MULTIMODAL ENSEMBLE SYSTEM FOR HETEROGENEOUS AUDIO CLASSIFICATION

Technical Report

Mehmet Atilay Kucukoglu

New York University, New York, USA
mk9649@nyu.edu

ABSTRACT

This technical report presents an ensemble approach to DCASE 2026 Challenge Task 1 on Heterogeneous Audio Classification, along with the experiments that led to it. Our reproduction of the HATR baseline system achieves 79.01% hierarchical F1 score on BSD10k-v1.2. Alternative audio encoders such as BEATs, ConvNeXt, MATPAC, and Whisper performed below CLAP, suggesting that CLAP's joint audio-text alignment provides a powerful grounding for this task. Incorporating hierarchical loss, contrastive loss, confidence weighting, and class weighting did not improve the hierarchical F1 score. Data augmentation methods such as CutMix, mixup, and cross-modal swap with Gaussian noise were also explored. The cross-modal swap achieved the best single model performance at 79.13% hF1. To address class imbalance and the low performance of specific classes, external data from FSD50K and ESC-50 mapped to the BST taxonomy, as well as the larger BSD35k-CS dataset, were explored but did not achieve higher F1 scores. The best results were observed using an ensemble of 5 models trained throughout this exploration that incorporate different encoders, loss strategies, and augmentation methods. The ensemble approach achieved 81.13% hierarchical F1 score on BSD10k-v1.2, a +2.12% improvement over the baseline.

***Index Terms*—** Heterogeneous audio classification, CLAP, ensemble, broad sound taxonomy, multimodal learning

1. INTRODUCTION

The DCASE 2026 Challenge Task 1 focuses on heterogeneous audio classification using the Broad Sound Taxonomy (BST) [1]. The challenge of this task comes from its heterogeneous nature, meaning that the proposed system should be able to classify sounds as diverse as instrument samples, crowd speech, urban soundscapes, and experimental sound effects. The official baseline of the challenge utilises the Hierarchical Audio-Text Representation (HATR) architecture [2] on top of audio and text embeddings extracted with the LAION-CLAP [3] encoder.

This technical report presents a systematic investigation to improve upon the baseline HATR system. First, alternative audio encoders (BEATs [4], ConvNeXt [5], MATPAC [6], and Whisper [7]) are evaluated to test whether an alternative audio backbone could improve classification. Second, several loss functions are explored, including hierarchical loss that utilises the two-level taxonomy and supervised contrastive loss. Third, training strategies such as confidence-based and class-based sample weighting are tested to address class imbalance. Fourth, data augmentation methods including mixup, CutMix, and cross-modal embedding swap are explored. Fifth, external datasets (ESC-50 [8] and FSD50K [9]) are mapped to the BST taxonomy and incorporated as additional training data. Finally, the models trained across these experiments are combined through ensemble logit averaging.

Experimental results showed that alternative audio encoders resulted in performance drops in both the audio-only and multimodal approaches. This suggests that CLAP's [3] joint audio-text pretraining provides a strong fundamental understanding that other audio encoders cannot match for this task. Experimental loss functions, training strategies, data augmentation methods, and external dataset integration also could not beat, or had only minimal gains over the multimodal baseline. The best single model observed was trained using cross-modal embedding swap with Gaussian noise, achieving a 0.12% gain over the baseline. The biggest improvement came from the ensemble method that was averaging the output logits of 5 trained models. This ensemble achieved a 2.12% gain over the baseline, suggesting that different approaches learn complementary information that benefits different classes.

2. DATASETS

The provided BSD10k-v1.2 dataset was used throughout the experiments [10, 2]. The dataset consists of 10,956 audio clips collected from Freesound and annotated according to the Broad Sound Taxonomy (BST) [1] by expert annotators. The BST taxonomy organizes sounds into a two-level hierarchy. The 5 top-level categories are Music, Instrument samples, Speech, Sound effects, and Soundscapes. These are further divided into 23 second-level classes. Each audio clip is provided with its audio file and textual metadata including a title, tags, and a free-text description. Annotations also include a confidence score from 1 to 5 indicating annotator certainty.

