MULTIFEATURE SEQUENTIAL LEARNING WITH CNN14 ENSEMBLE FOR DCASE 2026 TASK 7

Technical Report

Anderson Giacomini

Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP), São Carlos, Brazil
giacomini@usp.br

ABSTRACT

We present a system for DCASE 2026 Task 7 (Domain-Agnostic Incremental Learning) that combines a multi-feature cross-attention transformer (MultiFeatureSED) with the organizer-provided CNN14 baseline. MultiFeatureSED is trained from scratch on Domain 2 (D2) and fine-tuned on Domain 3 (D3) with L2 weight regularization to mitigate catastrophic forgetting. D1 knowledge is incorporated exclusively via the CNN14 checkpoint, whose domain-specific batch normalization and entropy-based domain routing complement the multi-feature model at inference. An equal-weight ensemble with domain-conditional per-class probability calibration, optimized on the development set, achieves 61.5% mean balanced accuracy versus 55.2% for CNN14 alone.

***Index Terms*—** incremental learning, catastrophic forgetting, sound event detection, cross-attention, calibration

1. INTRODUCTION

Domain-Agnostic Incremental Learning for sound classification [1] requires a model to absorb new domains sequentially without access to prior-domain data, while retaining performance on all domains seen so far [2]. The core challenge is catastrophic forgetting: gradient updates for new data tend to overwrite weights that encode prior-domain knowledge [1]. Our system addresses this with two complementary strategies.

Regularization-based retention. MultiFeatureSED—a cross-attention transformer over seven acoustic feature streams—is fine-tuned on each new domain using an L2 penalty that constrains parameters to stay close to the previous checkpoint, a simplified form of Elastic Weight Consolidation without Fisher weighting. This preserves D2 representations while adapting to D3.

Knowledge-preserving ensemble. The organizer-provided CNN14 baseline carries D1 knowledge that MultiFeatureSED cannot acquire under the sequential rule (D1 data is inaccessible during our training). Ensembling CNN14 with MultiFeatureSED therefore adds a complementary knowledge source rather than just averaging correlated predictions: CNN14 leads on D2 (where its domain-specific BN was trained with more data) while MultiFeatureSED D3 leads on D3.

Per-class probability calibration further corrects systematic class confusions that persist after ensembling. We extend joint calibration to domain-conditional calibration—separate weight vectors for clips assigned to D2 vs. D3—yielding an additional gain on D3 with minimal cost on D2.

The task involves 10 sound classes (alarm, baby_cry, dog_bark, engine, fire, footsteps, knocking, telephone_ringing, piano, speech) across three domains. MultiFeatureSED is trained on D2 (1530 train / 639 dev clips) and fine-tuned on D3 (1882 train / 806 dev clips). All clips are 4 seconds at 32 kHz. D1 data is not accessed; D1 knowledge enters only through the provided CNN14 checkpoint.

2. SYSTEM DESCRIPTION

2.1. MultiFeatureSED Architecture

MultiFeatureSED fuses seven complementary feature streams, all computed with a common hop of 512 samples (250 frames per 4-second clip at 32 kHz):

- **Log-mel spectrogram**: 512-pt FFT, 64 mel bins, 50–14 kHz
- **MFCC**: 512-pt FFT, 40 coefficients
- **Chroma**: 512-pt FFT, 12 bins
- **ZCR, RMS, Spectral centroid**: frame-level scalar streams
- **Statistical**: $[\mu, \sigma, \max, \min, p_{25}, p_{75}]$ per HOP-length frame (6-d)

Spectral streams are tokenized by a 2D CNN (32 channels); scalar streams by 1D CNNs (64 channels, 2 layers). All streams are linearly projected to $d_{\text{model}} = 128$. Fusion is by pairwise cross-attention: each stream attends to every other stream via $N_{\text{heads}} = 8$ multi-head attention, repeated for $N_{\text{layers}} = 4$ transformer blocks. Frame-level logits from intermediate layers 1 and 3 are averaged with the final layer (weight 0.3 each) as an auxiliary supervision signal during training. Clip-level predictions are obtained by linear-softmax MIL pooling [3] with temperature $\beta = 3.0$.

