Hierarchical and Multimodal Learning for Heterogeneous Sound Classification

Panagiota Anastasopoulou[1], *Francesco Ardan Dal Rí*[1,2], *Xavier Serra*[1], *Frederic Font*[3]

[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2]DISI - Dept. of Information Engineering and Computer Science, University of Trento, Italy
[3]Phonos Fundació Privada, Barcelona, Spain
panagiota.anastasopoulou@upf.edu, francesco.dalri-2@unitn.it, xavier.serra@upf.edu, frederic.font@upf.edu

***Abstract*—This paper investigates multimodal and hierarchical classification strategies to enhance performance in real-world sound classification tasks, centered on the two-level structure of the Broad Sound Taxonomy. We propose a framework that enables the system to consider high-level sound categories when refining its predictions at the subclass level, thereby aligning with the natural hierarchy of sound semantics. To improve accuracy, we integrate acoustic features with semantic cues extracted from crowdsourced textual metadata of the sounds such as titles, tags, and descriptions. During training, we utilize and compare pre-trained embeddings across these modalities, enabling better generalization across acoustically heterogeneous yet semantically related categories. Our experiments show that the use of text-audio embeddings improve classification. We also observe that, although hierarchical supervision does not significantly impact accuracy, it leads to more coherent and perceptually structured latent representations. These improvements in classification performance and representation quality make the system more suitable for downstream tasks such as sound retrieval, description, and similarity search.**

***Index Terms*—Classification, Hierarchical, Multimodal, Taxonomy**

1. INTRODUCTION

Automatic analysis of heterogeneous sound types remains a central problem in audio classification, spanning domains such as music, speech, and environmental acoustics [1]–[3]. Addressing this challenge requires broader classification frameworks and taxonomies capable of handling arbitrary input sounds. Unlike approaches tailored to particular types of sound, in heterogeneous sound classification, the goal is to develop a high-level classifier that generalizes across diverse acoustic inputs. This presents challenges due to the high intra-class variability, the varying audio qualities (quality, recording conditions, etc.), and the presence of ambiguous cases [4]–[7]. In this paper, we focus on a heterogeneous classification framework, where we use the Broad Sound Taxonomy (BST) [8] which organizes sounds into a two-level hierarchical structure, with 5 top-level and 23 second-level categories.

Human auditory perception naturally distinguishes sound classes at multiple abstraction levels [9], [10]. Therefore, the hierarchical information derived from the taxonomic structure can be valuable during the training of classifiers, e.g. [11]–[13]. In addition, in cases where acoustic signals are ambiguous or acoustically similar across different sources, introducing semantic information from human-annotated metadata can help disambiguate and improve classification. This is especially relevant with taxonomies such as the BST, where audio classes are defined according to sound semantics. Freesound [14], which contains a large and heterogeneous collection of user-contributed audio material, has recently adopted the BST as its organizational scheme. Freesound hosts audio recordings annotated with titles, tags, and free-text descriptions. The use of metadata is also essential in professional audio collections that involve music, instrument samples, and sound effects. Previous research shows that a common cause of misclassifications is acoustic ambiguity resulting from similarities in sound characteristics or shared sound sources between categories [6]. This is particularly relevant in scenarios involving heterogeneous classification, where classes and audio qualities vary widely. Thus, incorporating additional modalities and semantic context is expected to enhance model performance.

In our work, we explore both hierarchical and multimodal approaches to improve classifier performance, and also take into account annotation confidence scores (i.e. scores that reflect the certainty of the annotator for each sound annotation) to filter training data and explore its impact in the classifier performance. The experimental results show that the combination of text and audio embeddings improves classification, while hierarchical supervision helps produce latent representations that are more coherent and perceptually well-structured. Our approach aims not only to improve classification accuracy but also to make the system better applicable to different downstream tasks, such as computing context-informed similarity between sounds, generating descriptive characterizations, and enhancing sound retrieval.

