



Beyond Conventional Metrics: using Entropic Triangles to Explain Balancing Methods in Acoustic Scene Classification

Claudia Montero-Ramírez¹, Alba Martínez-Serrano¹, Jorge Garcelán-Gómez¹, Francisco J. Valverde-Albacete², Carmen Peláez-Moreno^{1,3}

¹Signal Theory and Communications Dept., Universidad Carlos III de Madrid, Spain

²DTSCSTC, Universidad Rey Juan Carlos, Spain

³Institute of Gender Studies, Universidad Carlos III de Madrid, Spain

clmonter@pa.uc3m.es, amerran@pa.uc3m.es, jgarcela@pa.uc3m.es,
francisco.valverde@urjc.es, cpelaez@ing.uc3m.es

Abstract

Understanding soundscapes is essential for making sense of real-world scenarios in complex environments. However, real-life conditions present significant challenges for AI-based methods, particularly due to the highly imbalanced nature of Audio Tagging problems. In this work, we investigate the impact of data imbalance on the training dynamics of state-of-the-art models for Acoustic Scene Classification. Using the DCASE TAU Urban Acoustic Scenes 2022 dataset and the CP-Mobile model, we introduce controlled imbalance scenarios and analyze their effect through the Entropic Triangle framework. Our findings reveal that the training dynamics are strongly influenced by the chosen balancing approach. It also suggests longer training periods with conventional optimizer for re-balanced classification metrics. In this way, this study provides new insights into the role of entropy-based analysis in developing robust Acoustic Scene Classification systems for real-world applications.

Index Terms: Audio Tagging, Acoustic Scene Classification, Data Imbalance, Entropic Triangle

1. Introduction

Any Audio Tagging (AT) problem applied to real-life conditions exhibits a highly imbalanced nature. An example of this is AudioSet [1], one of the most well-known databases in the field of Sound Event Detection (SED) and AT. It comprises 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos. Its majority sound labels, *Speech* and *Music*, present over 1 million samples, while other many sounds have less than 200. On the other hand, the DCASE community¹ launches every year several challenges related to Acoustic Scene Classification (ASC), SED and AT, among other topics. In relation to SED, the DESED dataset [2] has been used since DCASE 2020 Task 4. It contains domestic sounds and is also highly unbalanced. Similarly, WildDESED—an extension of the original DESED dataset—was created in 2024 to reflect a wider variety of domestic scenarios by incorporating complex and unpredictable background noises [3].

Regarding the ASC field, the dataset selected for the DCASE ASC task since 2022, is the TAU Urban Acoustic Scenes 2022 Mobile Development Data Set [4], which is almost completely balanced. This is far from reality; especially when we face Artificial Situational Awareness problems, where people's daily lives are tracked acoustically in order to provide security for those who need it.

In 2020, the research group UC3M4Safety² developed a database called *WELIVE*, designed to track the daily lives of women at risk of suffering gender-based violence [5, 6, 7]. This database collected both physiological variables and audio recordings of their routine environment. The monitoring carried out supported the widely accepted hypothesis that humans experience acoustic scenes in a highly imbalanced manner throughout their daily lives, resulting in a naturally imbalanced dataset, underscoring the importance of understanding its consequences.

To ensure public availability of the data and a sufficient amount of it for our experiments, however, the TAU Urban Acoustic Scenes 2022 Mobile [4] dataset has been used instead of *WELIVE* in the experiments presented in this work.

The goal of this paper is to use information-theoretic tools to explain balancing methods in ASC, as class imbalance is a challenge encountered when working with real-world conditions in AT problems. To address this, we use different class balancing methods and analyze the training process of ASC. Specifically, we focus on the use of inverse prior weighting in the loss function.

The contributions of this work are twofold: First, we present a novel application of Entropic Triangles (ET) to visualize and analyze the training process when applying various class balancing methods in ASC, providing unique insights into the dynamics of class re-balancing techniques. Second, we have seen, for the first time in the ET, that the entropy of predicted classes can exceed that of true classes at certain points during training when using a specific class re-balancing function, offering new perspectives on the behavior of these methods. Through this comprehensive analysis, we improve the understanding of class imbalance in ASC and pave the way for more effective strategies in handling imbalanced datasets in this domain.