The second development dataset experimented with was BSD35k-CS [11], which contains 33,829 sounds labeled by their Freesound uploaders. This makes its annotations noisier than BSD10k-v1.2. This dataset was incorporated as additional training data to test whether the larger data volume could improve performance.

In later experiments, two additional datasets were explored in an attempt to balance out some of the low performing classes. ESC-50 [8] is a collection of 2,000 environmental sound recordings organized into 50 balanced classes, each clip being 5 seconds long. FSD50K [9] is a larger dataset of over 51,000 sounds from Freesound annotated with labels from the AudioSet ontology. Since neither dataset follows the BST taxonomy, their labels were manually mapped to the 23 BST classes.

All experiments use BSD10k-v1.2 through 5-fold stratified

https://github.com/Atilik/dcase2026_task1_kucukoglu

cross-validation, with the same random seed (1821) across all experiments to ensure that fold splits are identical and results are directly comparable to the reproduced baseline. For experiments using external data, additional samples are added only to the training folds. Validation and test folds always consist of BSD10k-v1.2 sounds to preserve comparability. The macro hierarchical F1 score (hF1) with $\lambda = 0.75$ is used as the primary evaluation metric, as it is the main metric of the challenge.

3. METHOD

3.1. Baseline

The baseline system uses the HATR (Hierarchical Audio-Text Representation) [2] architecture with LAION-CLAP [3] as both the audio and text encoder. The baseline repository provides the precomputed embeddings extracted for both the audio and text using LAION-CLAP. During training, the audio and text representations are combined through an attention-based fusion module that learns per-sample weights for each modality. The fused representation is passed to a classification head containing residual blocks, which produces logits over the 23 second-level classes. Training uses the Adam optimizer [12] with early stopping and 5-fold stratified cross-validation.

3.2. Encoder

The audio encoder is an essential component of a classification system. To test whether a different encoder could improve performance, the first experiments replaced the baseline's CLAP [3] audio encoder with alternatives. Table 1 shows the performance of the HATR [2] backbone with four alternative encoders: BEATs [4], ConvNeXt [5], MATPAC [6], and Whisper [7]. Note that for the multimodal setting only the audio encoder is swapped, while CLAP is kept as the text encoder.

Table 1: Audio encoder comparison on BSD10k-v1.2 test split (5-fold Cross validation). Hierarchical F1 (%) score reported for audio-only and multimodal settings.

Encoder	Audio-only	Multimodal
CLAP (baseline)	**75.30 ± 1.14**	**79.01 ± 0.54**
ConvNeXt	73.03 ± 0.39	78.17 ± 0.60
MATPAC	–	77.63 ± 0.36
BEATs	70.38 ± 0.88	77.38 ± 1.24
Whisper	64.97 ± 0.96	77.49 ± 0.41

These results suggest that CLAP's audio-text pretraining makes it a stronger backbone for this task than encoders trained on audio alone.

A three-modality variant was also explored that combines CLAP audio, CLAP text, and Whisper audio embeddings through the same attention-based fusion. Whisper is trained for speech recognition, and the motivation was that it could add speech-related information to help with the underperforming speech classes and complement the more general CLAP audio embeddings. This did not help, reaching 78.83% hF1 with mixup data augmentation, which is slightly below the two-modality baseline.