2. BACKGROUND

Hierarchical classification has been widely applied across various domains, including musical instrument families, music genre recognition, sound effect organization, and acoustic scene analysis, where labels can be naturally organized into structured taxonomies [15]–[18]. In such tasks, these hierarchical structures align with the way humans perceive and categorize auditory information, moving from broad categories (e.g. *animal*) to more specific ones (e.g. *cat*). Unlike flat classifiers that treat all labels independently, hierarchical approaches leverage relationships among labels to improve generalization, accuracy, and interpretability. Early examples in the audio domain that included hierarchical information were carried out using models such as HMMs and GMMs [19], [20]. More recently, hierarchical information has been integrated in classification problems through hierarchical deep neural architectures, loss functions, and evaluation metrics. Many different hierarchy-aware loss functions have been used, adapting standard losses such as cross entropy, triplet, and contrastive [21]–[26], proposing custom losses such as rank-based loss [23], and combinations thereof [27]. Most of these approaches are paired with representation learning and deep embeddings. These methods are particularly valuable when addressing challenges such as class imbalance, semantic overlap, or sparse labels at deeper levels of the taxonomy.

Multimodal setups in representation learning are gaining significant traction in the audio domain. It is common practice to obtain and utilize accompanying metadata, such as textual descriptions, tags, or contextual information, that can provide additional semantic information in the acoustic signal. Recent advances in multimodal learning have led to methods that learn shared embedding spaces for audio and text, supporting seamless integration and cross-modal understanding of sound data. Models such as Contrastive Language-Audio Pretraining (CLAP) are context-agnostic and aim to learn general-purpose representations that are transferable across diverse audio-text tasks [28], [29]. Other approaches incorporate semantic supervision, training models to classify audio events while aligning them with text-derived embeddings, thereby encouraging semantic consistency in the learned representations [30], [31]. In some methods, audio and text embeddings are explicitly combined or fused as joint input for downstream classification tasks [32], [33], allowing the model to leverage complementary information from both modalities. In this paper, we incorporate hierarchy information using a custom hierarchy-aware loss function, and we leverage the CLAP audio-text space to include audio and auxiliary textual information in the classification pipeline.

3. METHODOLOGY

3.1. Model Formulation

We design a classifier to process both audio and text embeddings, denoted respectively as $\mathbf{a} \in \mathbb{R}^{d_a}$ and $\mathbf{t} \in \mathbb{R}^{d_t}$. Each input passes through a modality-specific encoder, implemented as a multilayer perceptron (MLP) comprising an input projection, a sequence of residual blocks, and an output projection. All layers use Leaky ReLU activations. The encoders extract modality-specific feature vectors:

$$\mathbf{h}_a = f\text{enc}^a(\mathbf{a}), \quad \mathbf{h}_t = f\text{enc}^t(\mathbf{t}) \tag{1}$$

When both modalities are used jointly, their representations are fused via an attention-based mechanism:

$$\mathbf{h}_f = \alpha_1 \mathbf{h}_a + \alpha_2 \mathbf{h}_t, \quad \boldsymbol{\alpha} = \text{Softmax}\left(W_2 \tanh(W_1[\mathbf{h}_a; \mathbf{h}_t])\right) \tag{2}$$

The resulting feature vector $\mathbf{h} \in \mathbf{h}_a, \mathbf{h}_t, \mathbf{h}_f$ is then passed through a latent projection and classification layer, producing the final logits:

$$\hat{\mathbf{y}} = f\text{cls}\left(f\text{proj}(\mathbf{h})\right) \tag{3}$$

The training objective is to minimize the cross-entropy loss between the predicted logits $\hat{\mathbf{y}}$ and the ground truth labels $\mathbf{y}$:

$$\mathcal{L} = -\sum_i y_i \log\left(\text{softmax}(\hat{\mathbf{y}})_i\right) \tag{4}$$

where i indexes the second-level classes.

3.2. Hierarchical Setting

In the hierarchical setting, we train the model adding two auxiliary losses to the standard cross-entropy in Eq. 4: a top-class penalty $\mathcal{L}_{\text{Top}}$ and a contrastive loss $\mathcal{L}_{\text{Contr}}$.