2. Related work

Several methods have been proposed to address class imbalance, which can be broadly categorized into data-level, algorithm-level and hybrid approaches [8, 9, 10]. *Data-level methods* modify the training data to balance class distributions, including oversampling minority classes, undersampling majority classes, and synthetic data generation techniques such as SMOTE. On the other hand, *algorithm-level techniques* adjust the learning process, often by modifying loss functions, incorporating class-weighting mechanisms, or using cost-sensitive learning to assign different importance to each class. Finally, *hybrid methods* integrate strategies from both approaches to cre-

¹<https://dcase.community/>

²<https://www.uc3m.es/institute-gender-studies/>

ate more robust solutions for imbalanced scenarios.

A key aspect of evaluating ASC models is the choice of metrics. Previous studies have extensively discussed common evaluation metrics in this domain [8, 9]. The most widely used include macro-F1, weighted-F1, Standard (or unweighted) Accuracy, and Weighted Accuracy. For the ASC task, the DCASE community uses Accuracy as evaluation metric [11]. Additionally, precision, recall and ROC are also commonly used in other multiclass imbalanced problems [12]. Finally, the DCASE 2024 challenge of SED used a modified AUC for the evaluation [13].

Although these metrics provide useful insights into model performance, they may not fully reflect the complexities introduced by data imbalance, reinforcing the need for alternative approaches [14].

3. Methods

3.1. Class Balancing Methods

One of the best known balancing methods is to apply class weights to the loss function during training [8], [9],

$$w_i = \frac{1}{p_i \cdot C} \quad (1)$$

where p_i represents the prior probability of each class i and C is the number of classes. We compare it with non-weighting, that is,

$$w_i = 1 \quad \forall i \quad (2)$$

3.2. Entropy Balance Equations and Triangles

The Channel Bivariate Entropy Balance Equation and Triangle is an information-theoretic framework for assessing the robustness of a model in multiclass problems [14, 15].

Given the actual label distribution $X \sim P_X$ and the predicted one $Y \sim P_Y$ ³ as marginalized from the empirical distribution P_{XY} obtained from a confusion matrix $N_{X,Y}$, the following *entropy balances*—unnormalized and normalized by $H_{U_X \cdot U_Y}$ —hold:

$$\begin{aligned} H_{U_X \cdot U_Y} &= \Delta H_{P_X \cdot P_Y} + 2 \cdot MI_{P_{XY}} + VI_{P_{XY}} \\ 1 &= \Delta H'_{P_X \cdot P_Y} + 2 \cdot MI'_{P_{XY}} + VI'_{P_{XY}} \end{aligned} \quad (3)$$

where U_X and U_Y are the uniform distributions of label and predicted labels, respectively, and we also have:

- The *mutual information* $MI_{P_{XY}}$ [16, 17].

$$MI_{P_{XY}} = H_{P_X} + H_{P_Y} - H_{P_{XY}} \quad (4)$$

- The *redundancy or divergence with respect to uniformity* $\Delta H_{P_X \cdot P_Y}$ [18].

$$\Delta H_{P_X \cdot P_Y} = (H_{U_X} - H_{P_X}) + (H_{U_Y} - H_{P_Y}) \quad (5)$$

- The *variation of information* $VI_{P_{XY}}$ [19].

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} \quad (6)$$

The 2-simplex described by the Entropy Balance Equation (3) when applied to evaluating classifiers can be represented by an aggregate ET as depicted in Figure 1 where dots describe specific types of classifiers. The call-outs situated in the center of the sides of the triangle apply to the whole side refer to

³It is more usual to have $Y \sim P_Y$ as the real labels, and $\hat{Y} \sim P_{\hat{Y}}$ as the predicted labels, but we conform here to the ETs notation.

the confusion matrices of the classifiers situated in them. E.g., for Figure 1, the perfect classifier (blue) would be the one that achieves a perfect diagonal in the confusion matrix, the useless classifier (brown) struggles with balanced data, and the worst classifier (red) deals with easy, unbalanced data, achieving high accuracy without learning, leading to the accuracy paradox as they become majority classifiers [14].