The following pretrained checkpoints were used for each encoder:

- **CLAP**: `630k-audioset-fusion-best.pt`
- **ConvNeXt**: `convnext_tiny_465mAP_BL_AC_70kit.pth`
- **MATPAC**: `matpac_10_2048.pt`
- **BEATs**: `BEATs_iter3_plus_AS2M_finetuned_on_AS2M_cpt1.pt`
- **Whisper**: `openai/whisper-base`

3.3. Loss Functions

The baseline system uses cross-entropy loss on the 23 second-level classes. For the second experiment, two additional loss functions were explored. First, a hierarchical loss was explored, which adds a second cross-entropy term on the 5 top-level classes through a separate classification head. The main metric of the challenge, hierarchical F1, penalizes wrong top-level predictions more heavily than wrong second-level predictions. Penalizing the model for these mistakes during training was therefore a promising idea. Different weights (0.3, 0.7, 1.0, and 1.5) were tried for this experiment.

The second explored loss function was a supervised contrastive loss [13] applied on the latent representation before the final classifier. The motivation behind this was to see whether pulling samples from the same top-level class closer together and pushing different classes apart could improve classification. Contrastive loss was explored only in the multimodal setting.

The hierarchical loss improved the audio-only result from 75.30% to 76.05% at weight 0.7, and it also reduced the variance across folds. However, this gain was not observed in the multimodal setting. Table 2 presents the results of this experiment.

Table 2: Loss function comparison on BSD10k-v1.2 test split (5-fold Cross validation). Hierarchical F1 (%) score reported for audio-only and multimodal settings.

Loss	Audio-only	Multimodal
Baseline (CE)	75.30 ± 1.14	**79.01 ± 0.54**
+ hloss (w=0.3)	75.98 ± 0.78	78.59 ± 0.81
+ hloss (w=0.7)	**76.05 ± 1.10**	78.50 ± 0.38
+ hloss (w=1.0)	75.63 ± 0.74	77.76 ± 1.11
+ hloss (w=1.5)	75.22 ± 1.03	78.28 ± 0.45
+ contrastive (λ=0.2)	–	78.52 ± 0.37
+ contrastive (λ=0.5)	–	78.42 ± 0.69
+ contrastive (λ=1)	–	77.92 ± 0.51

These results suggest that the text modality already provides enough grounding, so the extra hierarchical supervision is not very useful when text is present.

3.4. Training Strategies

Inspecting the distribution of the BSD10k dataset revealed significant class imbalance. More specifically, the largest class, `fx-o` (objects/appliances), contains 1,211 samples, whereas the smallest class, `fx-a` (animals), contains only 166 samples. Class counts also vary within the same top-level category. For example, `sp-s` (solo speech) has 806 samples, while `sp-c` (conversation) has only 177. To address this imbalance, two weighting strategies were explored. The first scales the loss of each class to give more weight to the smaller classes. The explored method was inverse square-root frequency, which weights each class by the inverse square root of

its sample count. This gives a softer correction than plain inverse frequency.

The second strategy utilized the confidence scores from 1 to 5 provided by the BSD10k metadata. The motivation behind this was that more confidently labeled sounds should influence training more than the noisier, low scored sounds. To achieve this the loss of each sample is scaled by its confidence score. Three mappings were explored. The linear mapping sets the weight to $w = c/5$, scaling from 0.2 at confidence 1 up to 1.0 at confidence 5. The shifted mapping uses $w = (c - 1)/4$, which spans the full range from 0.0 to 1.0 and removes confidence 1 samples entirely. The binary mapping sets $w = 1.0$ for samples with confidence ≥ 3 and $w = 0.3$ otherwise. A combination of inverse square-root class weighting and the linear confidence mapping was also explored.

None of these strategies significantly improved the hF1 over the multimodal baseline. However, confidence weighting did improve the audio-only result, with linear confidence weighting reaching 76.52% compared to the 75.30% baseline. As with the hierarchical loss, this gain did not transfer to the multimodal setting. The results of this experiment are presented in Table 3.

Table 3: Training strategy comparison on BSD10k-v1.2 test split (5-fold Cross validation). Hierarchical F1 (%) score reported for audio-only and multimodal settings.

