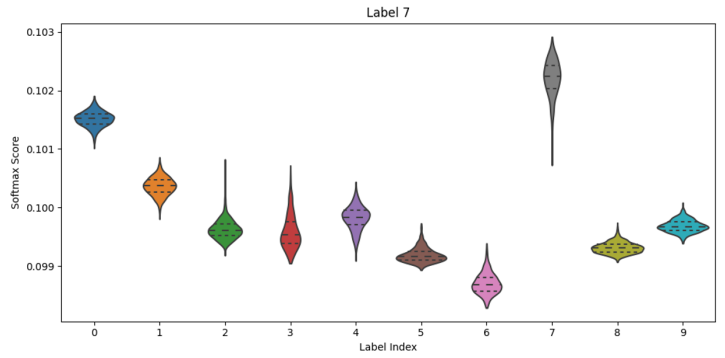

Figure 15: Softmax score output of EuiNet model on label 7

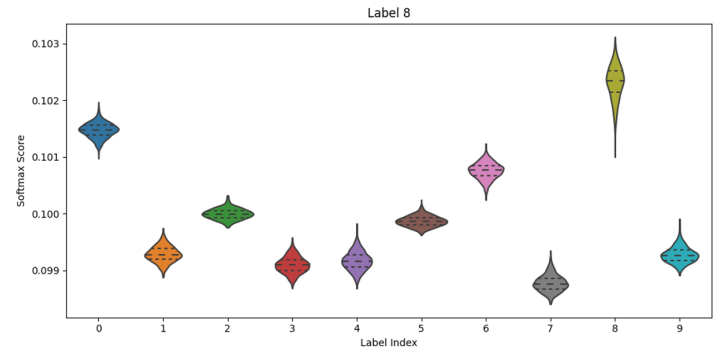

Figure 16: Softmax score output of EuiNet model on label 8

prediction result of the model discriminator does not corrupt the prediction scores of the classifier, thus results nearly 100% accuracy on in-distribution data, while successfully detecting the OOD data with nearly 97% accuracy.

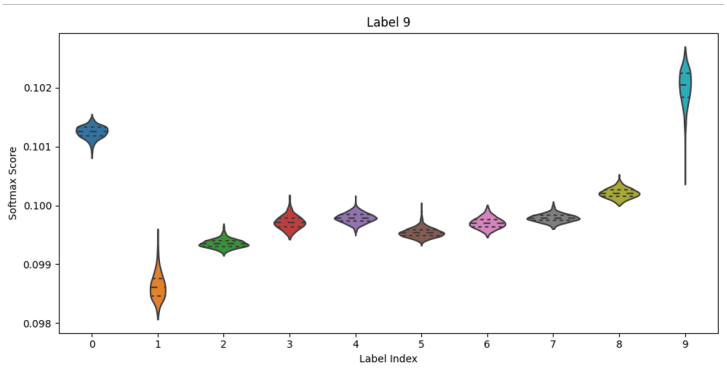

Figure 17: Softmax score output of EuiNet model on label 9

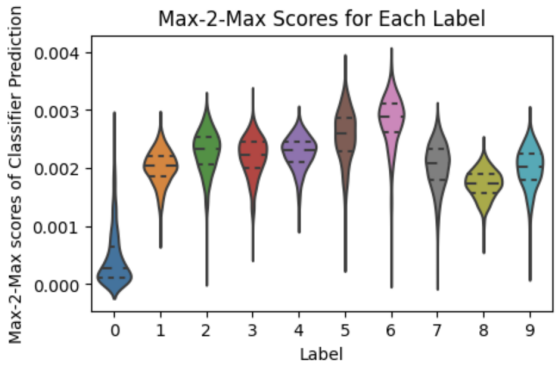

Figure 18: Max-2-Max score of the softmax output of EuiNet model on each label

Label	Discriminator Score Medians	Max-2-Max Score Medians
0	1.0	2.7290732e-04
1	6.051204e-02	2.034545e-03
2	5.6319936e-05	2.3321658e-03
3	1.9545769e-05	2.2342727e-03
4	0.00036736636	2.30439e-03
5	5.046315e-05	2.589155e-03
6	2.382453e-04	2.8864443e-03
7	9.088801e-04	2.0839795e-03
8	5.7197027e-05	1.7445087e-03
9	7.878914e-06	2.0312928e-03

Figure 19: Medians of model discriminator's prediction score and the Max-2-Max score of the softmax output of EuiNet model.

5 Conclusion

In this study, the proposed EuiNet model demonstrated the potential integration of the conceptual ODIN and discriminator-based generative OOD detection methods, and its experimentally confirmed outstanding performance. The hyperparameters identified in this research do not guarantee optimal results. This implies that there is ample room for performance improvement through fine-tuning of the hyperparameters. Furthermore, when training the model based on the MNIST dataset, it was observed that the model achieved rapid and superior performance with a small dataset size and low complexity, reaching the overfitting stage in a limited number of training iterations. Therefore, it is anticipated that the OOD detection in more complex datasets, rather than solely within this dataset, can be effectively conducted.

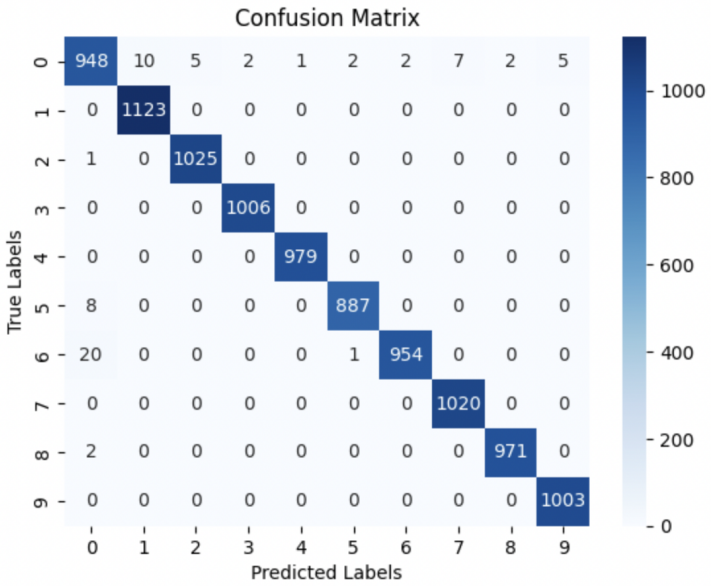

Figure 20: Confusion matrix of EuiNet model.

Label	0	1~9
Accuracy	96.9353	99.5853

Figure 21: Accuracy of EuiNet model

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[4] Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, and Jong Wook Kim. Q-learning algorithms: A comprehensive classification and applications. *IEEE access*, 7:133653–133667, 2019.

[5] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[6] Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, 2018.

[7] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.