BN-DISTANCE SOFT ROUTING FOR DOMAIN-AGNOSTIC INCREMENTAL AUDIO CLASSIFICATION

Technical Report — DCASE 2026 Task 7

Zhenping Zhang[1], *Wenxing Yang*[1,*], *Wenqiang Zhao*[1], *Qianyi Wang*[1], *Yuzhu Wang*[2]

[1]College of Oriental Pan-Vascular Devices Innovation, University of Shanghai for Science and Technology, Shanghai, China
[2]Signal Processing Research Center, Tampere University, Tampere, Finland
[*]Corresponding author: wenxingyang@usst.edu.cn

ABSTRACT

This report describes our submission to DCASE 2026 Task 7, Domain-Agnostic Incremental Learning for Audio Classification. The submitted system keeps the official CNN14-style acoustic classifier and its domain-specific batch-normalization (BN) branches unchanged. The core idea is to treat shallow BN running statistics as domain fingerprints. For each test clip, BN-distance is computed between the clip's pre-BN activation statistics and the stage-available BN branch statistics at selected shallow BN layers. These distances are converted to softmax branch weights and used to softly fuse the branch class posteriors. No learned router or selector is trained; only a deterministic routing rule is applied at inference time. The method uses no external data, no data augmentation, and no evaluation-set statistics. On the development set, after-D2 D2 macro accuracy improves from 58.60% to 65.03%, and after-D3 D2/D3 average macro accuracy improves from 52.50% to 58.87%.

***Index Terms*—** Domain-agnostic learning, incremental learning, audio classification, batch normalization, soft routing

1. INTRODUCTION

DCASE 2026 Task 7 studies domain-agnostic incremental learning for audio classification. At inference time, the system must predict the acoustic event class without knowing the domain label of the input clip. The official baseline addresses domain shift with domain-specific BN branches for D1, D2, and D3, while most convolutional and classifier parameters are shared.

A direct way to use such a model is hard branch selection: choose one BN branch and output the prediction from that branch. However, branch selection is uncertain when the input domain is unknown. A wrong hard decision discards useful predictions from the other branches. Our submission therefore uses BN-distance soft routing. The system does not learn a new branch selector. It only uses the BN running mean and variance already stored in the submitted acoustic dictionaries. The intuition is that the BN running statistics act as low-dimensional domain fingerprints: if a sample's activation statistics are close to a branch's running statistics, that branch should receive a larger contribution to the final prediction.

2. ACOUSTIC MODEL

The acoustic model is the official Task 7 CNN14-style BN-branch baseline. Audio is sampled at 32 kHz. Log mel-band energies are extracted with 64 mel bins, a 1024-sample analysis window, and a 320-sample hop size. The CNN contains six convolutional blocks followed by global pooling and a fully connected 10-class classifier. The convolutional and classifier weights are shared across domains, while each BN layer has separate D1, D2, and D3 branches.

We use the provided D2 and D3 model dictionaries without fine-tuning. The submitted post-processing rule has no learned parameters. It does not use the labels of D2 and D3 jointly to train a router. During after-D2 inference, each input clip is forwarded through the BN branches available in the D2 checkpoint to produce branch posteriors. During after-D3 inference, each input clip is forwarded through the D3 checkpoint with the D1, D2, and D3 BN branches, yielding three class-posterior vectors $p_1(y \mid x)$, $p_2(y \mid x)$, and $p_3(y \mid x)$.

3. BN-DISTANCE SOFT ROUTING

For a sample x, branch k in the stage-available branch set $\mathcal{K}$, and BN layer l, we compute a normalized distance between the sample's pre-BN activation mean and the branch running statistics:

$$d_l(x,k) = \frac{1}{C_l} \sum_{c=1}^{C_l} \frac{(\mu_{x,l,c} - m_{k,l,c})^2}{v_{k,l,c} + \epsilon}, \tag{1}$$

where C_l is the number of channels in layer l, $\mu_{x,l,c}$ is the channel-wise mean of the pre-BN activation for sample x, computed over the time-frequency/spatial dimensions for each individual input clip, and $m_{k,l,c}$ and $v_{k,l,c}$ are the running mean and running variance of BN branch k. The distance is computed in evaluation mode, so BN running statistics are never updated.

The submitted system uses the layer set

$$\mathcal{L} = \{\texttt{bn0}, \texttt{conv\_block1.bnF}, \texttt{conv\_block1.bnS}\}. \tag{2}$$

For each branch, the selected-layer distance is the average over the three layers:

$$D(x,k) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} d_l(x,k). \tag{3}$$

Table 1: Stage-specific configuration of the submitted BN-distance soft-routing system.

Stage	Checkpoint	BN branches	BN-distance layers
after-D2	`checkpoint_D2.pth`	D1, D2	`bn0+conv_block1.bnF+conv_block1.bnS`
after-D3	`checkpoint_D3.pth`	D1, D2, D3	`bn0+conv_block1.bnF+conv_block1.bnS`

The branch distances are then converted to soft routing weights. For after-D2, $\mathcal{K} = \{D1, D2\}$; for after-D3, $\mathcal{K} = \{D1, D2, D3\}$:

$$w_k(x) = \frac{\exp(-D(x,k)/\tau)}{\sum_{j\in\mathcal{K}} \exp(-D(x,j)/\tau)}, \tag{4}$$

with temperature $\tau = 0.5$. Smaller BN distance gives a larger weight. The final class posterior is

$$p(y \mid x) = \sum_{k\in\mathcal{K}} w_k(x) p_k(y \mid x), \tag{5}$$

where $p_k(y \mid x)$ is the softmax posterior produced by branch k. The predicted class is $\arg\max_y p(y \mid x)$.