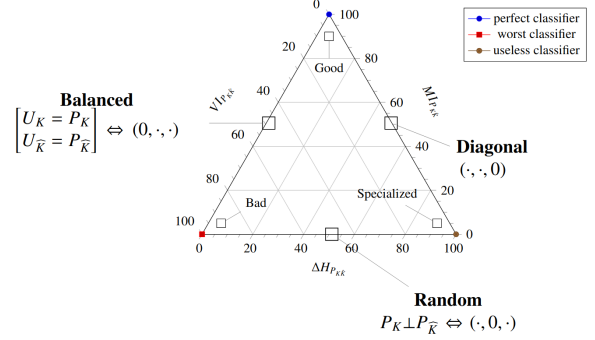


Figure 1: Schematic ET as applied to supervised classifier assessment, from [14].

An ideal classifier should maximize $MI_{P_{XY}}$ —therefore minimizing $VI_{P_{XY}}$ —while not making use of the imbalanced captured in $\Delta H_{P_X \cdot P_Y}$.

$$MI'_{P_{XY}} \rightarrow 1 \quad (7)$$

$$[H'_{P_{X|Y}}, \Delta H'_{P_X}] \approx [H'_{P_{Y|X}}, \Delta H'_{P_Y}] \quad (8)$$

where (7) maximizes the mutual information between the real and predicted labels, and (8) asserts that the actual and predicted labels have the same probability distribution. But since $\Delta H'_{P_X}$ does not change with training, but the other quantities in (8) do, it is very convenient to represent (3) side by side the *split balance equations* for real and predicted labels where the quantities in (8) appear explicitly:

$$1 = \Delta H'_{P_X} + MI'_{P_{XY}} + H'_{P_{X|Y}} \quad (9)$$

$$1 = \Delta H'_{P_Y} + MI'_{P_{XY}} + H'_{P_{Y|X}} \quad (10)$$

This can be represented in the ET just by renaming the axes appropriately, as seen in Figure 2.

4. Experimental Settings

4.1. Dataset and Model

The dataset used in this paper is the TAU Urban Acoustic Scenes 2022 Mobile [4], as it is the one used for the ASC task in DCASE. The model is the CP-Mobile [20], the baseline for the DCASE Task 1 Challenge 2024⁴.

4.2. Data Preprocessing

By default, the label distribution corresponding to the dataset used for this paper is almost uniform. Then, to illustrate the effects of real-life label distributions, the training set is linearly imbalanced and reduced. A 5-fold stratified partition is performed, so that the validation and training sets have the same

