POSTERIOR STACKING AND CONSERVATIVE METADATA GATING FOR HETEROGENEOUS AUDIO CLASSIFICATION

Technical Report

Minkyu Kil[1,2], Seunggyu Jeong[1,2], Seong-Eun Kim[1,2]

[1] Medisensing, Seoul, Korea
[2] Seoul National University of Science and Technology, Seoul, Korea
henry42471@gmail.com, wa3229433@gmail.com, sekim@seoultech.ac.kr

ABSTRACT

This technical report describes the Medisensing-SeoulTech submission to DCASE 2026 Task 1, Heterogeneous Audio Classification. The submitted systems classify each evaluation item into one of 23 second-level Broad Sound Taxonomy classes. The final package contains four systems. Systems 1 and 2 are high-scoring weighted posterior stacks built from frozen audio and audio-text embedding heads with conservative same-parent metadata gates. System 3 is a Larger-CLAP residual classifier with a TF-IDF metadata gate, included as a complementary metadata-aware variant with broader label-symbol coverage. System 4 is a complementary audio-only posterior ensemble. Across metadata-assisted systems, metadata can refine the second-level class only inside the audio-predicted top-level parent. No evaluation-set ground-truth labels or manual evaluation-set annotations are used for training, threshold selection, system selection, or reporting.

Index Terms— Heterogeneous audio classification, Broad Sound Taxonomy, posterior stacking, metadata gating, CLAP, TF-IDF

1. INTRODUCTION

DCASE 2026 Task 1 evaluates heterogeneous audio classification under the Broad Sound Taxonomy (BST) [1, 3, 4, 5]. Each system predicts one of 23 second-level labels under five parent categories: music, instrument samples, speech, sound effects, and soundscapes. Since the ranking metric is macro hierarchical F-score [1, 12], cross-parent mistakes are more costly than many within-parent confusions. Our final submission therefore emphasizes hierarchy-aware posterior combination and conservative correction rules.

The audio collection is heterogeneous in both content and metadata quality. Some classes are acoustically close, such as speech subtypes or soundscape subtypes, while other labels are separated mainly by semantic context. A robust system therefore needs strong acoustic representations, but also needs a cautious way to use text fields without letting text dominate the acoustic decision.

The final package contains four systems in the Kil_Medisensing_task1 label family, indexed from 1 to 4. We combine two strong posterior-stack systems with complementary systems based on different feature and decoding choices. The goal is to keep the best internal development variants while reducing the chance that all four submissions share the same decoding failure mode.

2. DATA AND EXTERNAL RESOURCES

BSD10k-v1.2 is used as the clean anchor data. It contains 10,956 rows with 23 second-level labels and 2,481 uploaders. The local BSD10k split used in several experiments follows an 80/10/10 seed-42 split with 8,756 training rows, 1,085 validation rows, and 1,115 test rows.

BSD35k-CS [6] is treated as noisy auxiliary data, not as equally clean ground truth. When it is used, it is handled with reliability filters, sample weights, or fixed data-source policies. Metadata fields are used only as noisy helper signals.

External rows are admitted only through fixed provenance and mapping rules. The submitted systems use curated FSD50K [7] support rows, and selected posterior-stack branches also use curated rows from UrbanSound8K [8], FSDKaggle2018 [9], and FSDKaggle2019 [10]. Frozen pretrained embedding models are used as feature extractors. No evaluation-set ground-truth labels or manual evaluation-set annotations are used for training, threshold selection, system selection, or reporting.

Table 1: Data resources and branch-specific filtered subsets. Rows are not additive across rows.

Source	Use	Rows
BSD10k-v1.2	clean anchor	10,956
BSD35k-CS	noisy auxiliary	33,829
BSD35k subset	parent specialist	29,214
External subset	mapped support	5,150
FSD50K text subset	metadata helper	9,071

For external support rows, class mapping is performed before model comparison. Rows with unresolved audio paths, ambiguous mapping status, review-required mapping, or missing feature caches are removed. This rule-based filtering is intentionally conservative: it sacrifices some auxiliary data to reduce the risk of injecting wrong labels into rare BST classes.

3. FEATURE REPRESENTATIONS

The posterior-stack systems use cached frozen representations from LAION-CLAP audio embeddings [11], Larger-CLAP audio embeddings, M2D-CLAP, BEATs, ATST, and PaSST. Additional CLAP-derived posterior heads are used only as fixed feature extractors and are not fine-tuned.

The metadata helper uses title, tags, and description. The main text helper is a sparse TF-IDF classifier over word 1–2 grams and character 3–5 grams. We prefer this conservative sparse helper because task metadata may be missing, promotional, or semantically broader than the acoustic event.