Let $\{\mathbf{z}_i\}_{i=1}^N$, with $\mathbf{z}_i \in \mathbb{R}^d$ and $\|\mathbf{z}_i\| = 1$, denote the normalized latent representations of a batch of N samples. Each sample i has a second-level label y_i and a corresponding top-class label $t_i = \text{top}(y_i)$. The top-class penalty is:

$$\mathcal{L}_{\text{Top}} = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\big(\text{top}(\hat{y}_i) \neq t_i\big), \tag{5}$$

where $\hat{y}_i$ is the predicted second-level class, and $\mathbb{1}(\cdot)$ is the indicator function, equal to 1 if the condition holds and 0 otherwise.

The use of contrastive loss [34] encourages samples with the same top-class label to be pushed toward the same position in the feature space:

$$\ell_i = -\frac{1}{|\mathcal{P}(i)|}\sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p/\tau)}{\sum_{a \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}, \tag{6}$$

where $\mathcal{P}(i) = \{p \neq i \mid t_p = t_i\}$ is the set of positive samples sharing i's top-class label, $\mathbf{z}_p$ are the corresponding representations, $\mathbf{z}_a$ are all other representations in the batch, and $\tau > 0$ is a temperature parameter. The total contrastive loss, weighted by a factor λ, is:

$$\mathcal{L}_{\text{Contr}} = \lambda \frac{1}{N}\sum_{i=1}^N \ell_i. \tag{7}$$

4. EXPERIMENTAL SETUP

4.1. Dataset and Preprocessing

For our experiments, we use BSD10k-v1.1, an updated version of the BSD10k dataset introduced in [6]. BSD10k-v1.1 comprises 10,956 sounds retrieved from Freesound, labeled into 5 top-level classes and 23 second-level classes, according to the hierarchical taxonomy presented in [8]. This version contains a more balanced dataset with increased representation of underrepresented classes. It was also refined by analyzing the training data to identify and correct misclassifications caused by human error. The dataset includes confidence scores assigned to each sound during the annotation process, ranging from 1 (very unconfident) to 5 (very confident). The distribution of annotation confidence levels is as follows: 1 (1%), 2 (6.9%), 3 (30.1%), 4 (55%) and 5 (7%). The updated version of the dataset is publicly accessible[1].

As part of the experimental setup for model training, we run part of the experiments with the sounds assigned a confidence score $\geq$3 and $\geq$4, which consist 92.1% and 62% of the dataset, respectively. As input features, we use representations from text and audio embeddings. In previous work, it was demonstrated that CLAP audio embeddings achieve better performance compared to purely audio-based embeddings in a heterogeneous setting [6]. In this work we extract CLAP embeddings as our audio feature representation. Additionally, we introduce a new modality by extracting text embeddings from the same CLAP model, using the metadata (textual data) in BSD10k-v1.1 originally provided by Freesound users. Text embeddings are extracted in two configurations: i) using sound title and tags; ii) using sound title, tags, and textual descriptions. Because the description is considered noisier, as it often includes unstructured or irrelevant information, we aim to evaluate its impact compared to the more concise metadata. All audio and text embeddings are extracted as 512-dimensional vectors.

4.2. Hyperparameters and Training

We trained each model for a maximum of 100 epochs, using early stopping in validation to prevent overfitting with $patience = 5$, Adam optimizer, fixed learning rate scheduler with an initial learning rate of 1e$-$3, and a batch size of 64. Due to the unbalanced class distribution in the dataset, we opted for a stratified 5-fold cross-validation [35], ensuring that each fold preserved the overall class proportions. For each fold, we split the data into 80%/20% train/test, further dividing the former into 90%/10% train/validation. In all experiments, the feature vectors h_a and h_t were set to a dimensionality of 128, while the latent representation vectors z had a fixed dimensionality of 64. In the hierarchical settings, we kept $\tau = 0.5$ and $\lambda = 1$ for the contrastive loss. Finally, we applied data augmentation on both audio and text embeddings, namely noise addition and random masking (up to 70%). In the context of pretrained embeddings, such augmentations act as a network regularizer similar to dropout [36]. Preliminary results indicate that these augmentations do not impact overall model performance but do lead to faster convergence, thereby reducing training time.

