Technical Assessment

Demonstrate your skills in data processing, ML pipeline development, and practical deployment by working with an acoustic dataset of your choice. You are free to use your favourite tech toolkit, whatever you deem most appropriate. You are also welcome to use AI tools to complete this task but be prepared to explain every part of your code during a live technical review. Please note the following task requirements:

1. **Dataset Selection & Preparation**
    - Select a raw acoustic dataset (likely .wav) from a public source (e.g., oceanographic repositories, bioacoustics databases, or audio classification datasets). Keep the size manageable — using a representative subset of the data is sufficient to demonstrate the approach.
    - Load, explore, and preprocess the data where necessary (i.e. raw .wav > ML features)
    - Document any quality issues, artifacts, or preprocessing decisions

2. **Data Splitting Strategy**
    - Provide a detailed written explanation of how you are splitting the acoustic data for model training to avoid:
        - Data leakage
        - Temporal bias
        - Class imbalance issues
        - Overfitting to specific recording conditions/equipment

      If you're using a pre-split dataset (e.g. a benchmark set), please explain how you would have split it.

3. **Model Architecture Selection**
    - In the scope of this task it is not required to train an ML model from scratch but please describe in brief the approach you would take to architecture selection, training, and tuning, and what machine learning architecture you would start with (with reasoning). Prepare some example features to fit this choice (no need to pre-process the full dataset, especially if large). Also think about the necessary computational requirements for your architecture, both for training and deployment.

4. **Visualisation**
    - Create a simple web-based visualisation interface to explore the dataset (no Notebooks)
    - Include relevant acoustic features (e.g., spectrograms, waveforms, feature distributions)
    - Add anything else you deem relevant or useful

5. **Code Quality**
    - Use version control (Git)
    - Write clean, documented code and include a README with setup instructions

Deliverables

Please deliver the results in a GitHub repository including your code and README and attach relevant documentation and decision-making. Include dataset description and source, setup instructions, data splitting and architecture rationale, and any assumptions or trade-offs made.

Results are to be submitted by **Wednesday the 29th of October, end of day**, afterwards we will book in a live technical code review session between November 6-7 for you to demonstrate your solution.

We will pay attention to data handling and preprocessing decisions, understanding of signal processing, architecture selection, code organisation, and rationale of explanations.

Good luck, we're looking forward to seeing what you come up with!

Any questions, you can drop us a message:

Astrid van Toor – avtoor@blueoasis.pt

Himanshu Singhal – hsinghal@blueoasis.pt