DIRNET: DOMAIN-AGNOSTIC INCREMENTAL ROUTING NETWORK

Technical Report

Jungyu Choi[1]*, Seowoo Kim*[1]*, Sungbin Im*[2]

[1] Dept. of IT Engineering, Soongsil University, Dongjak-gu, Seoul, Republic of Korea
cjg@soongsil.ac.kr, bananana1121@gmail.com
[2] School of Electronic Engineering, Soongsil University, Dongjak-gu, Seoul, Republic of Korea
sbi@ssu.ac.kr

ABSTRACT

This technical report presents DIRNET, our submission to DCASE 2026 Challenge Task 7 on domain-agnostic incremental sound classification. The task requires a model to learn sequential acoustic domains (D1, D2, and D3) while preserving previously acquired knowledge and performing inference without domain labels. To address this challenge, we propose a unified CNN14-based framework with two domain-routing strategies. DIRNET-OOD performs soft domain fusion using class-wise prototype distances and a D1 out-of-distribution residual prior. DIRNET-SCR estimates domain responsibility from confidence, prediction margin, and entropy after domain-specific calibration, followed by soft probability mixture across domain paths. Experimental results show that both systems substantially outperform the official baseline, demonstrating the importance of effective scoring and fusion strategies for domain-agnostic incremental sound classification.

***Index Terms*—** DCASE 2026 Task 7, domain-agnostic incremental learning, out-of-distribution, prototype routing, score-calibrated routing, soft probability mixture

1. INTRODUCTION

DCASE 2026 Task 7 addresses domain-agnostic incremental learning, where acoustic domains are learned sequentially in the order D1 $\rightarrow$ D2 $\rightarrow$ D3 while the target sound classes remain fixed [1]. During evaluation, test samples may originate from any domain, but domain labels are not provided. Therefore, the system must classify the input sound while determining how domain-specific knowledge should be utilized under unknown-domain conditions.

This setting is challenging because acoustic characteristics vary across domains and only current-domain data are available during incremental learning, making the model vulnerable to catastrophic forgetting. The official baseline employs domain-specific batch normalization and entropy-based path selection [2], achieving a D2/D3 macro-average accuracy of 52.5%, which highlights the difficulty of domain-agnostic inference.

To address this problem, we propose Domain-Agnostic Incremental Routing Network (DIRNET), a CNN14-based framework with domain-specific classification paths. Two scoring strategies are investigated. The first strategy employs out-of-distribution (OOD) residual routing, where D2/D3 class-wise prototype distances and a D1 residual prior are used for soft fusion. The second strategy employs score-calibrated routing (SCR), which combines confidence, prediction margin, and entropy to produce calibrated probability mixtures. Based on these strategies, we construct two system variants, DIRNET-OOD and DIRNET-SCR, respectively. Experimental results demonstrate the importance of domain-aware scoring and fusion for domain-agnostic incremental sound classification.

The remainder of this report is organized as follows. Section 2 describes the proposed framework and scoring strategies. Section 3 presents the experimental setup, Section 4 reports the results, and Section 5 concludes the report.

2. PROPOSED METHOD

2.1. Common CNN14-based model architecture

DIRNET extracts a 2048-dimensional clip-level embedding from a log-mel representation using a CNN14/PANNs-style backbone [3]. The backbone consists of six Conv(3$\times$3)–BN–ReLU–Conv(3$\times$3)–BN–ReLU–AvgPool blocks followed by pooling and a domain-specific classifier head for 10 sound classes.

The framework maintains separate D1, D2, and D3 paths. The D1 path is initialized from the provided checkpoint, while D2 and D3 are sequentially adapted from the preceding stage. During evaluation, all domain paths are evaluated in parallel, and the final prediction is determined by the corresponding domain-routing strategy.