[1] https://github.com/allholy/BSD10k

Table 1: Performance comparison across different hierarchy and modality configurations at three annotation confidence thresholds, including top-2 accuracy. All values are reported as percentages.

Confidence	Loss	Modality	Second-level	Top-level	Second-level (top-2)	Top-level (top-2)
≥ 1	Hierarchical	Audio	76.73 ± 0.47	87.67 ± 0.35	88.76 ± 0.20	94.44 ± 0.33
		Text	77.26 ± 0.32	87.11 ± 0.27	89.22 ± 0.10	94.14 ± 0.10
		Both	79.33 ± 0.03	88.75 ± 0.30	90.76 ± 0.13	95.19 ± 0.14
	Non-hierarchical	Audio	76.94 ± 0.23	87.50 ± 0.10	89.40 ± 0.58	95.29 ± 0.21
		Text	76.69 ± 0.26	86.54 ± 0.36	88.92 ± 0.27	94.55 ± 0.25
		Both	**79.80 ± 0.53**	**88.97 ± 0.11**	**91.40 ± 0.40**	**96.02 ± 0.13**
≥ 3	Hierarchical	Audio	78.83 ± 0.65	88.81 ± 0.29	89.50 ± 0.20	94.57 ± 0.31
		Text	78.16 ± 0.99	87.96 ± 0.54	89.86 ± 0.22	94.70 ± 0.26
		Both	81.07 ± 0.81	89.81 ± 0.52	91.64 ± 0.05	95.69 ± 0.19
	Non-hierarchical	Audio	78.77 ± 0.31	88.98 ± 0.12	90.05 ± 0.06	95.64 ± 0.18
		Text	78.50 ± 0.49	87.93 ± 0.18	90.49 ± 0.01	95.16 ± 0.10
		Both	**81.50 ± 0.91**	**90.00 ± 0.25**	**92.12 ± 0.13**	**96.23 ± 0.03**
≥ 4	Hierarchical	Audio	84.79 ± 1.32	92.57 ± 0.28	92.18 ± 1.00	95.87 ± 0.50
		Text	83.39 ± 1.59	91.48 ± 0.43	92.17 ± 0.56	95.40 ± 0.08
		Both	87.34 ± 0.83	**93.60 ± 0.34**	94.11 ± 0.42	96.58 ± 0.17
	Non-hierarchical	Audio	84.96 ± 0.42	92.23 ± 0.16	92.83 ± 0.81	96.74 ± 0.38
		Text	84.12 ± 0.77	91.61 ± 0.42	92.92 ± 0.46	96.49 ± 0.19
		Both	**87.36 ± 0.42**	93.42 ± 0.22	**94.76 ± 0.32**	**97.66 ± 0.13**

5. RESULTS

We compare against a baseline consisting of a KNN trained in audio-only embeddings [6], which we retrain with the latest version of the dataset and yield an accuracy of 77.45% in second-level classes and 87.65% in top-level classes. Our best performing model with the same setting (non-hierarchical, audio-only, non-filtered dataset by confidence levels) performs 77.51% on the second level and 87.84% on the top level. This demonstrates an incremental improvement.

Table 1 shows the results of all training configurations. For the text embeddings, we use the best-performing variant, which includes descriptions. The difference in accuracy is marginal, suggesting that a reduced amount of textual information may still be sufficient in this setting. Overall, both hierarchical and non-hierarchical models tend to exhibit comparable performance. Models trained with audio embedding achieve a small increase in accuracy than those using text embeddings, while both modalities significantly outperform the single-modality ones. Annotator confidence plays an important role in the overall accuracies: filtering the dataset by confidence level (including sounds with confidence scores of ≥ 3 and ≥ 4) removed the more ambiguous samples, resulting in a substantially improved accuracy across all configurations (up to ~2% and ~9% in the second level and ~1.5% and ~5% in the top level, respectively). The approach presented herein allows for a more thorough assessment of the model's learning performance. We have observed that human uncertainty is reflected in the model's behavior; specifically, the model occasionally makes errors similar to those of annotators due to the ambiguous boundaries inherent in the broadness of the task. The average confidence per class in the dataset ranges from 3.15 to 4.05, indicating moderate variations of annotation uncertainty in specific classes but suggesting the presence of ambiguous sounds in most classes.

