ACACIA: AUDIO CLASSIFICATION USING ATTENTION-BASED CLEANED MULTIMODAL EMBEDDINGS

Technical Report

Francesco Colotti[1,*], *Kerstin Markl*[1,*], *Riccardo Casciotti*[1,*], *Javier Naranjo-Alcazar*[2,*],

[1] Tampere University, Tampere, Finland
{francesco.colotti, kerstin.markl, riccardo.casciotti}@tuni.fi
[2] Instituto Tecnologico de Informatica (ITI), Valencia, Spain
{jnaranjo}@iti.es

ABSTRACT

This report presents our pipeline for Task 1 of DCASE 2026, the first edition of the Heterogeneous Audio Classification challenge based on the Broad Sound Taxonomy. Our approach exploits the complementarity of acoustic and textual information to perform audio classification using sound recordings, user-provided tags and textual descriptions. To improve the quality of the textual modality, we propose a pre-processing pipeline that excludes non-informative content from tags and descriptions prior to encoding. The proposed architecture comprises three branches that independently encode audio, tags, and descriptions, whose representations are fused into a shared embedding space. To exploit the hierarchical organization of the Broad Sound Taxonomy, the model performs classification using embeddings in the hyperbolic space. Experimental results show that our strategy surpasses the official baseline, demonstrating the potential of our text processing pipeline and of our multimodal approach for audio classification.

***Index Terms*—** Heterogeneous Audio Classification, DCASE, multi-modal

1. INTRODUCTION

Driven by deep learning architectures [1, 2] and the proliferation of large-scale audio datasets [3, 4, 5], environmental sound classification has advanced rapidly in recent years. Yet, most established benchmarks still rely on idealized conditions, assuming homogeneous data distributions and well-curated, flat annotations. To address this gap, DCASE 2026 Task 1: Heterogeneous Audio Classification [6] introduces a new benchmark built upon the Broad Sound Taxonomy (BST), marking the first edition of this task. BST defines a hierarchical label space composed of five top-level categories and twenty-three second-level categories, enabling evaluation at multiple semantic resolutions [7]. The challenge is supported by two complementary Freesound-derived datasets: a curated subset with higher annotation reliability [8] and a larger crowd-sourced collection [9] reflecting realistic noise, ambiguity, and labeling inconsistencies. Importantly, the official evaluation metric explicitly accounts for the taxonomy structure, penalizing hierarchical misclassifications more strongly than flat label errors, thereby encouraging methods that model both coarse and fine-grained semantics.

A characteristic of this task is the availability of textual metadata associated with audio recordings, including user-created tags and descriptions. While containing at times irrelevant information, such metadata provides semantic context that complements purely acoustic representations. This setting naturally motivates multimodal approaches that jointly leverage audio signals and textual information, particularly when combined with mechanisms for handling label noise and semantic ambiguity.

In this technical report, we present **ACACIA** (**A**udio **C**lassification using **A**ttention-Based **C**leaned Mult**I**mod**A**l Embeddings), a multimodal framework designed to exploit both acoustic and textual modalities while explicitly modelling the hierarchical structure of the dataset. A preprocessing stage is first applied to textual descriptions and tags to reduce noise and remove non-informative content before encoding. The proposed system then processes text-audio data through dedicated encoders for audio, cleaned tags, and cleaned textual descriptions, followed by fusion into a shared representation space, whose embeddings are mapped into hyperbolic space prior to classification. Generally, hyperbolic neural networks have led to performance improvements in various tasks using, or implying, hierarchical settings due to their capacity to faithfully represent hierarchical relationships in data [10] [11] [12], hence being a good fit for this task.

Apart from ACACIA, this report summarizes and analyses the alternative approaches explored during the development phase of the challenge, including additional features and a Mixture-of-Experts (MoE) scheme. Experimental comparisons highlight the relative strengths and limitations of these approaches under the heterogeneous and hierarchy-aware evaluation setting defined by the task. Overall, our findings indicate that carefully integrating cleaned textual metadata with acoustic representations yields consistent gains over audio-only baselines.

The rest of the manuscript is organised as follows: our methodology, including metadata preprocessing, the ACACIA architecture and other approaches, is presented in Section 2, results are shown and briefly discussed in Section 3 and Section 4 concludes the report.

2. METHOD

In this section, we describe the proposed ACACIA framework for heterogeneous audio classification, as well as a set of exploratory approaches investigated during the development of the system for DCASE 2026 Task 1.

We first introduce ACACIA (see Section 2.3), the main submitted system, which follows a multimodal architecture that combines

*Equal contribution.