⁴https://github.com/fschmid56/cpjku_dcaset3

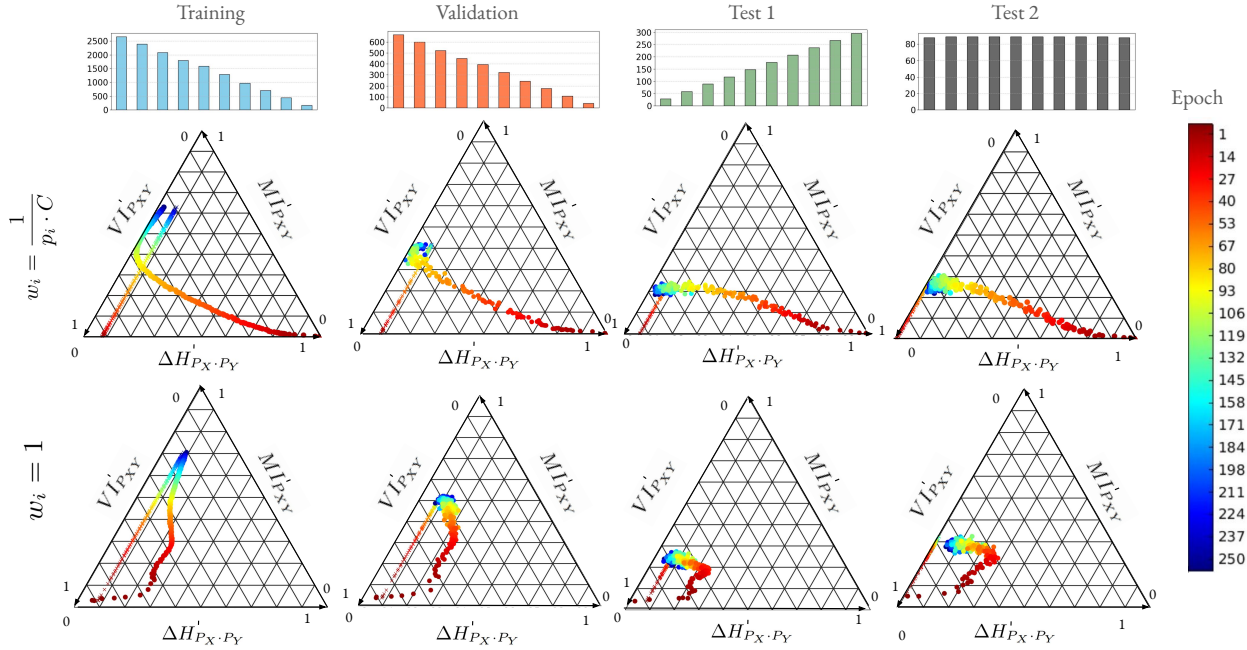


Figure 2: *ET* for the training set and validation for fold 1 and two independent tests. The top *ET*s represent the weighting of the cost function and the lower ones, present no weighting. Each epoch is illustrated with two elements aligned horizontally: the actual labels $X \sim P(X)$ are represented with \times and are given by $[\Delta H'_{P_X}, MI'_{P_{X|Y}}, H'_{P_X|Y}]$ coordinates, and predicted ones $Y \sim P(Y)$ are represented with \bullet and correspond to $[\Delta H'_{P_Y}, MI'_{P_{Y|X}}, H'_{P_Y|X}]$. The class distributions for the training, validation, test 1 and test 2 sets are shown at the top of the figure.

class distributions⁵. Any partition, data reduction and class-imbalancing is performed by preserving the original distribution of devices and cities. The training and validation sets are composed of 17, 616 samples, resulting in $3, 523.2 \pm 0.45$ samples per fold for validation, and $14, 092.8 \pm 0.40$ for training, expressed as $\mu \pm \sigma$.

In addition, two test sets with different distributions are evaluated to mimic real-world situations when the distributions are not known a priori:

- **Test 1** has the inverse distribution with respect to the training and validation sets, i.e. the sorting of the classes by occurrence gives exactly the inverse ordering.
- **Test 2** has almost an uniform distribution.

The upper row of Figure 2 shows an example of class distribution for the first fold. Note that there are 5 different training and validation sets since we used 5-fold cross-validation, while the two test sets are distinctive and unrelated.

4.3. Training

CP-Mobile is trained end-to-end during 250 epochs, using the Adam optimizer with a learning rate of $5 \cdot 10^{-3}$ and a weight-decay of $1 \cdot 10^{-4}$. As mentioned above, stratified 5 fold cross-validation is performed and, additionally, 2 different test sets are evaluated in each epoch and fold. Then, two different cost-weighting methods are applied in the training process for each fold, as described in Section 3.1:

- **Method 1:** The class-weighting parameter given by (1) is

⁵Find all partitions in this link: <https://github.com/clmonter/Interspeech-2025-Entropy-Triangle>

introduced in the cross-entropy loss. After the softmax activation the inverse a priori label distribution taken from the training set is multiplied to make the decision.

- **Method 2:** No class-weighting is applied during the training, i.e. as in (2).

