FINE-TUNED EXPERT AGGREGATION FOR DOMAIN-AGNOSTIC INCREMENTAL AUDIO CLASSIFICATION

Technical Report

Se-Min Heo[1], *Seunggyu Jeong*[1,2], *Seong-Eun Kim*[1,2],

[1] Seoul National University of Science and Technology, Seoul, Korea
[2] Medisensing, Seoul, Korea
smheo@seoultech.ac.kr, wa3229433@gmail.com, sekim@seoultech.ac.kr

ABSTRACT

This technical report describes our submission for Task 7, Domain-Agnostic Incremental Learning for Audio Classification, of the DCASE 2026 Challenge. The task requires an audio classification system to adapt to sequentially introduced domains while performing inference without access to domain labels. Our submitted systems use two separately fine-tuned MCnn14 experts for D2 and D3, both initialized from the official D1 checkpoint. The D2 expert is trained with device augmentation, and the D3 expert is trained with gain-shift augmentation. At inference time, each input sample is evaluated by both experts, and their probability outputs are combined without using domain labels. We evaluate entropy-guided soft aggregation, full-safe test-time augmentation, and mean probability averaging as inference variants. Using the class-wise macro accuracy protocol, the best submitted system obtains a final-stage average validation accuracy of 63.79%.

***Index Terms*—** Domain-Agnostic Incremental Learning, Domain-Dependent Augmentation, Domain-Specific Experts, Mixture of Experts, Entropy-Based Routing

1. INTRODUCTION

Audio classification systems are often affected by domain shifts caused by changes in recording devices, acoustic environments, signal levels, or recording conditions [1]. This issue becomes more challenging when new domains are introduced sequentially and the system must be updated without access to previous-domain training data, which is closely related to continual learning [2, 3].

Task 7 of the DCASE 2026 Challenge addresses this problem through a domain-agnostic incremental learning scenario for audio classification [4, 5]. The system starts from the provided D1 checkpoint and then learns from the newly introduced D2 and D3 domains. At inference time, the domain label of each input sample is not provided, so the system must make predictions without knowing the test-domain identity.

Our submission uses separately fine-tuned MCnn14 experts for D2 and D3. Both experts are initialized from the official D1 checkpoint and trained independently with domain-dependent augmentation. The D2 expert is trained with device augmentation, while the D3 expert is trained with gain-shift augmentation. During inference, both experts are evaluated for each input sample, and their probability outputs are combined without using domain labels. We evaluate entropy-guided soft aggregation, full-safe test-time augmentation, and mean probability averaging as inference variants.

2. PROPOSED SYSTEM

This section describes the proposed system used for our DCASE 2026 Task 7 submission. The submitted systems use two separately fine-tuned MCnn14 experts, one for D2 and one for D3. Both experts are initialized from the official D1 checkpoint. The final S1–S4 configurations do not use adapter modules; instead, they rely on full fine-tuned domain experts and combine their outputs at inference time without using domain labels.

2.1. Overall Pipeline

Figure 1 shows the overall pipeline of the proposed system. The official D1 checkpoint is used as the common initialization source for both the D2 and D3 experts:

$$\theta_2^{(0)} = \theta_{\mathrm{D1}}, \qquad \theta_3^{(0)} = \theta_{\mathrm{D1}}. \tag{1}$$

The D2 expert is fine-tuned using the D2 training data with device augmentation, while the D3 expert is fine-tuned using the D3 training data with gain-shift augmentation.

In the submitted systems, D1 is not used as a separate inference expert. D1 knowledge is inherited through checkpoint initialization. In preliminary validation, adding the D1 model as an additional inference branch did not improve final validation performance, so the final systems aggregate only the D2 and D3 experts.

2.2. Domain-Dependent Expert Training

The submitted systems use MCnn14 as the audio classification backbone, following the CNN14-style convolutional audio models introduced in PANNs [6]. Let f_{θ_2} and f_{θ_3} denote the D2 and D3 experts, respectively. Both experts have the same architecture, but they are trained and stored as separate parameter sets. This design reduces inter-expert interference because D3 training does not overwrite the stored D2 expert, while accepting the cost of maintaining multiple full checkpoints.