The posterior features are 23-way class probabilities. In the main stack, the input to the stacker is the concatenation of posterior vectors from candidate systems. The stacker is trained to learn which candidate probabilities are reliable for each class and which candidates should receive lower weight. This design keeps the final model interpretable at the level of candidate posteriors rather than raw neural activations.

4. SUBMITTED SYSTEMS

Table 2 summarizes the final four submitted systems. Systems 1 and 2 are the main high-scoring posterior-stack systems; Systems 3 and 4 are complementary variants selected for alternative output profiles.

Table 2: Submitted systems. Metadata fields are title, tags, and description.

System	Short	Role
task1_1	MGateStack	Weighted stack plus target-masked metadata gate
task1_2	ParentSpec	Stack plus raw parent-specialist correction
task1_3	MTFGate	Larger-CLAP residual MLP plus same-parent metadata gate
task1_4	AudEns	Complementary audio-only greedy ensemble

4.1. System 1: Target-masked posterior stack

System 1 is the main posterior stack. Each candidate model emits a 23-dimensional posterior vector. The stack receives the concatenated posterior features and learns a weighted combination using development data. A TF-IDF helper may change only selected leaf classes with positive out-of-fold evidence. The final target mask allows metadata-assisted changes only for fx-o, m-m, and ss-u, and only inside the audio-predicted parent.

The target mask is deliberately narrow. It is not a general text-fusion mechanism; it is a small correction layer for leaf labels where text helped consistently during development. If the metadata helper suggests a label outside the audio-predicted parent, the correction is rejected.

4.2. System 2: Raw parent specialist

System 2 starts from the same posterior-stack family and adds a raw parent-specialist branch. This branch is trained with cached raw embeddings and metadata-derived representations. The specialist is used as a correction layer, not as a replacement classifier. It can refine a second-level label only when the proposed class remains inside the same BST parent.

The parent-specialist branch was selected because many remaining errors in the high-scoring stack are local confusions within a parent. Instead of allowing a global override, the specialist compares candidate leaves only inside the audio-predicted parent.

4.3. System 3: Larger-CLAP with TF-IDF gate

System 3 is based on a residual MLP over frozen 512-dimensional Larger-CLAP audio embeddings. It combines BSD10k-v1.2, reliability-weighted BSD35k-CS, and a fixed FSD50K subset. A sparse TF-IDF metadata classifier is allowed to override the audio prediction only when the helper confidence and margin pass fixed thresholds and the proposed class remains in the same parent. The final output covers all 23 second-level classes.

This system is included as a complementary metadata-aware Larger-CLAP variant. In package validation, it provides broader label-symbol coverage than the main stack outputs. Its output differs from System 1 on about one third of evaluation rows, so it provides an alternative error profile while preserving the same broad data-use policy.

4.4. System 4: Audio greedy ensemble

System 4 is an audio-only posterior ensemble. It combines ridge and hierarchical ridge heads trained on CLAP, M2D-CLAP, CLAP-fused, and LAION-CLAP-derived embeddings. The ensemble is selected greedily on BSD10k validation hierarchical F-score and decoded by argmax. It does not use challenge metadata at inference time.

System 4 is the most different output in the package. It is a strong common-condition audio alternative and does not depend on metadata at inference time. Because it is selected from a different audio posterior family, it is less likely to share the exact same decoding failures as the main stack.

5. TRAINING, INFERENCE, AND GATING

For posterior-stack systems, candidate heads are trained on cached features and produce second-level probability vectors. The stacker combines these posteriors and keeps the predicted top-level parent as a constraint for later gates. Metadata gates operate after the audio or stacked posterior has selected a parent. A metadata helper can refine the leaf class only if the change stays under that parent and passes the configured confidence, margin, rank, or agreement checks.

Candidate heads use ridge, hierarchical ridge, residual MLP, and related lightweight classifiers depending on the representation. Hyperparameters and candidate subsets are selected on development folds with hierarchical F-score as the main criterion. Selection emphasizes stable improvements rather than isolated row-level changes, especially for helper gates.

At inference time, each system outputs a single CSV with id, predicted_bst_second_level_class, and prediction_score. Prediction scores are uncalibrated posterior confidence values from each system and are used only as optional

confidence outputs. No post-hoc class quota or evaluation-set label-distribution constraint is applied in the final package.

6. DEVELOPMENT EVIDENCE

Table 3 summarizes the common leakage-controlled BSD10k evidence used to check the final system choices. All rows are evaluated under the same 10k-data condition. Hierarchical F-score (hF), fine accuracy, and top-level accuracy are reported in percent. The delta column gives the absolute hF difference from the Larger-CLAP baseline.

Table 3: Common leakage-controlled BSD10k evidence. Metrics are in percent.

