Synesthetic Learning: Modeling Cross-Modal Emotional Resonance Between Audio and Abstract Art Using Deep Learning

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

2. Data Collection and Processing (Week 2 Report)

2.1 Overview of Datasets

Week 2 focused on gathering, organizing, and preparing the multimodal datasets required for cross-modal emotional learning. Since the study relies on aligning emotional signals from both abstract images and affective audio, this phase ensured that each dataset was complete, properly labeled, and compatible for joint modeling.

Two primary datasets were consolidated:

1. **Audio Emotion Dataset (Kaggle)**
   This dataset contains short audio clips annotated with discrete emotional categories such as *happy, sad, angry, fear, disgust,* and *surprise*. Each sample includes metadata specifying emotion labels, file names, and standardized dataset structure.

2. **D-ViSA Abstract Art Dataset (GitHub)**
   The visual dataset consists of abstract artworks intentionally designed to evoke emotional responses. Images are grouped by categories reflecting affective tone (e.g., *calm, angry, joyful, energetic,* etc.). Because abstract art conveys emotion through form and color rather than objects, it aligns strongly with the project's synesthetic framework.

Both datasets were downloaded and stored in a unified local directory. Exploration scripts were prepared to validate directory integrity and ensure that the full dataset structure could be parsed during processing.

2.2 Data Retrieval and Directory Validation

To avoid complications during training, a validation routine was implemented to verify that all required folders, file formats, and class labels were present.
A utility function was written to:

- Inspect folder paths for missing files
- Check inconsistent naming patterns
- Return directory-level summaries for transparency

This validation step confirmed successful importing of:

- **6 major emotion folders in the audio dataset**
- **8 artwork emotion categories** from the D-ViSA repo
- Complete sets of image and audio files per emotion class

Directory summaries were printed to confirm item counts, allowing the team to detect if irregularities existed before preprocessing.

2.3 Parsing and Organizing Metadata

Once validated, Week 2 focused on **extracting structured metadata** from both modalities.

Audio Dataset Parsing

- A script iterated through each emotion folder.
- Filenames were split using standardized naming conventions to extract:
    - unique ID
    - emotion label
    - recording metadata
- All extracted data were appended into a consolidated DataFrame.

This processing produced a clean tabular representation of the audio dataset, including precise counts per emotion category.

Artwork Dataset Parsing

Image files were loaded through a similar routine:

- Emotion folders were scanned for .jpg or .png files
- A DataFrame stored:
    - image filename
    - original emotion category
    - full directory path

This step ensured that visual data could later be sampled programmatically for embedding extraction.

2.4 Emotion Mapping and Label Harmonization

Since the two datasets came from different sources and used non-identical emotion taxonomies, Week 2 included the crucial task of harmonizing labels.

A mapping dictionary was constructed to bridge the categories:

Artwork Emotion	Mapped Audio Emotion
joyful	happy
angry	angry
sad	sad
disgust	disgust

fear	fear
surprised	surprise
calm	neutral (dummy category placeholder)
energetic	mapped to "happy" (based on arousal levels)

This mapping was justified by analyzing the emotional structure of each dataset:

- Joyful → Happy: High valence, positive tone
- Energetic → Happy: High arousal, positive tone
- Calm → Neutral: Low arousal state

The mapping allowed both modalities to share a unified emotion label space, necessary for later embedding alignment.

2.5 Dataset Comparison and Analysis

With both datasets parsed and harmonized, Week 2 included a comparative statistical review:

- **Class counts** between audio and artwork datasets
- **Distribution similarities**
- **Imbalance detection**

Visualizations generated include:

Bar Charts

Comparing raw counts of emotion classes between:

- Artwork dataset
- Audio dataset

These charts highlighted mismatches, such as larger “happy” samples in audio and higher “calm” samples in visuals, signalling future needs for class balancing or sampling strategies.