In addition, we report the top-2 accuracy for both taxonomic levels, defined as the proportion of samples for which the correct label appears among the model's top two predictions. We support that this is particularly relevant, since the case of edge-case sounds falling between two categories could be semantically plausible, as already highlighted in [6]. Therefore, we can assess its ability to capture these ambiguities and provide insight into how often it nearly arrives at the correct decision, even when its top prediction is incorrect. The accuracy across all configurations trained with the unfiltered dataset improves by ~12% in the top level and ~7% in the second level. During the experiments, this proved also useful from an analytical perspective, as it allowed us to evaluate clearly wrong-classified samples. In further comment on such ambiguity, Fig. 1 shows an example of a confusion matrix relative to a single fold in hierarchical, multimodal training. Overall, some classes always exhibit a higher degree of confusion. Misclassifications are prominent within the *Music* (*m*) top class, particularly with the *Multiple instruments* (*m-mi*) subclass frequently being assigned to the other two subclasses, while within the *Soundscapes* (*ss*) class, the *Indoors* (*ss-i*) class gets incorporated into *Urban* (*ss-u*). In cases where the top-level class is also misclassified, certain second-level classes remain persistently challenging even as the annotation confidence score filtering increases. Specifically, misclassifications occur from *Speech→Conversation/Crowd* (*sp-c*) to *Soundscapes→Urban*, and from *Sound effects→Animals* (*fx-a*) to *Soundscapes→Nature* (*ss-n*). This could be due to certain classes having more complex definitions or involving the concept of single/multiple sound sources, which the model may not effectively capture.

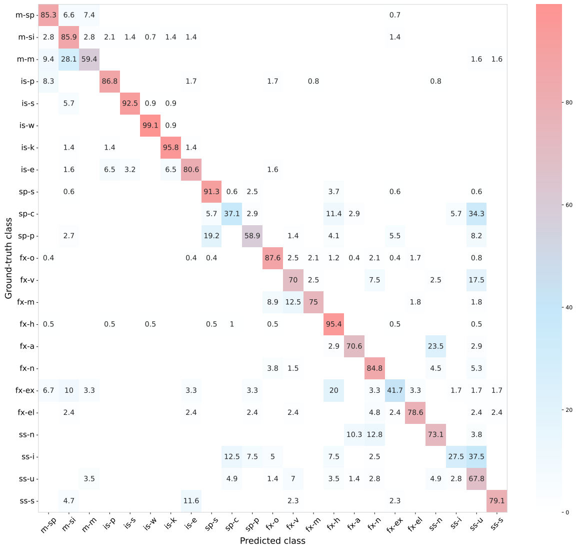
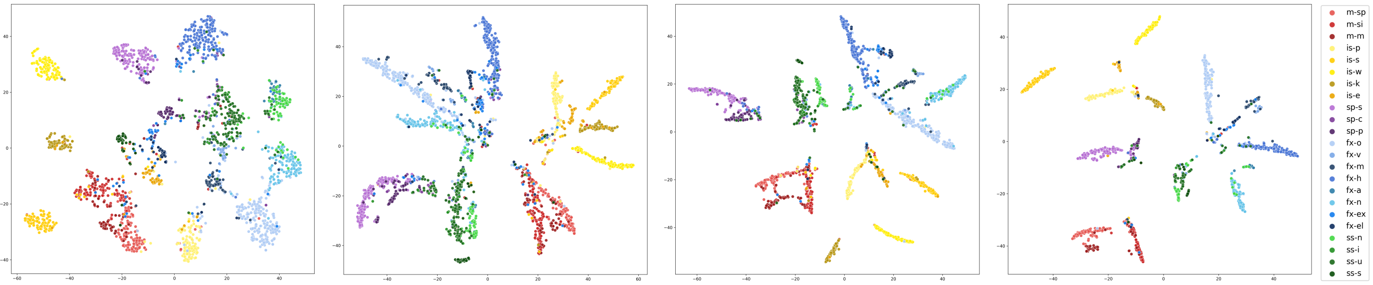


Fig. 1: Confusion matrix showing class-wise accuracies for the hierarchical configuration (H, A+T, C $\geq$ 1, training fold 1).