2.2. Sequential Training

D2 training starts from random initialization using only D2 audio. Loss is binary cross-entropy with per-class weights proportional to the inverse square root of class frequency, correcting the class imbalance in D2. Augmentation includes MixUp ($\alpha = 0.4$) and label smoothing ($\varepsilon = 0.05$). Optimizer: AdamW (lr=10^{-4}, weight decay 10^{-2}) with ReduceLROnPlateau (patience 8, factor 0.1). Early stopping on D2 balanced accuracy (patience 16); converged at epoch 109, reaching 65.2% D2 BAcc.

D3 fine-tuning loads the D2 checkpoint and trains exclusively on D3 data (no D2 data accessed). To mitigate catastrophic forget-

ting, the loss includes an L2 parameter-distance penalty:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \sum_i \left(\theta_i - \theta_i^{D2}\right)^2 \tag{1}$$

where θ^{D2} are the frozen D2 checkpoint parameters. The penalty strength λ is selected by ablation (Table 1). Early stopping monitors D3 balanced accuracy only (patience 16); the best checkpoint ($\lambda = 0.001$) converged at epoch 56.

2.3. CNN14 Baseline

The CNN14 baseline [2] uses domain-specific batch normalization (one BN set per domain). At inference, the model is run twice—once with D2 BN and once with D3 BN—and the branch producing the lower output entropy is selected:

$$\text{domain}^* = \arg\min_{d \in \{D2, D3\}} H(p_d), \quad H(p) = -\sum_k p_k \log p_k \tag{2}$$

This soft routing requires no explicit domain label at test time. On the evaluation set (3755 clips), 2927 clips (78%) were routed to D2 BN and 828 (22%) to D3 BN.

2.4. Ensemble and Calibration

Probability ensemble. Predictions from both models are averaged with equal weight ($\alpha = 0.5$, selected by grid search; Table 2):

$$p_{\text{ens}}[k] = 0.5 \cdot p_{\text{MF-D3}}[k] + 0.5 \cdot p_{\text{CNN14}}[k] \tag{3}$$

Equal weighting is optimal because the two models have complementary strengths: CNN14 leads on D2 (+18.1% over MF-D3) while MF-D3 leads on D3 (+10.1% over CNN14).

Domain-conditional calibration. A per-class scaling vector $\mathbf{w} \in R^{10}_{>0}$ is applied before argmax to correct systematic class biases:

$$\hat{y} = \arg\max_k \; w_k \cdot p_{\text{ens}}[k] \tag{4}$$

We optimize two separate vectors, $\mathbf{w}_{D2}$ and $\mathbf{w}_{D3}$, via Nelder-Mead minimization of the negative mean balanced accuracy on the respective development sets. At inference, the CNN14 entropy-based domain assignment (Section 2.3) selects which vector to apply per clip. This domain-conditional approach adds +0.8% mean BAcc over joint calibration (a single shared $\mathbf{w}$ for all clips) by better adapting to the class confusion patterns that differ between domains.

3. EXPERIMENTAL RESULTS

3.1. L2 Regularization Ablation

Table 1 shows sensitivity to λ in Eq. (1). All three values incur substantial D2 forgetting relative to the 65.2% achieved before fine-tuning. Increasing λ trades D3 performance for D2 retention monotonically: $\lambda = 0.003$ recovers 1.9% D2 BAcc over $\lambda = 0.001$ (45.3% vs. 43.4%) but loses 5.0% on D3; $\lambda = 0.01$ recovers a further 5.3% D2 but collapses D3 to 44.2%. In terms of the mean score optimized by the submission, $\lambda = 0.001$ is strictly best (50.5% vs. 49.0% and 47.4%). The selected checkpoint relies on the CNN14 ensemble to compensate for the D2 forgetting it accepts.

Table 1: Effect of L2 regularization strength λ on D3 fine-tuning (dev BAcc %).

λ	D2	D3	Mean
0.001	43.4	**57.5**	**50.5**
0.003	45.3	52.6	49.0
0.010	**50.6**	44.2	47.4

3.2. Ensemble Weight Ablation

Table 2 shows the effect of ensemble weight α (CNN14 fraction). Equal weighting ($\alpha = 0.5$) maximizes mean balanced accuracy. Increasing α beyond 0.5 improves D2 but degrades D3 faster than D2 improves, reflecting CNN14's comparative weakness on the newer domain.