2.2. Incremental training and selective plasticity

At the D2 stage, the model is initialized from the D1 checkpoint and adapted to D2 data. In DIRNET-OOD, branch isolation is strongly applied to preserve the D1 branch while learning the D2 branch and the corresponding classifier/routing components. After training, D2 class-wise prototypes are computed using a clean forward pass with augmentation disabled. At the D3 stage, the model adapts to D3 based on the learned D2 branch or checkpoint, and D2/D3 prototypes are recomputed after D3 training. In DIRNET-SCR, teacher-student knowledge distillation and weak feature distillation are used to reduce forgetting during incremental adaptation [4, 5]. Before training stage t, the previous-stage model is fixed as a teacher, and a KL-divergence loss is added so that the current student model produces similar outputs to the teacher on available previous-domain heads. The overall loss combines the current-domain cross-entropy loss, logit distillation, and weak feature distillation.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (RS2024-00355759, Factory facility anomaly detection and localization technology based on multi-noise removal).

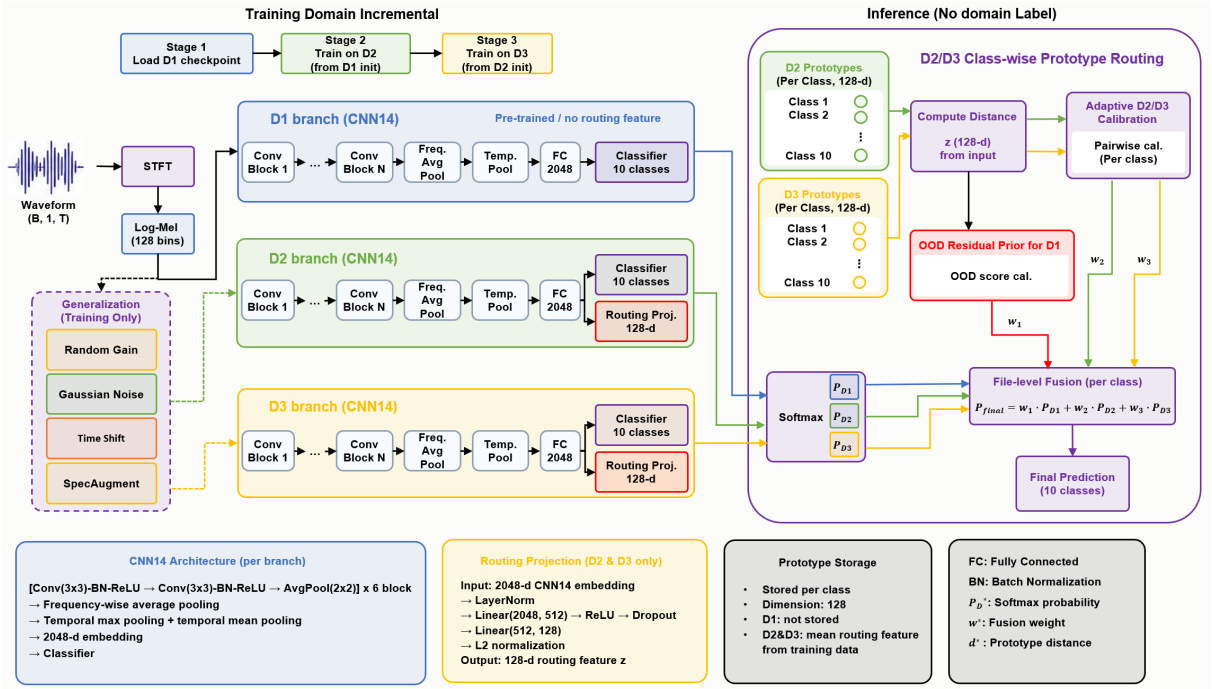

Figure 1: DIRNET-OOD architecture.

In DIRNET-SCR, for D3 adaptation, we additionally use D2-anchored selective channel plasticity. At the end of D2 training, D2 activation statistics are accumulated from Conv5 and Conv6 and stored in the D2 checkpoint as channel masks. During D3 training, these stored masks are used to identify the D2-important channel set K_{D2}. In the final DIRNET-SCR system, K_{D2} is defined as the set of channels with the top 60% activation magnitudes in each layer. Before the optimizer update, channel-wise gradients are rescaled as

$$\tilde{g}_j = \begin{cases} \rho g_j, & j \in K_{\mathrm{D2}}, \\ \lambda g_j, & j \notin K_{\mathrm{D2}}, \end{cases} \tag{1}$$

where g_j is the original gradient of channel j. We use $\rho = 0.25$ to reduce updates on D2-important channels and $\lambda = 1.1$ to preserve adaptation capacity for the remaining channels. This selective channel plasticity mitigates damage to D2 representations during D3 training [6].