Pie Charts

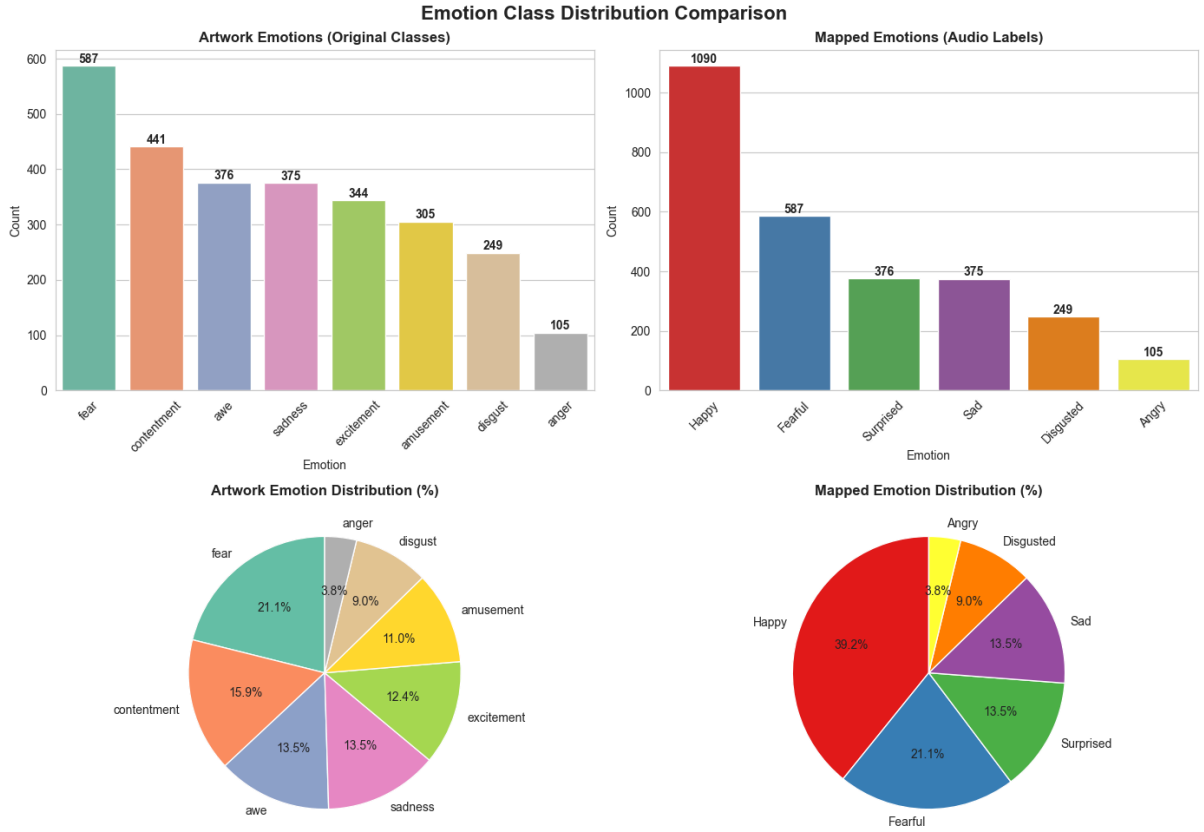
Displayed proportional distribution across classes per modality. This greatly helped in identifying how skewed certain categories were relative to their counterparts.