Table 2: Ensemble weight grid search: $\alpha \cdot p_{\text{CNN14}} + (1-\alpha) \cdot p_{\text{MF-D3}}$ (dev BAcc %).

α	D2	D3	Mean
0.0 (MF-D3 only)	43.4	**57.5**	50.5
0.25	49.1	57.3	53.2
0.50	61.5	54.2	**57.8**
0.75	**63.7**	49.2	56.5
1.0 (CNN14 only)	62.8	47.5	55.2

3.3. System Comparison and Calibration Analysis

Table 3 shows the full system progression. Joint calibration contributes +2.9% mean BAcc over the uncalibrated ensemble; domain-conditional calibration adds a further +0.8%, primarily by improving D3 recognition (+2.4%).

Table 3: Development set balanced accuracy (%) by system component. The D2-only row is evaluated before D3 fine-tuning; D3 performance at that stage is not measured.

System	D2	D3	Mean
CNN14 baseline	62.8	47.5	55.2
MultiFeatureSED D2-only	**65.2**	—	—
MultiFeatureSED D3 seq. ($\lambda = 0.001$)	43.4	57.5	50.5
Ensemble $\alpha = 0.5$ (uncalibrated)	61.5	54.2	57.8
+ joint calibration	63.1	58.3	60.7
+ domain-cond. calibration (submitted)	62.3	**60.7**	**61.5**

Table 4 shows all three calibration weight vectors. Values below 1 indicate over-prediction; values above 1 indicate under-prediction by the uncalibrated ensemble. *Footsteps* is the most consistently over-predicted class across both domains ($w_{D2} = 0.269$, $w_{D3} = 0.331$), reflecting a strong tendency for the ensemble to assign footstep-like sounds incorrectly. *Fire* and *speech* are also systematically over-predicted.

The domain-conditional vectors differ substantially from each other, validating the domain-conditional approach. *Baby_cry* requires a large upweight in D3 ($w_{D3} = 2.363$) but only moderate correction in D2 ($w_{D2} = 1.124$), suggesting the D3 domain makes this class significantly harder to detect. Conversely, *engine* switches

from under-predicted in D2 ($w_{D2} = 1.291$) to over-predicted in D3 ($w_{D3} = 0.817$), indicating domain-specific confusion patterns that a single shared vector cannot capture.

Table 4: Per-class calibration weights (joint and domain-conditional). Values <1: over-predicted; values >1: under-predicted by the uncalibrated ensemble.

Class	Joint	w_{D2}	w_{D3}
alarm	0.945	1.331	0.797
baby_cry	1.527	1.124	**2.363**
dog_bark	1.547	1.592	1.002
engine	1.584	1.291	0.817
fire	0.423	0.595	0.465
footsteps	**0.371**	**0.269**	**0.331**
knocking	0.761	0.633	0.546
tel. ringing	1.240	1.403	1.536
piano	1.096	1.763	1.181
speech	0.506	0.512	0.426

4. CONCLUSION

We presented a two-component system for DCASE 2026 Task 7 that addresses domain-agnostic incremental learning through regularization-based forgetting prevention and a complementary ensemble. MultiFeatureSED's seven-stream cross-attention architecture leverages diverse acoustic cues, while its L2 regularized sequential training preserves D2 representations during D3 adaptation. The CNN14 ensemble provides D1 knowledge inaccessible to MultiFeatureSED under the sequential constraint, and its entropy-based domain routing enables calibration without explicit domain labels at test time. The full system achieves 61.5% mean development set balanced accuracy, a +6.3% improvement over CNN14 alone.

A key limitation is that L2 regularization with uniform λ across all parameters is a weak proxy for parameter importance: parameters critical to D2 and irrelevant to D3 should be frozen more tightly. Future work could apply Fisher-weighted EWC [4] or parameter-efficient fine-tuning (e.g., LoRA adapters per domain) to better separate domain-specific from shared representations, potentially recovering more of the D2 performance lost during D3 fine-tuning.

5. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[3] Y. Wang, N. J. Bryan, M. Cartwright, J. P. Bello, and J. Salamon, "A comparison of deep learning methods for weakly supervised sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2655–2669, 2021.

[4] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.