Strategy	Audio-only	Multimodal
CLAP (baseline)	75.30 ± 1.14	79.01 ± 0.54
Class weight (inverse sqrt)	75.28 ± 0.41	78.42 ± 0.32
Confidence (linear)	**76.52 ± 0.52**	**79.06 ± 0.65**
Confidence (shifted)	76.18 ± 0.57	78.99 ± 0.40
Confidence (binary)	75.61 ± 1.14	78.66 ± 1.12
Combination	76.16 ± 0.59	78.77 ± 0.42

These results suggest that the attention-based fusion and the text modality already handle much of the difficulty caused by class imbalance and label noise. This makes the explicit weighting redundant when text is present.

3.5. Data Augmentation

After the different weighting strategies did not work, several embedding-level data augmentation methods were explored. The first was mixup [14], which creates new training samples by taking a linear combination of two embeddings and their labels, controlled by a parameter α. Values of 0.1, 0.2, 0.3, and 0.4 were explored. The second method was CutMix [15], which replaces a contiguous segment of one embedding vector with the corresponding segment from another sample. The third was cross-modal swap, where the text embedding of a sample is replaced with the text embedding of another sample from the same class, combined with Gaussian noise ($\sigma = 0.001$) added to both modalities. The motivation here was to prevent the model from memorizing exact audio-text pairings and encourage it to map to the shared semantic content of each class. A combination of these methods (mixup, CutMix, and noise injection) was also tested.

Table 4 presents the results of the data augmentation experiments. Mixup with $\alpha = 0.2$ and cross-modal swap were the only methods that improved the multimodal performance over the baseline, with cross-modal swap producing the best individual model at $79.13\% \pm 0.49\%$ hF1. For the audio-only setting, mixup consistently improved performance across all tested α values, and the best audio-only model was obtained using mixup with $\alpha = 0.3$ ($76.11\% \pm 0.69\%$ hF1).

Table 4: Data augmentation comparison on BSD10k-v1.2 (5-fold CV). Hierarchical F1 (%) score reported for audio-only and multimodal settings.

Augmentation	Audio-only	Multimodal
CLAP (baseline)	75.30 ± 1.14	79.01 ± 0.54
Cross-modal swap + noise	–	**79.13 ± 0.49**
Mixup (α=0.1)	76.04 ± 0.58	78.72 ± 0.73
Mixup (α=0.2)	75.60 ± 0.53	79.11 ± 0.63
Mixup (α=0.3)	**76.11 ± 0.69**	78.35 ± 1.17
Mixup (α=0.4)	75.73 ± 0.85	78.71 ± 0.42
CutMix (α=0.2)	–	78.70 ± 1.20
Combined	–	78.80 ± 0.81

Cross-modal swap with noise provided the best single model (non-ensemble) result for the task and was included as the third submission system (NYU_Single).

3.6. Additional Training Data

The lowest class-level F1 scores were on Conversation/Crowd (`sp-c`) and Indoor soundscapes (`ss-i`). These classes are also among those with the fewest samples in the dataset. To improve performance on these classes, two sources of additional training data were explored.

The first source was external data from ESC-50 and FSD50K. Since neither dataset uses the BST taxonomy, their labels were manually mapped to the 23 BST classes. An overlap check with BSD10k removed 5,067 sounds shared through Freesound to prevent data leakage. The external datasets do not include Freesound-style metadata such as titles, tags, and descriptions. To generate the text embeddings, a single generic description was assigned to each BST class. For example, "animal sound recording" was assigned to all added `fx-a` samples. Three strategies were tried. The first added external samples only for the two lowest scoring classes, `sp-c` and `ss-i`. The second added all mapped external data. The third added all external data but down weighted the external samples in the loss to 0.3. In all three, the external data was added only to the training folds, while the validation and test folds stayed BSD10k only for a correct comparison to the baseline.