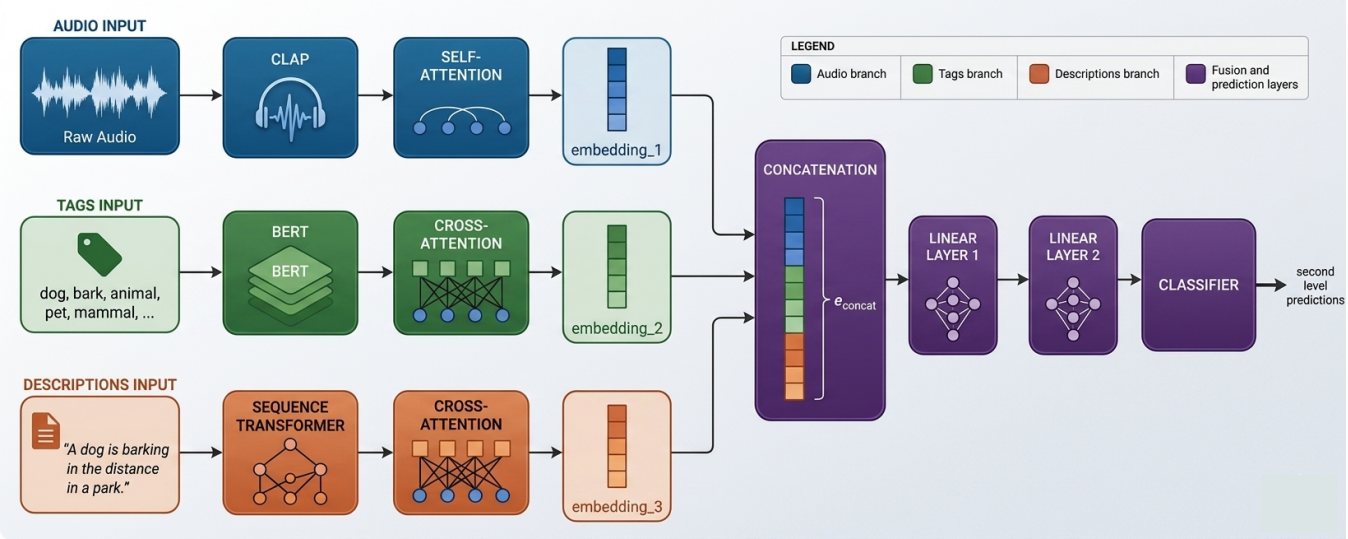

Figure 1: Overview of the proposed ACACIA architecture. The model consists of three parallel branches encoding audio, tags, and textual descriptions. The audio signal is processed using CLAP [13] embeddings followed by a self-attention module, while tags are encoded using BERT [14] and descriptions using a sequence transformer; both textual branches are further refined through cross-attention layers. The resulting embeddings are concatenated and passed through fully connected layers to produce hierarchical predictions at both the top-level and second-level of the Broad Sound Taxonomy.

audio, user-provided tags, and textual descriptions through dedicated encoding branches.

In addition to the main approach, we also present several alternative configurations explored during the challenge period in Subsection 2.4), including variations in text encoding strategies, and fusion mechanisms. These exploratory models provide additional insights into the design choices that are most effective for heterogeneous and weakly structured audio metadata.

2.1. Dataset

For our experimental evaluation, we utilized the BSD10k-v1.2 dataset [15, 16]. We deliberately discarded the larger BSD35k-CS dataset [17], as our preliminary tests showed no significant improvement in model performance; furthermore, the ground-truth annotations of BSD35k-CS have not been officially verified, raising concerns about label noise. In addition to the standard cross-validation protocol implemented in the baseline code with 5 folds, we introduced a rigorous, user-aware train/val/test partition. By explicitly structuring the splits based on user identity, we prevented any user overlap across the partitions. This validation scheme ensures a more realistic and robust evaluation, closely mimicking actual deployment scenarios where the system must generalize to entirely unseen users. For clarity in the subsequent discussion and results, we hereafter refer to the 5-fold cross-validation scheme as the *Baseline* partition, and the proposed user-aware split as the *Uploader* partition.

2.2. Preprocessing Step

The provided metadata contains tags and descriptions that contain irrelevant information (e.g. Digital Audio Workstation in tags, social media links in descriptions). We thus create a second version of the metadata by excluding tags not in the Open English WordNet [18] list of words, and by substituting descriptions with less verbose LLM-rewritten versions that exclude information not directly related to sounds in the clip. We use the publicly available Meta LLaMA 3.2-3B-Instruct with few-shot prompting for what concerns the descriptions. We note that the process may exclude relevant tags and produce partially hallucinated descriptions; nonetheless, we train all the presented systems with the processed metadata.