2.3. DIRNET-OOD: prototype residual routing

DIRNET-OOD performs prototype-based routing using class-wise D2/D3 prototypes in a 128-dimensional normalized feature space derived from the 2048-dimensional CNN14 embedding. As shown in Fig. 1, D2/D3 domain evidence is estimated from prototype distances, while the D1 path is represented by an out-of-distribution residual prior. The final prediction is obtained through soft fusion of branch probabilities.

Let x denote an input audio clip, $r \in \mathbb{R}^{128}$ the routing feature, and $p_{d,c}$ the prototype of class c in domain $d \in \{D2, D3\}$. Let $P_d(c|x)$ denote the class posterior probability produced by domain branch d. The prototype-based distance to domain d is computed as

$$\zeta_d(x) = \sum_{c=1}^{C} P_d(c|x) \frac{1 - r \cdot p_{d,c}}{\|r\| \, \|p_{d,c}\|}. \tag{2}$$

Instead of relying only on the hard predicted class, the distance is computed as a probability-weighted average over all class prototypes, allowing multiple classes to contribute to the routing decision when the classifier output is uncertain.

Since D1 raw data are unavailable, D1 prototypes cannot be computed in the same way as D2/D3 prototypes. Therefore, DIRNET-OOD does not directly include D1 in the prototype-distance computation and instead models it as an OOD residual prior [7]. The raw routing scores for D2 and D3 are defined from prototype distances as

$$s_d^{\mathrm{OOD}} = \exp(-\zeta_d / T), \quad d \in \{D2, D3\}, \tag{3}$$

where T is the temperature parameter for prototype distance. A fixed residual prior is assigned to the D1 branch:

$$s_{D1}^{\mathrm{OOD}} = \lambda_{OOD}. \tag{4}$$

The initial routing weights of the three branches are normalized as

$$w_d^{\mathrm{OOD}} = \frac{s_d^{\mathrm{OOD}}}{s_{D1}^{\mathrm{OOD}} + s_{D2}^{\mathrm{OOD}} + s_{D3}^{\mathrm{OOD}}} \quad d \in \{D1, D2, D3\}. \tag{5}$$

If an input is far from both D2 and D3 prototypes, the D2/D3 scores decrease and the relative contribution of the D1 residual weight increases after normalization. Conversely, if the input is close to a specific D2 or D3 prototype, the corresponding domain weight increases and the D1 contribution decreases. When D2 and D3 distances are very close, a simple softmax can assign nearly equal D2:D3 weights. To address this issue, we use adaptive D2/D3 pairwise calibration. We first define the total D2+D3 mass as

$$m_{D2+D3} = w_{D2} + w_{D3}. \tag{6}$$

A pairwise temperature parameter $T_{D2,D3}$ is introduced to control how strongly the distance difference between D2 and D3 affects the