5. Results and Discussion

5.1. Metrics

Figure 3 shows the evolution of the results in terms of F1-macro, F1-weighted, UA and WA during the training process for each of the epochs. Regarding these conventional metrics, it can be seen that usually, the non-weighting method performs better in the validation set. This is because, by not applying any weighting on the loss function, the system implicitly learns the distribution of the training set, which makes it perform better when tested on a set that has the same distribution. On the other hand, we find the opposite case with respect to Test 1. These data have a distribution that is the inverse of that of the training and validation sets, so weighting Method 1, which forces the system to weight the loss corresponding to each of the classes according to their inverse a priori probability, exhibits a better behavior. We hypothesize that the model trained this way is more robust to the variations of the data distributions. The results with the uniform distribution of Test 2 show no significant differences between both weighting methods.

5.2. Entropy Balance Equations

Recall that $N_{X,Y}$ is The confusion matrix obtained from classification, with $X \sim P(X)$ and $Y \sim P(Y)$ representing the

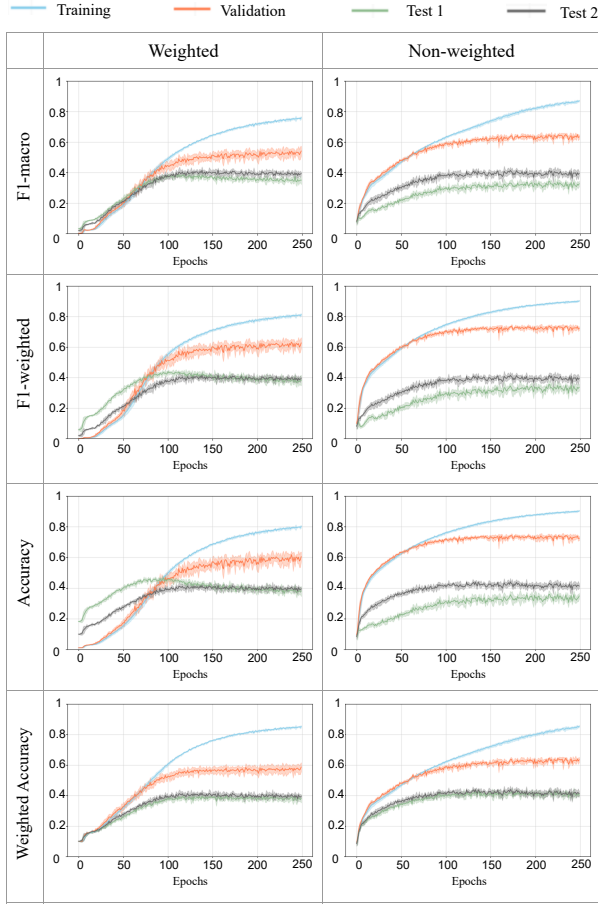


Figure 3: Evolution in terms of F1-macro, F1-weighted, UA and WA, for 5 folds, expressed as $\mu \pm \sigma$. The left column represents the evolution using the class-weighting in the loss function, whereas the right column is related with the non-weighting method.

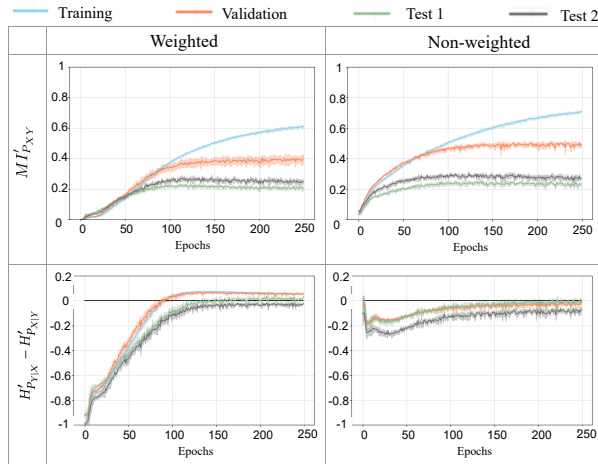


Figure 4: Evolution of $MI'_{P_{XY}}$ and $H'_{P_{Y|X}} - H'_{P_{X|Y}}$, for 5 folds expressed as $\mu \pm \sigma$. The left column represents the evolution using the class-weighting in the loss function, whereas the right column is related with the non-weighting method. Note that ideally $MI'_{P_{XY}} \rightarrow 1$ (7) and $|H'_{P_{Y|X}} - H'_{P_{X|Y}}| \rightarrow 0$ (8).

actual and predicted labels, respectively.