The second source was the larger BSD35k-CS dataset. Unlike ESC-50 and FSD50K, BSD35k-CS already uses the BST classes, so no label mapping was needed. Instead of adding the whole dataset, a class equalization strategy was used. For each fold, the sample count of each class in the BSD10k training split was computed. If a class had fewer samples than a target count, the missing samples were drawn from BSD35k-CS, with oversampling and replacement when needed. Two targets were tried. Median equalization raises each class to the median class count of the BSD10k split. Maximum equalization raises each class to the maximum class count.

Table 5 presents the results. None of the external data strategies improved the multimodal hF1 over the baseline. The drop was largest when adding all external data. This is likely because the generic class descriptions produce identical text embeddings for

every added sample in a class, which gives much weaker semantic contrast than the unique user written metadata in BSD10k. The manual label mapping also adds noise. For BSD35k-CS, maximum equalization added too many out of domain samples and lowered the multimodal hF1 to 76.90%. Median equalization worked better, reaching 78.63% multimodal hF1.

Table 5: Additional training data comparison on BSD10k-v1.2 test split (5-fold Cross validation). Hierarchical F1 (%) score reported for multimodal setting.

Strategy	Multimodal hF1 (%)
CLAP (baseline)	**79.01 ± 0.54**
+ External (weak classes)	78.78 ± 0.37
+ External (all, weighted)	78.12 ± 0.47
+ External (all)	77.99 ± 0.72
+ BSD35k-CS (median eq.)	78.63 ± 0.67
+ BSD35k-CS (maximum eq.)	76.90 ± 0.94

3.7. Ensemble

The final experiment explored combining several models through ensemble logit averaging. All the experiments produced models trained with different encoders, loss functions, training strategies, and data augmentation methods. These models make different errors and predictions across the 23 classes. The main motivation of this approach is that averaging their predictions will make a more robust classifier than any single model. To test this, the output logits of all models are averaged before the argmax is applied to get the final prediction.

The best ensemble was selected through a two-phase search performed using a Python script. First, all combinations of size two to five were evaluated by accuracy across the five cross-validation folds. The hierarchical F1 score was then computed for the top fifty candidates. A greedy forward selection step was also applied to test larger ensembles, but adding a sixth model did not improve performance, so the best ensemble contains five models.

Table 6 shows the results for the top five ensembles and their performance on the BSD10k test split. The first ranked ensemble consists of mixup (α=0.3), mixup (α=0.2) with hierarchical loss, balanced median sampling, ConvNeXt with mixup, and the fine-tuned CLAP model. The fine-tuned CLAP model refers to a version of the CLAP audio encoder fine-tuned end-to-end on BSD10k, which improved the audio-only accuracy but did not surpass the frozen baseline in the multimodal setting. This ensemble achieves 81.13% hF1, the largest improvement observed over the reproduced baseline of 79.01%, a gain of 2.12%.

4. RESULTS

Four systems were submitted to the challenge. Table 7 presents their hierarchical F1 scores on the BSD10k-v1.2 dataset. For diversity, only two ensemble systems were submitted. The third submission was the best performing single model, and the fourth was an experimental model that utilizes Whisper and mixup data augmentation to help the system differentiate speech classes better.

Per-class results for the first ensemble model reveal a gap between the strongest and weakest classes. The model performs very well on the instrument sample classes, reaching 97.63% on Wind (`is-w`), 95.89% on Piano/Keyboard (`is-k`), and 94.68% on String (`is-s`). These classes have clear and consistent acoustic features that the audio and text embeddings capture well.

The weakest classes are Conversation/Crowd (`sp-c`) at 52.16% and Indoor soundscapes (`ss-i`) at 61.99%. These are also among the least represented classes in the dataset. Their low scores are likely caused by both the small number of training samples and the overlap of their acoustic content with other classes. Crowd and indoor sounds can contain speech, music, and effects at the same time, which makes their classification more ambiguous.

Table 6: Top ensemble combinations on BSD10k-v1.2 (5-fold CV) test split. Hierarchical F1 (%) score reported for multimodal setting.