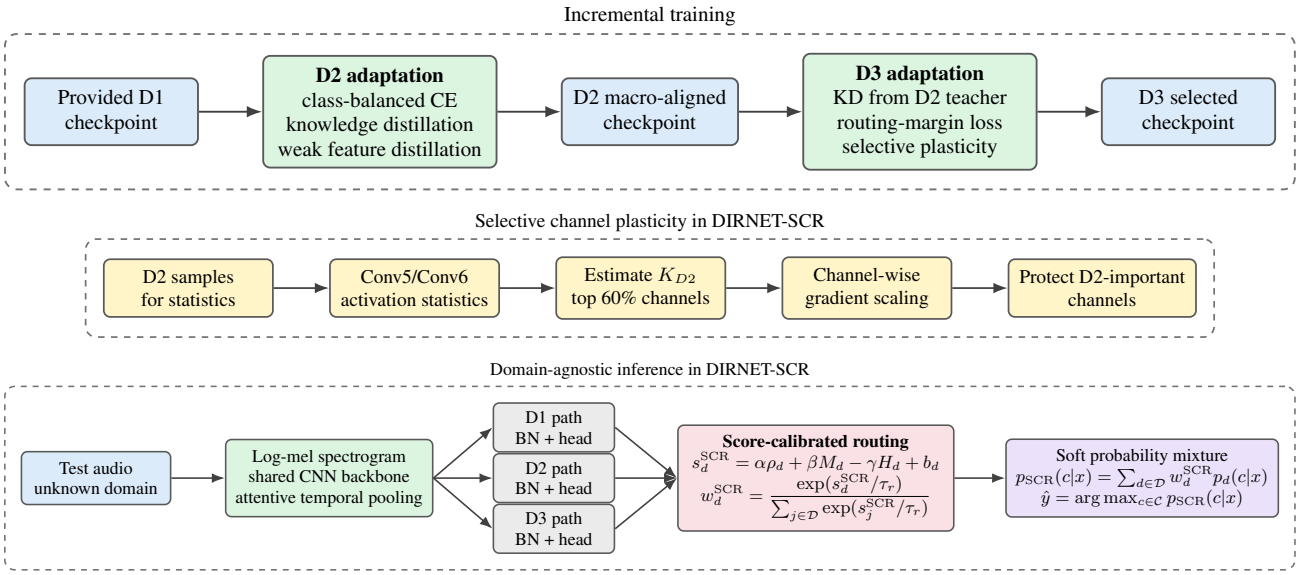

Figure 2: Overview of DIRNET-SCR. During incremental training, the model starts from the provided D1 checkpoint and is sequentially adapted to D2 and D3. Before D3 adaptation, D2-important channels are estimated from D2 activation statistics, and selective channel plasticity is applied in DIRNET-SCR to reduce updates on the protected channels. During inference, all domain-specific paths are evaluated in parallel. Their calibrated outputs are converted into SCR routing scores and soft mixture weights, and the final class prediction is obtained from the mixed probability $p_{\mathrm{SCR}}(c|x)$.

routing weights. Smaller values of $T_{D2,D3}$ produce sharper D2/D3 assignments, whereas larger values yield softer distributions. Then only the D2:D3 ratio is recalibrated using a temperature-scaled pairwise softmax:

$$q_d = \frac{e^{(-\zeta_d/T_{D2,D3})}}{e^{(-\zeta_{D2}/T_{D2,D3})} + e^{(-\zeta_{D3}/T_{D2,D3})}}, \quad d \in \{D2, D3\}. \tag{7}$$

The pairwise temperature $T_{D2,D3}$ is set using the development-set median of the distance gap between D2 and D3:

$$\Delta_{D2,D3} = |\zeta_{D2} - \zeta_{D3}|. \tag{8}$$

Given a target pairwise confidence η, the temperature is computed as

$$T_{D2,D3} = \frac{\mathrm{median}(\Delta_{D2,D3})}{\log(\eta/(1-\eta))}. \tag{9}$$

We then apply $T_{D2,D3} \leftarrow \mathrm{clip}(T_{D2,D3}, T_{\min}, T_{\max})$. The final D2/D3 routing weights are

$$\hat{w}_{D2}^{\mathrm{OOD}} = m_{D2+D3} \cdot q_{D2}, \quad \hat{w}_{D3}^{\mathrm{OOD}} = m_{D2+D3} \cdot q_{D3}. \tag{10}$$

Since this procedure recalibrates only the relative ratio between D2 and D3, the D1 routing weight remains independent of the D2+D3 mass redistribution.

2.4. DIRNET-SCR: score-calibrated soft probability routing

DIRNET-SCR estimates inference-time domain-path contribution using output score statistics instead of prototype distance. Fig. 2 shows the architecture of DIRNET-SCR. For each input audio clip, the D1, D2, and D3 domain-specific paths are evaluated in parallel. Each path produces calibrated class probabilities, from which confidence-related statistics such as maximum confidence, top-1/top-2 margin, and entropy are computed. These statistics are then combined into a routing score, and the final class probability is obtained by a soft mixture of the calibrated path probabilities.