The results from 5-fold cross-validation in terms of $MI'_{P_{XY}}$ and $H'_{P_{Y|X}} - H'_{P_{X|Y}}$ are given in Figure 4, but the full ET for the first fold is shown in Figure 2. As seen in both figures, the training process tries to optimize two equations at the same time—(7) and (8)—resulting in $H'_{P_{Y|X}} - H'_{P_{X|Y}} \rightarrow 0$ and $\Delta H'_{P_Y} - \Delta H'_{P_X} \rightarrow 0$ (8).

Observing the evolution of the training set during the first epochs for both weighting Methods 1 & 2, we see that $H'_{P_{Y|X}} - H'_{P_{X|Y}} < 0$. Solving the equation of divergence with respect to uniformity, we obtain $H_{P_Y} > H_{P_X}$. That is, the entropy of the predicted classes is larger than the entropy of the actual classes, which means that the system assumes that the classes are more imbalanced than they actually are. This makes sense since at this training stage the model still tends to predict classes in an unbalanced manner. This condition is maintained throughout the training process with Method 2.

However, when we apply the inverse prior weighting on the loss function according to Method 1, something we have not seen before happens around the 100th epoch: $H'_{P_{Y|X}} - H'_{P_{X|Y}} > 0$. That is, the entropy of the predicted labels is smaller than the entropy of the actual ones, $H_{P_Y} < H_{P_X}$. At this training stage, the system tends to predict a uniform distribution, since this is the apparent distribution that weighting Method 1 is enforcing. Note that if the distribution of the classes in real operating conditions is unknown, this is the best way to obtain unbiased models that exclusively learn from the intrinsic characteristics of the task at hand and not from the distributions.

After that, the model is trying again to optimize $H'_{P_{Y|X}} - H'_{P_{X|Y}} \rightarrow 0$ (8) and $\Delta H'_{P_Y} - \Delta H'_{P_X} \rightarrow 0$ (8). However, if the model is trained applying Method 1, then $H'_{P_{Y|X}} - H'_{P_{X|Y}} \rightarrow 0^+$ and $\Delta H'_{P_Y} - \Delta H'_{P_X} \rightarrow 0^+$. In contrast, with Method 2, we have $H'_{P_{Y|X}} - H'_{P_{X|Y}} \rightarrow 0^-$ and $\Delta H'_{P_Y} - \Delta H'_{P_X} \rightarrow 0^-$.

Regarding mutual information, the non-weighting method achieves a higher performance for the validation set, as is the case of conventional metrics shown in Figure 3. However, for both Test 1 and Test 2 there is no statistically significant difference between the weighting methods.

6. Conclusions and Future Work

This work highlights the importance of addressing class imbalance in AT and ASC problems, which naturally arise in real-world scenarios. Our findings show that results are highly dependent on the distribution used for training and testing, as well as on the balancing method applied.

In particular, we observed that the training dynamics are strongly influenced by the chosen balancing approach. The use of an inverse prior weighting method can lead the system to assume a uniform distribution that does not reflect reality, which requires careful consideration. As part of future work, we aim to leverage the ET to re-adjust the weighting function before the epoch when $H_{P_Y} < H_{P_X}$.

Furthermore, we plan to extend the training process for more epochs to analyze the convergence behavior of $H'_{P_{Y|X}} - H'_{P_{X|Y}} \rightarrow 0^+$ and $\Delta H'_{P_Y} - \Delta H'_{P_X} \rightarrow 0^+$. Exploring different datasets and models is another key aspect of our future work, along with integrating additional weighting functions, balancing strategies and training methods (e.g., self-supervised learning). Ultimately, we aim to apply these insights to tackle real-world AT and ASC challenges.