Models	hF1 (%)
Mixup 0.3, Mixup 0.2 + Hloss, Balanced median, ConvNeXt + Mixup, Fine-tuned CLAP	**81.13 ± 0.57**
Hloss 0.7, Mixup 0.2 + Class weight, Balanced median, Combined augmentation, ConvNeXt + Mixup	81.12 ± 0.75
Confidence weight, Mixup 0.2 + Class weight, Combined augmentation, ConvNeXt + Mixup, External weighted	81.11 ± 0.83
Confidence weight, Conf + Hloss, Mixup 0.3, Balanced median, MATPAC	81.11 ± 0.58
Mixup 0.2 + Class weight, Cross-modal swap, Combined augmentation, CutMix, MATPAC	81.10 ± 0.40

Table 7: Final submitted systems on BSD10k-v1.2 (5-fold CV). Hierarchical Precision, Recall, and F1 (%) scores reported for multimodal setting.

System	hP (%)	hR (%)	hF1 (%)
Ensemble rank 1	82.18	80.49	**81.13 ± 0.57**
Ensemble rank 2	81.81	80.85	81.12 ± 0.75
Single (cross-modal swap)	79.84	78.99	79.13 ± 0.49
Three-modality + Mixup	79.49	78.67	78.83 ± 1.14

5. CONCLUSION

This technical report presented a systematic investigation to improve upon the HATR baseline for the DCASE 2026 Task 1 on heterogeneous audio classification. Several directions were explored, including alternative audio encoders, loss functions, training strategies, data augmentation, and the incorporation of additional training data. No significant improvements were observed over the multimodal baseline's hF1 score. Alternative encoders performed worse than CLAP. This suggests that CLAP's joint audio-text pretraining provides a good grounding for this task, especially when text is involved. The biggest improvement came from combining diverse models through ensemble logit averaging. The best ensemble reached 81.13% hF1, a 2.12% improvement over the reproduced baseline. Notably, this ensemble included models that were individually weaker than the baseline. This shows that diversity across encoders and training strategies matters more than the strength of each model. Future work could focus on the weakest classes like Conversation/Crowd and Indoor soundscapes, which remain difficult due to limited data and acoustic overlap with other classes.

6. REFERENCES

[1] P. Anastasopoulou, X. Serra, and F. Font, "A general-purpose sound taxonomy for the classification of heterogeneous sound collections," In press. [Online]. Available: https://www.researchsquare.com/article/rs-7206795/v1

[2] P. Anastasopoulou, F. A. Dal Rí, X. Serra, and F. Font, "Hierarchical and multimodal learning for heterogeneous sound classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2025.

[3] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[4] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," 2022. [Online]. Available: https://arxiv.org/abs/2212.09058

[5] T. Pellegrini, I. Khalfaoui-Hassani, E. Labbé, and T. Masquelier, "Adapting a convnext model to audio classification on audioset," 2023. [Online]. Available: https://arxiv.org/abs/2306.00830

[6] A. Quelennec, P. Chouteau, G. Peeters, and S. Essid, "Masked latent prediction and classification for self-supervised audio representation learning," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2025, p. 1–5. [Online]. Available: http://dx.doi.org/10.1109/ICASSP49660.2025.10887666

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[8] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[9] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[10] P. Anastasopoulou, J. Torrey, X. Serra, and F. Font, "Heterogeneous sound classification with the Broad Sound Taxonomy and Dataset," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024.

[11] P. Anastasopoulou and F. Font Corbera, "Bsd35k-cs (broad sound dataset 35k - crowd sourced)," March 2026. [Online]. Available: https://doi.org/10.5281/zenodo.19187100

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[13] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2018. [Online]. Available: https://arxiv.org/abs/1710.09412

[15] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," 2019. [Online]. Available: https://arxiv.org/abs/1905.04899