Let $\mathcal{D} = \{D1, D2, D3\}$ denote the set of domain paths, and let $\mathcal{C} = \{1, \ldots, C\}$ denote the set of class indices, where C is the number of classes. For an input audio clip x, let $\mathbf{z}_d(x)$ be the class-logit vector produced by domain path $d \in \mathcal{D}$. Domain-specific temperature scaling is first applied [8]:

$$p_d(c|x) = \mathrm{softmax}(\mathbf{z}_d(x)/T_d)_c, \quad c \in \mathcal{C}. \tag{11}$$

Here, T_d compensates for confidence-scale differences across domain paths. From each calibrated probability vector, we compute maximum confidence ρ_d, top-1/top-2 margin M_d, and predictive entropy H_d:

$$\rho_d = \max_{c\in\mathcal{C}} p_d(c|x), \tag{12}$$

$$M_d = p_d(c_d^{(1)}|x) - p_d(c_d^{(2)}|x), \tag{13}$$

$$H_d = -\sum_{c\in\mathcal{C}} p_d(c|x) \log p_d(c|x), \tag{14}$$

where $c_d^{(1)}$ and $c_d^{(2)}$ are the top-1 and top-2 predicted classes of path d, respectively. The SCR routing score is defined as

$$s_d^{\mathrm{SCR}} = \alpha\rho_d + \beta M_d - \gamma H_d + b_d, \tag{15}$$

where α, β, and γ are the confidence, margin, and entropy weighting coefficients, respectively, with $\alpha = 1.0$ fixed for both submissions. Maximum confidence represents the strongest class confidence, margin reflects the separation between the top and second classes, and entropy captures the uncertainty of the overall class distribution. The term b_d is a domain-specific bias; in the second submission, $b_{D1} > 0$ is used to increase the contribution of the D1 path. Unlike the entropy-only baseline or hard score routing,

the final DIRNET-SCR submission does not select a single path. Instead, it converts SCR routing scores into soft mixture weights:

$$w_d^{\mathrm{SCR}} = \frac{\exp(s_d^{\mathrm{SCR}}/\tau_r)}{\sum_{j\in\mathcal{D}} \exp(s_j^{\mathrm{SCR}}/\tau_r)}, \tag{16}$$

where τ_r is the routing temperature. The final class probability is obtained by taking a weighted sum of calibrated probabilities from the D1, D2, and D3 paths:

$$p_{\mathrm{SCR}}(c|x) = \sum_{d\in\mathcal{D}} w_d^{\mathrm{SCR}} p_d(c|x), \qquad \hat{y} = \arg\max_{c\in\mathcal{C}} p_{\mathrm{SCR}}(c|x). \tag{17}$$

Therefore, D1 is not excluded from routing. Although D1 development labels are unavailable and direct calibration is difficult, the D1 path is retained as a soft mixture candidate because the hidden evaluation set may include D1-like samples. DIRNET-SCR-S1 is the setting with the highest local D2/D3 macro accuracy, whereas DIRNET-SCR-S2 is a D1-aware calibrated submission that intentionally increases the D1 mixture weight.

Table 1 summarizes the two submitted systems. Both systems share the same CNN14-based incremental framework and differ only in the scoring strategy used during domain-agnostic inference. DIRNET-OOD uses prototype-based residual fusion, whereas DIRNET-SCR uses score-calibrated soft probability mixture.

Table 1: Overview of the submitted systems.

Model	Scoring type	Final decision
DIRNET-OOD	Prototype residual	Soft fusion
DIRNET-SCR	Score-calibrated	Soft probability mixture

3. EXPERIMENT SETUP

The experiments follow the DCASE 2026 Task 7 protocol, where the model starts from the provided D1 checkpoint and sequentially learns the D2 and D3 development domains. Input audio is processed as a 32 kHz mono waveform and converted into a baseline-compatible log-mel representation with 64 mel bands, a 1024-sample window, and a 320-sample hop size. No external audio data, labels, or pretrained embedding models are used.

Training augmentation includes random gain, Gaussian noise, time shift, and SpecAugment [9]. For prototype recomputation, augmentation is disabled to obtain stable routing references. Since D1 development labels are unavailable, checkpoint selection and routing calibration are performed using the official-style macro average on the local D2/D3 split, where class-wise accuracies are first averaged within each domain and then across domains.