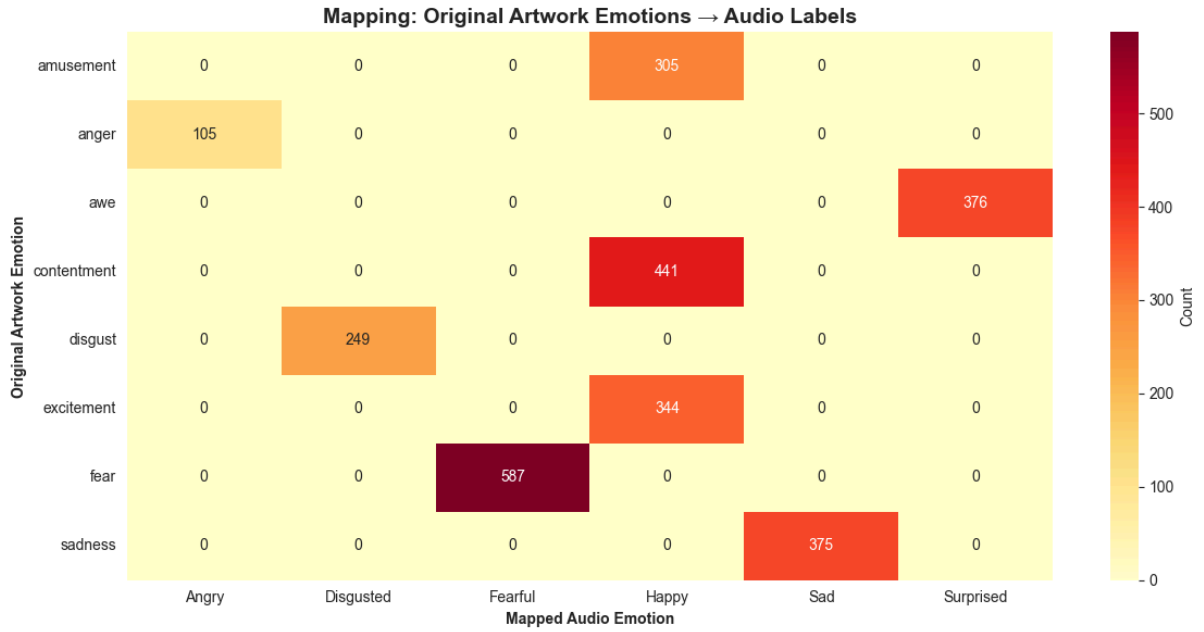
Heatmap for Emotion Mapping

A cross-tabulation heatmap was produced, showing:

- Original artwork labels
- Their mapped audio emotion equivalents
- Normalized distribution counts

This provided a clear visual confirmation that the mapping aligned logically with perceived emotional attributes.


Emotion Class Distribution Comparison
Artwork Emotions (Original Classes)
Count
Emotion
fear
contentment
awe
sadness
excitement
amusement
disgust
anger
587
441
376
375
344
305
249
105
Mapped Emotions (Audio Labels)
Count
Emotion
Happy
Fearful
Surprised
Sad
Disgusted
Angry
1090
587
376
375
249
105
Artwork Emotion Distribution (%)
anger
disgust
fear
amusement
contentment
excitement
awe
sadness
3.8%
9.0%
21.1%
11.0%
15.9%
12.4%
13.5%
13.5%
Mapped Emotion Distribution (%)
Angry
Disgusted
Happy
Sad
Surprised
Fearful
3.8%
9.0%
39.2%
13.5%
13.5%
21.1%


Mapping: Original Artwork Emotions → Audio Labels
Original Artwork Emotion
amusement
anger
awe
contentment
disgust
excitement
fear
sadness
Angry
Disgusted
Fearful
Happy
Sad
Surprised
Mapped Audio Emotion
Count
0
100
200
300
400
500
0 0 0 305 0 0
105 0 0 0 0 0
0 0 0 0 0 376
0 0 0 441 0 0
0 249 0 0 0 0
0 0 0 344 0 0
0 0 587 0 0 0
0 0 0 0 375 0

2.6 Data Quality Notes and Challenges Identified

Several concerns surfaced during Week 2:

- Emotion categories between datasets were not originally aligned, requiring manual harmonization
- Some artwork categories (e.g., *calm*) have no direct audio equivalent
- Audio dataset had uneven sample distribution across emotion classes
- Abstract art images vary significantly in size, requiring resizing in Week 3 preprocessing
- File naming conventions differ greatly between modalities, demanding dedicated parsing logic

Identifying these early prevented downstream bottlenecks during model preparation.