From Sound to Sight: An End-to-End Audio-to-Art Generative Intelligence Pipeline

Visaviern Mosqueda, Chris Jallaine Mugot, Frency Rayne Montesclaros, and Qylle Christian Quiño

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

Advances in artificial intelligence have increasingly enabled systems to move beyond single-modality understanding toward richer, multimodal reasoning. In this project, we present an end-to-end audio-to-art generative pipeline that transforms emotional cues extracted from sound into visually expressive artwork. Rather than focusing solely on artistic generation, the system was designed as a comprehensive predictive-and-generative analytics framework, emphasizing emotional consistency, representation alignment, and evaluability across all stages of the workflow.

The project begins with the recognition that human emotion is inherently multimodal. Audio signals, particularly speech and vocal expressions, carry rich affective information that can be computationally extracted and interpreted. Visual art, on the other hand, has long been used as a medium to convey emotion. Bridging these two modalities therefore presents an opportunity to explore how learned emotional representations can be transferred, aligned, and ultimately synthesized into novel creative outputs.

The primary audio dataset used in this study consists of over six thousand labeled audio samples categorized into three core emotional classes: Angry, Happy, and Sad. These classes were chosen to balance expressive diversity with modeling feasibility. In parallel, a curated artwork dataset was prepared, containing thousands of images annotated with corresponding emotional labels. To ensure semantic consistency across datasets, label normalization was performed, resolving discrepancies such as mapping “fearful” annotations into the “sad” category. This early harmonization step proved critical in maintaining alignment throughout the pipeline.

Before any modeling took place, a rigorous audio preprocessing workflow was applied. Raw audio files vary widely in duration, amplitude, and silence distribution, all of which can introduce noise into learning algorithms. To address this, each audio sample was resampled to a fixed sampling rate, normalized using root-mean-square scaling, trimmed to remove silence, and padded or truncated to a uniform length. These processed waveforms were then transformed into Mel-spectrograms, which provide a perceptually meaningful time–frequency representation of sound. By normalizing these spectrograms and storing them as tensors, the pipeline achieved both computational efficiency and reproducibility.

The first major modeling stage involved audio emotion classification. Several convolutional neural network architectures were evaluated using the preprocessed spectrograms as input. Each model was trained to predict emotional class labels while simultaneously learning latent embeddings that capture emotional structure. Performance was evaluated using accuracy and macro-averaged F1 scores to account for class imbalance. Among the tested models, a CNN-based architecture emerged as the most effective baseline, achieving a validation accuracy exceeding eighty percent. This model provided a strong foundation for subsequent optimization and representation learning.

To further improve predictive performance, hyperparameter tuning was conducted using an automated optimization framework. Key parameters such as learning rate, dropout probability, and architectural depth were systematically explored. This tuning process led to a notable improvement in validation accuracy, increasing performance to approximately eighty-seven percent. Beyond the numerical gain, this stage reinforced the importance of principled optimization in predictive analytics, demonstrating that model architecture alone is insufficient without careful parameter calibration.

Once an optimized classifier was obtained, attention shifted from prediction to representation. The learned audio embeddings were extracted from the trained network and treated as compact emotional descriptors. In parallel, image embeddings were extracted from the artwork dataset using a pretrained EfficientNet backbone, repurposed as a feature extractor. These image embeddings captured high-level visual semantics while remaining agnostic to the downstream generative task.

A key contribution of the project lies in the construction of a shared latent space that aligns audio and image embeddings. Separate projection networks were trained to map both modalities into a common 128-dimensional space. The objective was not merely dimensionality reduction, but semantic alignment: embeddings representing the same emotion, regardless of modality, should cluster together. Visualization using dimensionality-reduction techniques confirmed that this objective was largely achieved, with emotion emerging as the dominant organizing factor rather than modality.

To quantitatively evaluate this shared representation, cross-modal retrieval experiments were performed. For each audio embedding, the system retrieved the nearest image embedding based on cosine similarity. The resulting retrieval accuracy exceeded ninety percent, indicating that the shared space successfully encoded emotional correspondence between sound and image. This result served as a critical validation step, demonstrating that the pipeline had learned meaningful multimodal representations rather than superficial correlations.

Building upon this aligned representation, the project progressed to its generative phase through the implementation of a conditional Generative Adversarial Network. The GAN was designed to generate artwork images conditioned on projected audio embeddings, effectively translating emotional information from sound into visual form. Architectural choices such as conditional batch normalization and projection-based discrimination were employed to ensure stable training and strong conditioning signals. The GAN was trained using hinge loss functions over multiple epochs, with careful monitoring of generator and discriminator dynamics.

Evaluating generative models poses inherent challenges, particularly when the goal extends beyond visual realism to semantic fidelity. To address this, an emotion-consistency metric was introduced. Generated images were embedded using the same visual feature extractor and compared against real artworks in feature space. The proportion of generated images whose nearest real neighbors shared the intended emotion served as a proxy for semantic correctness. The final model achieved an emotion-consistency score of approximately seventy-five percent, indicating that most generated outputs preserved the emotional intent of the input audio.

The culmination of the project is a fully integrated, end-to-end pipeline capable of accepting raw audio input and producing emotionally aligned artwork. From preprocessing and emotion classification to embedding projection, retrieval, and image generation, each component feeds seamlessly into the next. Importantly, the pipeline also provides interpretability through nearest-neighbor retrieval, allowing users to contextualize generated outputs within the existing artwork dataset.

Overall, this project demonstrates how predictive analytics, representation learning, and generative modeling can be combined into a coherent multimodal system. Rather than treating classification, retrieval, and generation as isolated tasks, the pipeline emphasizes continuity and validation across stages. The results highlight the feasibility of emotion-aware generative systems and underscore the value of rigorous workflow design in applied data science. Through this work, the audio-to-art pipeline stands not only as a creative application, but as a structured demonstration of end-to-end machine learning system development.