2.3. ACACIA

We design a multimodal architecture to classify the provided clips using both sonic and textual information, with audio, tags and descriptions as inputs. The model consists of three encoder/attention paths, one per input, whose outputs are concatenated and fed to a Multilayer Perceptron (MLP). An overview of the system is visible in Fig 1. The encoders are input-specific: raw audio is resampled and encoded using CLAP[1] [13], each tag is encoded with BERT[2] [14] and the mean of the embeddings is used for classification, descriptions are encoded with MPNet[3] [19]. The encoded audio and text embeddings have dimension 512 and 768, respectively. The attention stage consists in a self-attention layer for audio, and cross-attention between audio and text embeddings for the tags and description paths; post-attention embeddings, each having dimension 512, are concatenated and given as input to the MLP. Prior to classification, the embeddings are mapped to hyperbolic space, using the Lorentz model. A Lorentz Multinomial Logistic Regression layer [20] is used to perform the final classification on the finer class level of the taxonomy.

[1] Using the *630k-audioset-fusion-best.pt* checkpoint from LAION-AI
[2] Using the *google-bert/bert-base-uncased* pretrained weights
[3] From the *sentence-transformers* package `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

Submission #	Splits	Model	Mode	Hier. Accuracy	Hier. F1
-	*Baseline*	Baseline	Audio	77.36% ± 0.71%	76.11% ± 0.45%
-	*Baseline*	Baseline	Multimodal	79.71% ± 0.82%	78.76% ± 0.79%
1	***Baseline***	**ACACIA**	**Multimodal**	**84.09% ±1.01 %**	**83.52% ±1.13 %**
2	*Uploader*	ACACIA	Multimodal	80.25%	79.76%
3	*Baseline*	EBC	Multimodal	80.85% ± 0.82%	79.67% ± 1.00%
-	*Uploader*	EBC	Multimodal	80.68%	78.57%
-	*Baseline*	H-MoE	Multimodal	83.52±0.32%	79.58±0.47%
4	*Uploader*	H-MoE	Multimodal	81.69%	73.10%

Table 1: Performance comparison on BSD10k between audio-only and multimodal baseline configurations. Results are reported as mean ± standard deviation.

2.4. Other approaches

Apart from the aforementioned architecture, we experimented with two other solutions using, respectively, additional features and a Hierarchical Mixture-of-Experts approach.

2.4.1. Enhanced Base Classifier

The idea behind the Enhanced Base Classifier (EBC) is to inform the model with additional text information that is derived from psychoacoustic features of the audio clips. We extract multiple features commonly used for audio- and speech classification, encompassing duration, loudness, loudness range, spectral contrast, spectral bandwidth, spectral centroid, onset strength variance, spectral roll-off, zero crossing rate, root mean square energy, harmonic- and percussive ratio.[4] Missing values are replaced by averaging over the respective feature column in the dataset.

The feature values are then mapped to pre-defined textual categories according to feature-related threshold, and subsequently used as inputs to a Sentence-BERT [21] model[5]. Relying on the pre-implemented attention fusion mechanism of the Baseline Classifier, we fuse the additional Sentence-BERT text embeddings with CLAP [6] [13] audio embeddings. In parallel, we merge separate text embeddings for tags and description as described in Section 2.3 with the baseline attention fusion mechanism. In a second step, we fuse the resulting audio and text embeddings with the same pre-implemented method.

2.4.2. Hierarchical Mixture of Experts

Hierarchical Mixture of Experts is an approach that solves the problem of hierarchical learning through separate network classifiers for each top category of the dataset. It is made up of a router network, which is trained only to distinguish the category a sample belongs to, and a series of experts, each trained only on the classes within the corresponding category. The network receives as input the CLAP [13] embeddings obtained from audio, tags, and descriptions pre-processed as specified in Section 2.2.

The three separate features obtained from audio, tags, and descriptions are then fused through feature-wise multimodal attention, which calculates the importance of each feature, and then goes through GLU feed-forward blocks [22], projections in GLU feed-forward blocks [23], and ReZero [24] inspired residual scaling. At this point, they are fed to the router network: a deep gated residual feature processing network, which will calculate the corresponding expert (category) to use for the current sample through a linear classifier. Each expert is composed of residual GLU Blocks and a classifier head to predict the class for the current sample.