(a) NH, A+T, C $\geq$ 1 (b) H, A+T, C $\geq$ 1 (c) H, A+T, C $\geq$ 3 (d) H, A+T, C $\geq$ 4

Fig. 2: Latent space visualizations of the various settings. Plots are relative to the same training fold. NH = non-hierarchical; H = hierarchical; A+T = audio + text modality; C = annotation confidence score.

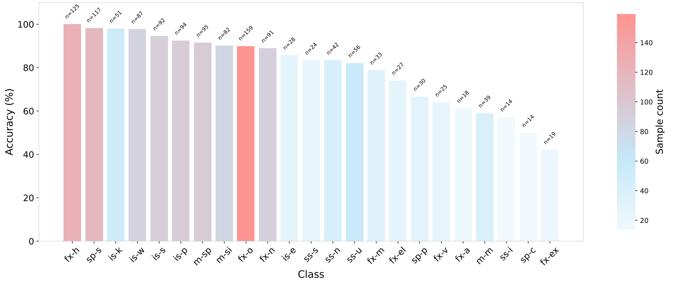


Fig. 3: Sorted average class accuracy color-coded by number of samples in the test set (NH, A+T, C $\geq$ 3, test fold 1).

Although the classification accuracy across the same modalities remains comparable between hierarchical and non-hierarchical approaches, the structure of the latent space differs significantly. We apply PCA to reduce the 64-dimensional representations to a 2D latent space. As seen in Fig. 2, the hierarchical latent space (2b) demonstrates a more coherent and well-separated representation, where classes sharing the same top-level class cluster closer together. This spatial organization reflects the underlying taxonomy of the data and highlights the positive impact of hierarchical priors on representation learning. Another artifact is the elongation of latent clusters, which may be attributed to the dimensionality reduction algorithm. Further examination can be conducted to identify which samples are located on opposite sides of the intra-class space. This behavior contrasts with the non-hierarchical latent space illustrated in Fig. 2a, where such clear organization is absent. For example, *Instrument samples →Percussion* (*is-p*) are positioned closer to *Sound effects* (*fx*), likely due to shared acoustic characteristics such as short, non-pitched textures. Another example is that *Soundscapes→Synthetic* (*ss-s*) are located close to *Sound effects→Electronic* (*fx-el*) and *Instrument samples→Electronic* (*is-e*), likely because they often include computer-generated textures and may share common tag vocabularies. This could indicate that the model first focuses on timbral characteristics (e.g. texture) rather than on properties such as the number of sources or time-domain characteristics. Interestingly, the conceptually important distinction between mono-source and multi-source content, such as that between *Sound effects* and *Soundscapes*, does not appear to be prioritized.

Figs. 2c and 2d, which depict the latent spaces for multimodal hierarchical training with annotation confidence scores $\geq$3 and $\geq$4, respectively, show a similar coherent structure where classes that share the same top level are closer. Additionally, as confidence increases, the clusters become sharper with larger gaps between second-level classes inside the same top-level class, fewer intrusions from other classes, and a greater tendency for each class to occupy a well-defined region of the latent space.