We use domain-dependent augmentation to specialize the two experts. Device augmentation is used for D2 to account for device- or recording-condition variation, which is a known source of domain shift in audio classification [1]. Gain-shift augmentation is used for D3 to account for signal-level variation. Let $a_2(\cdot)$ denote the device augmentation for D2 and $a_3(\cdot)$ denote the gain-shift augmentation for D3.

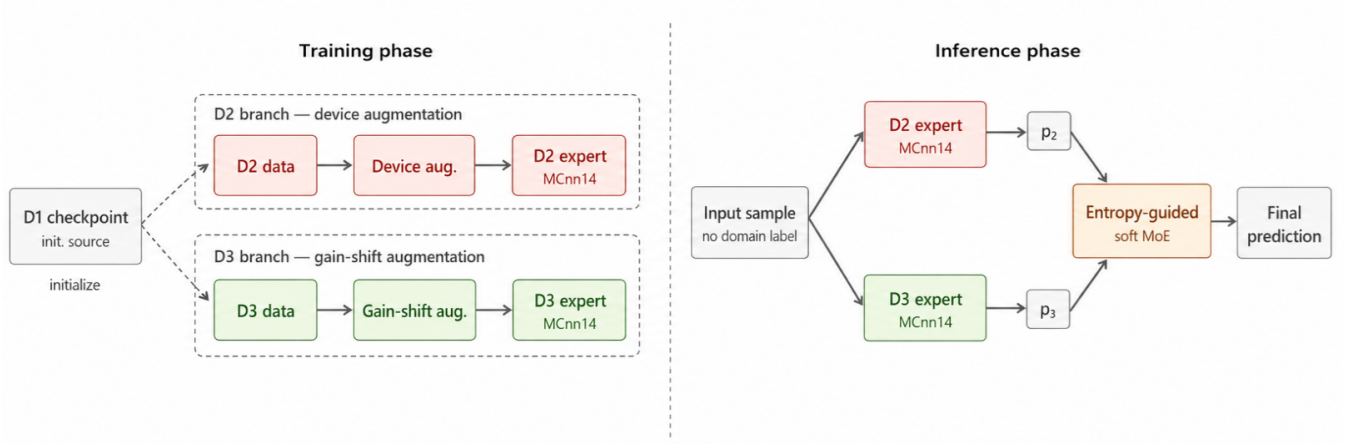

Figure 1: Overview of the proposed fine-tuned expert system. The official D1 checkpoint is used as the initialization source for both D2 and D3 experts. The D2 expert is trained with device augmentation, and the D3 expert is trained with gain-shift augmentation. During inference, both experts are evaluated for each input sample, and their probability outputs are combined without using domain labels.

For domain $i \in \{2, 3\}$, each expert is trained with the standard cross-entropy objective:

$$\mathcal{J}_i(\theta_i) = \frac{1}{N_i} \sum_{n=1}^{N_i} \mathcal{L}_{\mathrm{CE}} \left(f_{\theta_i}(a_i(x_n^{(i)})), y_n^{(i)} \right). \tag{2}$$

The optimized parameters are denoted by

$$\theta_i^* = \arg\min_{\theta_i} \mathcal{J}_i(\theta_i), \qquad i \in \{2, 3\}. \tag{3}$$

No additional routing loss, gating loss, knowledge distillation loss, pseudo-label loss, or consistency loss is used. The expert aggregation mechanism is applied only at inference time.

2.3. Domain-Agnostic Inference

At inference time, the domain label of the input sample is not available. The system evaluates both experts and obtains their probability outputs:

$$p_2 = \mathrm{softmax}(f_{\theta_2^*}(x)), \qquad p_3 = \mathrm{softmax}(f_{\theta_3^*}(x)). \tag{4}$$

The main submitted system uses entropy-guided soft aggregation. This follows the general idea of combining multiple expert predictions, as in mixture-of-experts models [7], but computes the aggregation weights from predictive entropy instead of using a learned gating network.

The predictive entropy of expert i is computed as

$$H_i = -\sum_{c=1}^{C} p_i(c) \log(p_i(c) + \epsilon), \qquad i \in \{2, 3\}. \tag{5}$$

The entropy values are converted into expert weights using a softmax over negative entropy:

$$w_i = \frac{\exp(-H_i/\tau)}{\exp(-H_2/\tau) + \exp(-H_3/\tau)}, \qquad i \in \{2, 3\}. \tag{6}$$

Here, τ is the temperature parameter. A smaller value makes the aggregation closer to hard expert selection, while a larger value produces smoother expert weights.