Additionally, this approach uses a teacher-enforced hierarchical training procedure, where the correct expert for the current sample is enforced during training, regardless of the prediction of the router; this ensures each expert has seen only samples belonging to its category.

The loss of the whole architecture is calculated by combining the loss of the router and the loss of the expert, and the final accuracy is obtained based on the accuracy of the router and the accuracy of the expert. In particular, a sample is considered correctly classified only if both the prediction of the router and the prediction of the expert are correct. During the evaluation phase, the "performance score" is calculated by multiplying the "performance score" of the expert by the "performance score" of the router.

3. RESULTS

We present the results of all the mentioned approaches in Table 1, specifying whether they are based on the *Baseline* or the *Uploader* dataset splits.

In general, all approaches trained on the *Baseline* split outperformed the baseline in terms of hierarchical accuracy and hierarchical F1-score. The ACACIA architecture obtained the overall best Hierarchical F1 score, showing the effectiveness both of our metadata preprocessing pipeline and the attention-based encoders, and outperformed the baseline performance on the *Baseline* splits also when trained and tested on the *Uploader* splits, highlighting a good capacity for generalization on unseen data.

Results achieved on the *Uploader* splits are lower than those of the same models trained on the *Baseline* splits but provide a better estimate of how the models would perform on truly unseen data. This confirms our initial hypothesis that shared uploaders bias the results.

While not included, we mention that multimodal inputs proved to be essential for achieving better performance, with all models showing significantly degraded performance when trained and evaluated solely with audio; this is expected due to the complementary information provided by sound recordings, tags and descriptions.

[4] Where possible, features are extracted using the *librosa* package: `https://librosa.org/doc/0.11.0/index.html`. Note that most of the features are represented as mean values.

[5] From the *sentence-transformers* package `https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased`

[6] Using the *630k-audioset-fusion-best.pt* checkpoint from LAION-AI

4. CONCLUSION

This report outlined our three presented approaches for the DCASE 2026 Task 1. The triad shares a metadata pre-processing pipeline designed to discard irrelevant information from tags and descriptions prior to embedding extraction, but each solution differs architecturally. We show that our approaches surpass the baseline performance.

5. ACKNOWLEDGMENT

The work carried out by Javier Naranjo-Alcazar was partially funded by the Spanish Ministry of Science, Innovation and Universities through the José Castillejo mobility grant for young doctors 2024 (CAS24/00150), supporting a research stay at a foreign higher education and research institution and Valencian Institute for Business Competitiveness (IVACE).

The team would like to thank, in no particular order, Michael Neri, David Diaz-Guerra, and Manu Harju for their support and incredible insight.

6. REFERENCES

[1] 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.

[2] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[3] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[5] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 626–630.

[6] 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

[7] 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.

[8] 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.

[9] 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

[10] M. GhadimiAtigh, J. Schoep, E. Acar, N. v. Noord, and P. Mettes, "Hyperbolic Image Segmentation," Mar. 2022, arXiv:2203.05898 [cs]. [Online]. Available: http://arxiv.org/abs/2203.05898

[11] F. G. Germain, G. Wichern, and J. L. Roux, "Hyperbolic Unsupervised Anomalous Sound Detection," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2023, pp. 1–5, iSSN: 1947-1629. [Online]. Available: https://ieeexplore.ieee.org/document/10248092/

[12] D. Petermann, G. Wichern, A. Subramanian, and J. L. Roux, "Hyperbolic Audio Source Separation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/10094943/

[13] 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 *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[15] 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.

[16] 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.

[17] 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

[18] J. P. McCrae, A. Rademaker, F. Bond, E. Rudnicka, and C. Fellbaum, "English WordNet 2019 – an open-source WordNet for English," in *Proceedings of the 10th Global Wordnet Conference*, P. Vossen and C. Fellbaum, Eds. Wroclaw, Poland: Global Wordnet Association, July 2019, pp. 245–252. [Online]. Available: https://aclanthology.org/2019.gwc-1.31/

[19] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 16 857–16 867. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

[20] A. Bdeir, K. Schwethelm, and N. Landwehr, "Fully hyperbolic convolutional neural networks for computer vision," in *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., vol. 2024, 2024, pp. 47 687–47 711. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2024/file/d0e83f2f6efb967577a7ec4239edefd5-Paper-Conference.pdf

[21] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[22] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[23] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[24] T. Bachlechner, B. P. Majumder, H. Mao, G. Cottrell, and J. McAuley, "Rezero is all you need: Fast convergence at large depth," in *Uncertainty in artificial intelligence*. PMLR, 2021, pp. 1352–1361.