7. Acknowledgements

This work was supported by a predoctoral contract (PIPF) funded by Universidad Carlos III de Madrid, under the Programme for Predoctoral Researchers in Artificial Intelligence (Resolution 24/06/2024). This work is also part of the project Vital-IoT funded by INCIBE (Ministry of Digital Transformation and Public Function) and the European Union NextGenerationEU in the framework of the Recovery and Resilience Facility (RRF), Grant PID2021-125780NB-I00 funded by the Spanish Ministerio de Ciencia, Innovación y Universidades MICIU/AEI/10.13039/501100011033 and by “ERDF/EU”.

8. References

- [1] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio Set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780, <https://research.google.com/audioset/>.
- [2] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.
- [3] Y. Xiao and R. K. Das, “WildDESED: An llm-powered dataset for wild domestic environment sound event detection system,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, 2024, pp. 196–200.
- [4] T. Heittola, A. Mesaros, and T. Virtanen, “Tau urban acoustic scenes 2022 mobile, development dataset,” Mar. 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6337421>
- [5] C. Montero-Ramírez, E. Rituerto-González, and C. Peláez-Moreno, “Building artificial situational awareness: Soundscape classification in daily-life scenarios of gender based violence victims,” *Engineering Applications of Artificial Intelligence*, 2025, submitted for review.
- [6] A. Martínez-Serrano, C. Montero-Ramírez, and C. Peláez-Moreno, “Authenticity at risk: Key factors in the generation and detection of audio deepfakes†,” *Applied Sciences (2076-3417)*, vol. 15, no. 2, 2025.
- [7] E. R. González, “Multimodal affective computing in wearable devices with applications in the detection of gender-based violence,” Doctoral Dissertation, Universidad Carlos III de Madrid, Madrid, Spain, feb 2023.
- [8] J. M. Johnson and T. M. Khoshgoftaar, “Survey on deep learning with class imbalance,” *Journal of big data*, vol. 6, no. 1, pp. 1–54, 2019.
- [9] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, “Review of classification methods on unbalanced data sets,” *Ieee Access*, vol. 9, pp. 64 606–64 628, 2021.
- [10] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, “The class imbalance problem in deep learning,” *Machine Learning*, vol. 113, no. 7, pp. 4845–4901, 2024.
- [11] DCASE, “Data-Efficient Low-Complexity Acoustic Scene Classification - Challenge results,” 2024. [Online]. Available: <https://dcase.community/challenge2024/task-data-efficient-low-complexity-acoustic-scene-classification-results>
- [12] D. M. W. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” 2020. [Online]. Available: <https://arxiv.org/abs/2010.16061>
- [13] DCASE, “Sound Event Detection with Heterogeneous Training Dataset and Potentially Missing Labels,” 2024. [Online]. Available: <https://dcase.community/challenge2024/task-sound-event-detection-with-heterogeneous-training-dataset-and-potentially-missing-labels-results>
- [14] F. J. Valverde-Albacete and C. Peláez-Moreno, “100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox,” *PLOS ONE*, vol. 9, no. 1, pp. 1–10, 01 2014. [Online]. Available: <https://doi.org/10.1371/journal.pone.0084217>
- [15] F. J. Valverde-Albacete and C. Peláez-Moreno, “A Framework for Supervised Classification Performance Analysis with Information-Theoretic Methods,” *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 32, no. 11, pp. 2075–2087, Sep. 2020.
- [16] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [17] ———, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 623–656, 1948.
- [18] F. J. Valverde-Albacete and C. Peláez-Moreno, “Two information-theoretic tools to assess the performance of multi-class classifiers,” *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1665–1671, Sep. 2010.
- [19] M. Meila, “Comparing clusterings by the variation of information,” *Learning theory and kernel machines*, pp. 173–187, Jan. 2003.
- [20] F. Schmid, T. Morocutti, S. Masoudian, K. Koutini, and G. Widmer, “CP-JKU submission to dcase23: Efficient acoustic scene classification with cp-mobile.” DCASE2023 Challenge, Tech. Rep., May 2023.