4. RESULTS AND DISCUSSION

Tables 2 and 3 summarize the main parameters of the final submitted systems. DIRNET-OOD uses D2/D3 prototype distances and a D1 OOD residual prior. DIRNET-SCR uses the same acoustic modeling pipeline but performs soft probability mixture using pathwise temperatures, score coefficients, routing temperature, and domain bias.

Table 4 presents the local development results at the D3 stage. Since D1 development labels are not available under the Task 7 protocol, only D2 and D3 accuracies are reported. All DIRNET variants substantially outperform the official baseline, improving the D2/D3 macro-average accuracy from 52.5% to over 70%. Among the submitted systems, DIRNET-OOD achieves the best overall performance, reaching 74.89% average accuracy in the S2 configuration. DIRNET-SCR also provides consistent improvements over the baseline, obtaining up to 70.76% average accuracy. These results demonstrate that domain-aware scoring and fusion are critical for domain-agnostic incremental sound classification.

Table 2: Parameters of DIRNET-OOD.

Model	λ_{OOD}
DIRNET-OOD-S1	0.35
DIRNET-OOD-S2	0.70

Table 3: Parameters of DIRNET-SCR. Both submissions use score-calibrated soft probability mixture over D1, D2, and D3 paths. $\alpha = 1.0$ for both.

Model	(T_1, T_2, T_3)	α	β	γ	τ_r	(b_1, b_2, b_3)
DIRNET-SCR-S1	$(1.70, 1.35, 1.10)$	1.0	0.75	0.05	0.50	$(0, 0, 0)$
DIRNET-SCR-S2	$(1.30, 1.20, 1.00)$	1.0	0.45	0.05	1.00	$(0.08, 0, 0)$

Table 4: Local development macro accuracy (%) at the D3 stage. D1 accuracy is not reported as development labels are unavailable.

Model	D2	D3	Avg.
Baseline	59.0	46.1	52.5
DIRNET-OOD-S1	80.06	69.65	74.86
DIRNET-OOD-S2	79.74	70.04	74.89
DIRNET-SCR-S1	80.07	61.45	70.76
DIRNET-SCR-S2	78.63	61.91	70.27

DIRNET-OOD estimates domain responsibility using class-wise prototype distances and a D1 OOD residual prior. This enables class-conditional modeling of domain shift and yields the highest overall performance. However, its effectiveness can decrease when D2 and D3 prototype distances become highly similar. In contrast, DIRNET-SCR performs routing using confidence, prediction margin, and entropy, followed by soft probability mixture across domain paths. Although its performance is slightly lower than that of DIRNET-OOD, it provides a simpler routing mechanism that does not require prototype storage or distance computation. The comparison between S1 and S2 suggests that increasing the D1 contribution improves robustness to potential hidden D1-like samples but may slightly reduce local D2/D3 performance.

5. CONCLUSION

This report presented two families of submitted systems for DCASE 2026 Task 7 under the unified CNN14-based framework DIRNET. DIRNET-OOD performs soft domain fusion using D2/D3 prototypes, a D1 OOD residual prior, and adaptive pairwise calibration. DIRNET-SCR combines maximum confidence, top-1/top-2 margin, entropy, and domain-specific calibration, and softly mixes D1, D2, and D3 calibrated probabilities. Both approaches show that domain-agnostic scoring and fusion are essential for incremental audio classification without test-time domain labels. The results further suggest that soft mixture strategies, which retain all domain paths during inference, are more robust than hard path selection under unknown-domain conditions.

6. REFERENCES

[1] R. Casciotti, M. Mulimani, M. Harju, J. R. Jensen, and A. Mesaros, "Domain-Agnostic Incremental Learning for Sound Classification: A DCASE 2026 Challenge Task," 2026.

[2] M. Mulimani and A. Mesaros, "Domain-Incremental Learning for Audio Classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[6] R. Casciotti, F. De Santis, A. Antonietti, and A. Mesaros, "Incremental learning for audio classification with Hebbian deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 14 777–14 781.

[7] T. Han and Y.-F. Li, "Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles," *Reliability Engineering & System Safety*, vol. 226, p. 108648, 2022.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 1321–1330.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019, pp. 2613–2617.