Week 3 Narrative Report: Synesthetic Learning Project

Chris Jallaine Mugot

Department of Data Science - College of Information Technology and Computing
University of Science and Technology of Southern Philippines
Lapasan, Cagayan de Oro City 9000, Philippines

This week's work advanced the technical backbone of the Synesthetic Learning project by transitioning from dataset preparation into **model design and initial experimental setup**. Building on the earlier stages where we validated the audio and abstract art datasets and aligned their emotional labels, Week 3 focused on defining the learning pipeline, conducting exploratory modeling decisions, and implementing the first experimental components of the system.

Our primary technical objective for this week was to determine how the model will process audio and map it to a corresponding emotion class, which will later be used to retrieve or match abstract artworks. To begin, I outlined and refined the initial model architecture. After reviewing existing literature on audio-based emotion recognition, we decided on a pipeline that converts raw audio into **Mel-spectrogram representations**, which provide an efficient and expressive time–frequency image suitable for convolutional neural networks. This choice balances interpretability, computational efficiency, and compatibility with standard deep learning architectures.

My contributions this week centered on designing the core system structure and initiating the first model experiments. I developed preprocessing scripts that load raw audio files, normalize sampling rates, generate Mel-spectrograms, and prepare them as standardized image-like tensors for training. These preprocessing functions form a crucial foundation for the training pipeline, ensuring that all audio files—despite coming from multiple datasets—are represented in a consistent and comparable form.

I also explored multiple candidate deep learning architectures for the emotion classifier, comparing the advantages of **CNN-based models** against more modern approaches using **pretrained audio backbones** such as PANNs and Audio Spectrogram Transformers. For the initial experiment, we settled on a baseline model using a lightweight convolutional architecture. This provides a manageable starting point for benchmarking and helps us understand how well simple models can capture the emotional structure in the audio dataset before moving to more complex, pretrained, or transformer-based models in later weeks.

To evaluate feasibility, I implemented and ran a small-scale initial training experiment using a subset of the data. This experiment allowed us to verify that the spectrogram generation pipeline works, the model can load and process batches, and the training loop runs without errors. The early results showed reasonable convergence behavior and confirmed that the emotional categories are learnable from spectrogram inputs, validating the soundness of the overall approach.

Another important aspect of this week's progress involved formalizing how the system will perform **cross-modal alignment**. I designed the retrieval mechanism where the predicted audio emotion will be matched to an artwork whose mapped emotional label corresponds to the classification output. Although the deeper synesthetic alignment model—where audio and art share a unified embedding

space—will be implemented in later phases, this week's design ensures that the baseline retrieval system is well-structured and computationally feasible.

Overall, my contributions in Week 3 focused on structuring and implementing the foundational model pipeline, exploring architecture choices, and conducting the initial experimental validation needed to progress into full-scale training. This work ensures that the system is technically grounded, scalable, and ready for iterative refinement in the subsequent weeks, where we will begin to evaluate model performance, optimize architecture selection, and prepare for cross-modal embedding experiments.