This differs from hard BN routing. Hard routing chooses only the branch with minimum BN distance. Soft routing keeps the branch decision continuous, so a sample can still use information from non-selected branches. This is useful because BN distance is an imperfect domain indicator and because the true-domain branch is not always the branch with the best class prediction.

4. CHOICE OF BN LAYERS

The selected layers are shallow BN layers: `bn0`, `conv_block1.bnF`, and `conv_block1.bnS`. This choice is based on the role of BN statistics in the network. Shallow activations are closer to input acoustic characteristics such as spectrum shape, channel response, background noise, and recording condition. These are the factors most directly related to domain shift. Therefore shallow BN running statistics are suitable as domain fingerprints.

Deeper BN layers are more class-dependent. Their activations reflect higher-level semantic content as well as domain properties. Adding such layers can mix class variation into the domain-distance estimate. Development observations showed that `bn0` is already a strong domain cue and that adding the first convolutional block gives a stable shallow fingerprint. These observations support using shallow BN statistics rather than averaging over many layers.

5. IMPLEMENTATION AND DATA USE

The development analysis uses two types of precomputed files. Branch probabilities are produced by running the acoustic model through the domain-specific BN branches available in the corresponding checkpoint and saving the class posterior for each branch. BN-distance files are produced by recording pre-BN activation means and comparing them with each branch's running mean and variance. For the after-D2 evaluation, D2 development results are generated with `checkpoint_D2.pth` and the BN branches available in that D2 checkpoint. For the after-D3 evaluation, D2 and D3 development results are generated with `checkpoint_D3.pth`, using the D1, D2, and D3 BN branches stored in the D3 checkpoint. These files are used only for deterministic evaluation of the fixed soft-routing rule.

Table 2: Development-test macro accuracy in the official reporting format.

System and stage	D2 (%)	D3 (%)	Avg. (%)
Official baseline after D2	58.60	–	58.60
BN-distance soft routing after D2	**65.03**	–	**65.03**
Official baseline after D3	59.00	46.10	52.50
BN-distance soft routing after D3	**68.15**	**49.59**	**58.87**

Table 3: Class-wise macro components for BN-distance soft routing after D3. A dash indicates that the class is not present in the corresponding domain.

Class	D2 (%)	D3 (%)
alarm	40.77	66.13
baby_cry	–	45.83
bark	86.96	46.75
engine	82.61	63.53
fire	27.78	37.84
footsteps	79.59	26.87
knock	80.56	–
telephone_ringing	–	38.71
piano	67.01	95.89
speech	79.90	24.73
macro avg.	68.15	49.59

The method does not fit a classifier or tune a learned module on the union of D2 and D3 samples. The only manually selected hyperparameters are the BN layer set and the softmax temperature $\tau = 0.5$. The layer set was chosen from development observations about shallow BN statistics. Table 1 summarizes the stage-specific checkpoints, branch sets, and selected BN layers used in the final evaluation. The evaluation-set audio is used only to produce predictions; no labels or evaluation-set statistics are used.

6. RESULTS

Following the official metric, we report class-wise macro accuracy. For each domain, accuracy is first computed separately for each class present in the domain, then averaged over classes. Table 2 compares the official baseline with the proposed soft-routing system.

The proposed system improves the after-D2 D2 macro accuracy from 58.60% to 65.03%. After D3, it improves the D2/D3 average from 52.50% to 58.87%. The D2 score remains high after adding D3 because probability fusion avoids making a brittle branch decision. The D3 improvement is smaller, which suggests that some D3 samples still receive non-negligible weight from D1 or D2 branches. This indicates that distance-based soft fusion is useful, but the remaining D3 errors also show that BN-distance is an imperfect proxy for classification reliability. Some D3 classes, especially `speech` and `footsteps`, remain challenging, and routing uncertainty still exists for D3 samples whose shallow BN statistics overlap with D1 or D2 branches.

7. SUBMISSION FILES

The submitted system is `Yang_USST_task7_1`, abbreviated BSBR-BND. The output file is a tab-separated file without a header, containing the evaluation filename and predicted class label. The output file contains 3755 rows, matching the number of evaluation audio files.

8. CONCLUSION

We presented a deterministic BN-distance soft-routing method for domain-agnostic incremental audio classification. The method trains no learned router or selector. Instead, it converts shallow BN-statistic distances into softmax branch weights and fuses the class posteriors from the existing BN branches. The selected shallow layer set, `bn0+conv_block1.bnF+conv_block1.bnS`, is used as a fixed deterministic configuration for both stages. The final system improves the official after-D3 average macro accuracy from 52.50% to 58.87% while keeping the acoustic model unchanged.

9. REFERENCES

[1] DCASE 2026 Challenge, "Task 7: Domain-Agnostic Incremental Learning for Audio Classification," 2026. [Online]. Available: `https://dcase.community/challenge2026/`

[2] M. Mulimani and A. Mesaros, "Domain-Incremental Learning for Audio Classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 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.