The final class probability distribution is computed as

$$p_{\mathrm{final}} = w_2 p_2 + w_3 p_3. \tag{7}$$

This inference procedure does not require domain labels, a learned domain classifier, or evaluation-set statistics.

2.4. Submitted Systems

We submit four systems using the same fine-tuned D2 and D3 expert checkpoints. S1 uses entropy-guided soft mixture of experts with $\tau = 3.0$. S2 uses the same aggregation rule with a higher temperature of $\tau = 4.0$. S3 combines entropy-guided soft aggregation with full-safe test-time augmentation. In this report, full-safe TTA refers to applying only label-preserving test-time transformations and averaging the resulting predictions before expert aggregation. Test-time augmentation can improve prediction stability without retraining the model [8]. S4 uses a simple mean probability average of the D2 and D3 expert outputs.

3. EXPERIMENTS AND RESULTS

3.1. Dataset Processing

We use the official development split files provided for DCASE 2026 Task 7. The training and validation metadata are loaded from the official split files, and samples are filtered by domain during expert training. The D2 expert is trained using D2 samples, and the D3 expert is trained using D3 samples.

The D1 training data are not reused in our submitted systems. Instead, the official D1 checkpoint is used only as the initialization source for both experts. The D2 and D3 experts are trained independently, and previous-domain training data are not reused during later expert training.

Audio files are loaded as mono waveforms at 32 kHz and prepared as 4-second inputs. The MCnn14 front end then converts the waveform into spectrogram and log-mel features before classification.

Table 1: Validation accuracy across incremental stages.

System	D2	D2 after D3	D3	Avg.
S1: Soft MoE, $\tau = 3.0$	77.03	76.93	49.52	63.22
S2: Soft MoE, $\tau = 4.0$	77.03	76.83	49.64	63.24
S3: Soft MoE + TTA	77.03	76.05	50.42	63.23
S4: Mean prob. avg.	77.03	76.68	50.90	63.79

3.2. Experimental Setup

All experiments are conducted using the official development data of DCASE 2026 Task 7. The submitted systems use two separately fine-tuned MCnn14 experts, one for D2 and one for D3. Both experts are initialized from the official D1 checkpoint. The D2 expert is trained with device augmentation, while the D3 expert is trained with gain-shift augmentation. Each expert is trained using the standard cross-entropy loss.

All submitted systems share the same trained D2 and D3 expert checkpoints. Therefore, the difference between S1–S4 comes only from the inference configuration. At inference time, the domain label is not used. Each input sample is evaluated by both experts, and the final prediction is obtained by combining their probability outputs according to the inference strategy of each system.

For reproducibility, each expert is fine-tuned for 120 epochs using AdamW with a batch size of 32 and an initial learning rate of 1×10^{-4} decayed with a cosine annealing schedule. The same training protocol is used for the D2 and D3 experts, except for the domain-dependent augmentation strategy. The submitted systems do not use an additional learned router, pseudo-labeling, knowledge distillation, or consistency regularization. The aggregation mechanism is applied only at inference time.

3.3. Evaluation Protocol

We report validation accuracy across the incremental stages. Following the official evaluation protocol, domain-wise accuracy is computed as the average of class-wise accuracies within each domain. For a domain d, the domain-wise accuracy is computed as

$$\mathrm{Acc}d = \frac{1}{|\mathcal{C}d|} \sum_{c \in \mathcal{C}d} \mathrm{Acc}d, c, \tag{8}$$

where $\mathcal{C}d$ denotes the set of classes appearing in domain d, and $\mathrm{Acc}d, c$ is the accuracy for class c in domain d.

For the final-stage validation results, the average score is computed from the final D2 and D3 domain-wise accuracies:

$$\mathrm{Avg.} = \frac{\mathrm{Acc}_{\mathrm{D2\ after\ D3}} + \mathrm{Acc}_{\mathrm{D3}}}{2}. \tag{9}$$

3.4. Validation Results

Table 1 reports the validation accuracy across incremental stages. S1 and S2 use entropy-guided soft MoE with $\tau = 3.0$ and $\tau = 4.0$, respectively. S3 adds full-safe TTA to S1, and S4 uses mean probability averaging. The D2 column denotes the validation accuracy after D2 expert training, while D2 after D3 and D3 denote the final-stage validation accuracies.