System	hF	Δ hF	Fine acc.	Top acc.
Larger-CLAP baseline	77.5156	0.0000	79.4619	88.0717
task1_1	80.4864	+2.9708	81.8834	89.6861
task1_2	80.4370	+2.9214	81.9731	89.6861
task1_3	72.8909	-4.6247	75.4260	86.3677
task1_4	80.9279	+3.4123	82.7803	90.6726

Systems 1, 2, and 4 improve over the Larger-CLAP baseline under the common condition. System 3 is weaker under this BSD10k metric and is submitted as a complementary metadata-aware Larger-CLAP variant. The final CSV sanity check confirms broader label-symbol coverage and a substantially different output profile, without using evaluation-set ground-truth labels.

7. SUBMISSION PACKAGE CHECKS

The submission package follows the official DCASE 2026 submission structure [2]: one Task 1 technical report in the task folder and four system folders, each containing the corresponding meta-information YAML and output CSV. Each submitted CSV contains 3,246 rows with the required fields id, predicted_bst_second_level_class, and prediction_score. All IDs are unique, all predicted labels belong to the official 23 second-level BST classes, and all scores are finite values within $[0, 1]$.

The submitted outputs were also checked for complementary output profiles. System 1 and System 2 differ on 29 of 3,246 rows. System 3 differs from System 1 on 1,080 rows (33.27%), System 4 differs from System 1 on 1,873 rows (57.70%), and System 2 differs from System 3 on 1,080 rows (33.27%). These differences indicate that the complementary systems provide alternative output profiles without using evaluation-set ground-truth labels.

8. COMPLEXITY AND REPRODUCIBILITY

Most submitted systems operate on cached embeddings and posterior files from several frozen feature extractors. A single end-to-end parameter or MAC count is therefore not directly comparable to a single neural network. The trainable Larger-CLAP plus metadata-gate pipeline reports 10,704,631 trainable parameters and 10,686,953 MACs, excluding the frozen embedding extractor. The remaining stack components are lightweight linear or calibration heads over cached features.

Development splits, selected thresholds, output manifests, and CSV validation summaries are stored in the local experiment records. External rows are selected by fixed rules before system comparison, and evaluation-set labels are not used for training, threshold selection, system selection, or report numbers.

9. CONCLUSION

The final Medisensing-SeoulTech submission combines strong posterior-stacking systems with complementary metadata-aware and audio-only alternatives. Systems 1, 2, and 4 improve over the Larger-CLAP baseline under the common BSD10k evidence. System 3 adds broader label-symbol coverage with a conservative Larger-CLAP plus TF-IDF gate. System 4 adds a substantially different audio-only greedy ensemble. Across all metadata-assisted systems, same-parent gating is used to exploit helpful text cues without allowing text to override the audio-predicted top-level taxonomy.

10. REFERENCES

[1] DCASE Community, "DCASE 2026 Challenge, Task 1: Heterogeneous Audio Classification," 2026. [Online]. Available: https://dcase.community/challenge2026/task-heterogeneous-audio-classification

[2] DCASE Community, "DCASE 2026 Challenge submission instructions and technical report template," 2026. [Online]. Available: https://dcase.community/challenge2026/submission

[3] P. Anastasopoulou, X. Serra, and F. Font, "A General-Purpose Sound Taxonomy for the Classification of Heterogeneous Sound Collections," Research Square preprint, in press, 2026, article rs-7206795/v1.

[4] P. Anastasopoulou, J. Torrey, X. Serra, and F. Font, "Heterogeneous Sound Classification with the Broad Sound Taxonomy and Dataset," in Proc. DCASE Workshop, 2024.

[5] P. Anastasopoulou, F. A. Dal Ri, X. Serra, and F. Font, "Hierarchical and Multimodal Learning for Heterogeneous Sound Classification," in Proc. DCASE Workshop, 2025.

[6] P. Anastasopoulou and F. Font Corbera, "BSD35k-CS (Broad Sound Dataset 35k - Crowd Sourced)," Zenodo, Mar. 2026, doi: 10.5281/zenodo.19187100.

[7] E. Fonseca et al., "FSD50K: An Open Dataset of Human-Labeled Sound Events," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.

[8] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in Proc. ACM Multimedia, 2014.

[9] DCASE 2018 Challenge Task 2, General-purpose audio tagging of Freesound content with AudioSet labels.

[10] DCASE 2019 Challenge Task 2, Audio tagging with noisy labels and minimally supervised data.

[11] B. Elizalde et al., "CLAP: Learning Audio Concepts from Natural Language Supervision," 2023.

[12] S. Kiritchenko, S. Matwin, and A. F. Famili, "Functional Annotation of Genes Using Hierarchical Text Categorization," in Proc. ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, 2005.