Finally, we observe a common tendency for less-represented classes to exhibit lower performance, as shown in Fig. 3. Although class imbalance is a well-known factor contributing to such disparities, we argue that this alone does not fully explain such behavior. Indeed, the taxonomy used underlies semantically or structurally complex phenomena that are inherently harder to learn, stemming from intra-class variability or a dependence on context-specific features. Consequently, the limited number of training examples may not fully represent this variability and lead to poor generalization on unseen data in the test split. Moreover, preliminary experiments involving traditional approaches to addressing dataset imbalance (e.g. focal loss [37]) yield only negligible improvements. Some underrepresented classes performed slightly better, but the improvements were not consistent across different folds. Enhancing performance on such classes may, therefore, require not only rebalancing the dataset but also increasing the diversity of training samples. Notably, the confidence of the model for each class did not show a correlation with the number of samples in the training set, and training with fewer data in all classes did not significantly impact the accuracy.

6. CONCLUSION AND FUTURE WORK

This paper investigates multimodal and hierarchical strategies to enhance classification performance in real-world tasks. The insertion of textual information demonstrated an improvement in classification accuracy. We also foresee that the shown benefits of hierarchical latent representations extend beyond representational structure, proving especially useful in organizing similarity relationships with a better balance between semantic and acoustic features while encoding high-level conceptual information. Filtering data based on annotation confidence scores proved helpful in understanding what the model learns and identified sources of model confusion. To improve generalization, incorporating this information into the training process, such as weighting samples by their confidence, could be beneficial.

Future work includes leveraging more data for the training and evaluation of heterogeneous classification tasks. Since the introduction of the BST in Freesound, contributors have been manually annotating their newly uploaded sounds with BST classes, thereby providing data for further experiments. Even though crowdsourced annotations may be inconsistent, since edge cases can be interpreted differently by each user, such models can still help identify potentially incorrect labels. We also plan to explore the effect of inserting hierarchical information directly into the embedding learning process.

7. ACKNOLEDGMENT

This research was partially funded by the Secretaria d'Universitats i Recerca del Departament de Recerca i Universitats de la Generalitat de Catalunya (2024FI-100252) under the Joan Oró program, the IA y Música: Cátedra en Inteligencia Artificial y Música (TSI-100929-2023-1) by the Secretaría de Estado de Digitalización e Inteligencia Artificial and the European Union-Next Generation EU under the program Cátedras ENIA 2022, and the IMPA project (PID2023-152250OB-I00) by the Ministry of Science, Innovation and Universities of the Spanish Government, the Agencia Estatal de Investigación (AEI) and co-financed by the European Union.

REFERENCES

[1] N. Narkhede, S. Mathur, A. Bhaskar, and M. Kalla, "Music genre classification and recognition using convolutional neural network," *Multimedia Tools and Applications*, vol. 84, no. 4, pp. 1845–1860, 2025.

[2] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised automatic speech recognition: A review," *Speech Communication*, 2022.

[3] H. Xu, Y. Tian, H. Ren, and X. Liu, "A lightweight channel and time attention enhanced 1D CNN model for environmental sound classification," *Expert Systems with Applications*, vol. 249, 2024.

[4] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[5] J. Zhang, S. Jayasuriya, and V. Berisha, "Learning repeatable speech embeddings using an intra-class correlation regularizer," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 425–76 443, 2023.

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

[7] M. Goulão, L. Bandeira, B. Martins, and A. L. Oliveira, "Training environmental sound classification models for real-world deployment in edge devices," *Discover Applied Sciences*, vol. 6, no. 4, 2024.

[8] P. Anastasopoulou, X. Serra, and F. Font, "A General-Purpose Sound Taxonomy for the Classification of Heterogeneous Sound Collections," *Research Square*, 2025.

[9] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT press, 1994.

[10] S. Kumar, K. E. Stephan, J. D. Warren, K. J. Friston, and T. D. Griffiths, "Hierarchical processing of auditory objects in humans," *PLoS computational biology*, vol. 3, no. 6, 2007.

[11] O. Bones, T. J. Cox, and W. J. Davies, "Sound categories: Category formation and evidence-based taxonomies," *Frontiers in Psychology*, vol. 9, 2018.

[12] A. Jimenez, B. Elizalde, and B. Raj, "Sound event classification using ontology-based neural networks," in *Proc. Annual Conference on Neural Information Processing Systems*, vol. 9, 2018.