The results show that D2 accuracy is largely preserved after introducing D3, while D3 remains substantially more challenging. The D2 after D3 accuracy remains around 76–77% across the submitted systems, whereas the D3 accuracy remains around 49–51%. This indicates that final performance is mainly limited by D3 adaptation.

Among the submitted systems, S4 obtains the highest final-stage average validation accuracy of 63.79%. This result suggests that the D2 and D3 experts provide complementary predictions, and that simple mean probability averaging is competitive in this validation setting. S3 improves D3 accuracy compared with S1 and S2, but the improvement is not sufficient to achieve the best final average because its D2 after D3 accuracy is lower.

S1 and S2 show similar performance, indicating that changing the temperature from $\tau = 3.0$ to $\tau = 4.0$ has only a limited effect. Overall, these results suggest that predictive entropy alone is not always a reliable routing signal, and that the main benefit of the submitted systems comes from using multiple fine-tuned experts rather than from entropy weighting alone.

Overall, the submitted results show that inference-time aggregation affects final validation performance even when the same D2 and D3 experts are used. The best result is obtained by combining soft expert aggregation with full-safe test-time augmentation, mainly due to improved D3 performance.

4. CONCLUSION

In this technical report, we described our submission for Task 7 of the DCASE 2026 Challenge. The submitted systems address domain-agnostic incremental audio classification by using separately fine-tuned MCnn14 experts for D2 and D3. Both experts are initialized from the official D1 checkpoint. The D2 expert is trained with device augmentation, while the D3 expert is trained with gain-shift augmentation. The submitted systems do not use adapter modules in the final S1–S4 configurations. Instead, they use full fine-tuned domain experts and combine their outputs at inference time.

The D1 checkpoint is used as the common initialization source for the D2 and D3 experts, rather than as a separate inference branch. This design allows the experts to inherit the representation learned from D1 while adapting to the newly introduced domains. Each expert is trained using the standard cross-entropy loss, and no additional routing loss, gating loss, knowledge distillation loss, pseudo-label loss, or consistency loss is used. The domain-agnostic inference stage is performed only at test time by combining the probability outputs of the D2 and D3 experts.

We submitted four inference configurations using the same trained experts. The entropy-guided soft mixture system with $\tau = 3.0$ obtained a Step3 average validation accuracy of 67.37%, and the high-temperature version with $\tau = 4.0$ obtained 67.42%. The mean probability averaging system obtained 67.88%. The best validation result was obtained by the full-safe test-time augmentation system, which achieved a Step3 average validation accuracy of 67.95%. This system improved the D3 accuracy to 56.08%, which led to the highest average score among the submitted systems.

The results suggest that inference-time aggregation affects the final performance even when the same D2 and D3 experts are used. In particular, D3 remains more challenging than D2 across all submitted systems, and improving D3 prediction stability has a large effect on the final average score. Future work includes reducing the parameter cost of maintaining multiple full experts, improving uncertainty calibration for entropy-based aggregation, and exploring

more compact domain-specific adaptation methods such as adapters or other parameter-efficient modules.

5. REFERENCES

[1] T. Morocutti, F. Schmid, K. Koutini, and G. Widmer, "Device-robust acoustic scene classification via impulse response augmentation," in *Proceedings of the 31st European Signal Processing Conference*, 2023, pp. 176–180.

[2] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3366–3385, 2022.

[3] 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, pp. 1–5.

[4] DCASE Community, "DCASE 2026 Challenge Task 7: Domain-Agnostic Incremental Learning for Audio Classification," https://dcase.community/challenge2026/task-domain-agnostic-incremental-learning-for-audio-classification, 2026, accessed: 2026-06-15.

[5] R. Casciotti, M. Mulimani, M. Harju, J. R. Jensen, and A. Mesaros, "Domain-agnostic incremental learning for sound classification: A dcase 2026 challenge task," *arXiv preprint arXiv:2606.02173*, 2026.

[6] 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.

[7] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[8] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1214–1223.