[13] S. Zhang, R. Xu, C. Xiong, and C. Ramaiah, "Use all the labels: A hierarchical multi-label contrastive learning framework," in *Proc. of the Conference on Computer Vision and Pattern Recognition (CVF)*, 2022.

[14] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. 21st Int. Conference on Multimedia (ACM)*, 2013, pp. 411–412.

[15] E. Costa, A. Lorena, A. Carvalho, and A. Freitas, "A review of performance evaluation measures for hierarchical classifiers," in *Proc. AAAI Workshop*, 2007, pp. 1–6.

[16] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data mining and knowledge discovery*, vol. 22, pp. 31–72, 2011.

[17] D. Moffat, D. Ronan, and J. D. Reiss, "Unsupervised taxonomy of sound effects," in *Proc. 20th Int. Conference on Digital Audio Effects (DAFx)*, 2017.

[18] E. Tieppo, R. R. dos Santos, J. P. Barddal, and J. C. Nievola, "Hierarchical classification of data streams: A systematic literature review," *Artificial Intelligence Review*, pp. 1–40, 2022.

[19] T. Zhang and C.-C. J. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6. IEEE, 1999, pp. 3001–3004.

[20] J. J. Burred and A. Lerch, "A hierarchical approach to automatic musical genre classification," in *Proc. Int. Conference on Digital Audio Effects (DAFx)*, 2003, pp. 8–11.

[21] J. Liang, H. Phan, and E. Benetos, "Learning from taxonomy: Multi-label few-shot classification for everyday sound recognition," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 771–775.

[22] L. Sun, Z. Lian, B. Liu, and J. Tao, "HiCMAE: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition," *Information Fusion*, vol. 108, 2024.

[23] I. Nolasco and D. Stowell, "Rank-based loss for learning hierarchical representations," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3623–3627.

[24] L. Pham, I. McLoughlin, H. Phan, R. Palaniappan, and A. Mertins, "Deep feature embedding and hierarchical classification for audio scene classification," in *Proc. Int. Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[25] P. Hase, C. Chen, O. Li, and C. Rudin, "Interpretable image recognition with hierarchical prototypes," in *Proc. AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, 2019, pp. 32–40.

[26] A. Jati, N. Kumar, R. Chen, and P. Georgiou, "Hierarchy-aware loss function on a tree structured label space for audio event detection," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6–10.

[27] H. Tian, S. Lattner, B. McFee, and C. Saitis, "Hybrid losses for hierarchical embedding learning," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[28] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[29] 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 *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[30] M. Esposito, G. Valente, Y. Plasencia-Calaña, M. Dumontier, B. L. Giordano, and E. Formisano, "Semantically-informed deep neural networks for sound recognition," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[31] Y. Xiao, Y. Ma, S. Li, H. Zhou, R. Liao, and X. Li, "Semanticac: Semantics-assisted framework for audio classification," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[32] Y. Zhang, M. Wu, and X. Cai, "A dynamic cross-modal learning framework for joint text-to-audio grounding and acoustic scene classification in smart city environments," *Digital Signal Processing*, 2025.

[33] B. Elizalde, S. Zarar, and B. Raj, "Cross modal audio search and retrieval with joint embeddings based on text and audio," in *Proc. Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4095–4099.

[34] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Proc. Int. Conf. Advances in Neural Information Processing Systems*, vol. 33, 2020.

[35] J. R. Almonteros and J. B. Matias, "Integration of stratified kfold cross validation to enhance prediction accuracy: A comparison study," in *Proc. Int. Conference on Data Analytics for Business and Industry (ICDABI)*. IEEE, 2024, pp. 81–85.

[36] Z. Liu, Y. Chen, J. Li, M. Luo, P. S. Yu, and C. Xiong, "Improving contrastive learning with model augmentation," *arXiv preprint arXiv:2203.15508*, 2022.

[37] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating deep neural networks using focal loss," *Advances in neural information processing systems*, vol. 33, pp. 15 288–15 299, 2020.