UE23AM342AA1 : Advanced Foundations for ML

Multimodal Generative Augmentation and Cross-Modal Learning for Bipolar Disorder

Review 2

Project Guide – Dr. Arti Arya

Mentor – Princia D'souza

Team 01

Abhishek P - PES2UG23AM002

Harsha - PES2UG23AM042

Lohit J - PES2UG23AM054

Literature Survey

1. Multimodal Temporal Machine Machine Learning for Bipolar Disorder and Depression Recognition (Ceccarelli & Mahmoud, 2022)

Used video, audio and text together, with RNNs/LSTM to model temporal evolutions of features for BD vs depression.

Demonstrates the value of multimodal and temporal modeling in BD; useful baseline for your fusion and cross-modal idea.

2. The PRIORI Emotion Dataset: Dataset: Linking Mood to Emotion Emotion Detected In-the-Wild (Khorram et al., 2018)

Collected smartphone conversational speech from individuals with BD; showed correlation between emotion (activation/valence) and mood states.

Important dataset and method paper for BD audio modality; supports your use of use of speech and gives a landmark for generative augmentation.

3. Acoustic and Facial Features From From Clinical Interviews for Machine Machine Learning–Based Psychiatric Psychiatric Diagnosis: Algorithm Algorithm Development (Birnbaum (Birnbaum et al., 2022)

Extracted facial and acoustic features from clinical interviews of BD and schizophrenia; used schizophrenia; used ML to differentiate diagnoses.

Shows how facial and audio data can work together for psychiatric diagnosis. Useful to justify using those modalities in your project.

4. Portable technologies for digital digital phenotyping of bipolar disorder: A systematic review (2021)

Systematic review of smartphone apps, wearables, audio/video in BD; noted small small sample sizes, heterogeneity, class imbalance.

Very relevant for your “gap” section — confirms issues of data scarcity and imbalance, imbalance, which your project aims to address.

5. Speech analysis for mood mood state characterization in characterization in bipolar patients (Agurto et al., 2013) 2013)

Focused on voice segmentation and spectral features (pitch, jitter) to classify mood states in BD.

Useful for your audio feature engineering; gives features to start with and with and baseline for augmentation.

6. Differentiation between depression and bipolar bipolar disorder in child and adolescents by voice voice features (2023)

Used voiceprint features in children/adolescents (MDD vs BD vs controls) and ML classifiers.

Adds context about voice modality in BD vs other disorders and supports the need for specialized data and augmentation for for underrepresented classes.

7. A Hybrid Model For Bipolar Disorder Classification From Visual Visual Information (Abaei & Al Osman, 2020)

Used CNN and LSTM on video facial features (structured interviews) to classify BD into remission, hypomania, and mania.

Useful for your video modality — shows how visual patterns can help classify BD states. classify BD states. Good for your multimodal pipeline.

8. Multimodal machine learning for language and and speech markers identification in mental health health (2024)

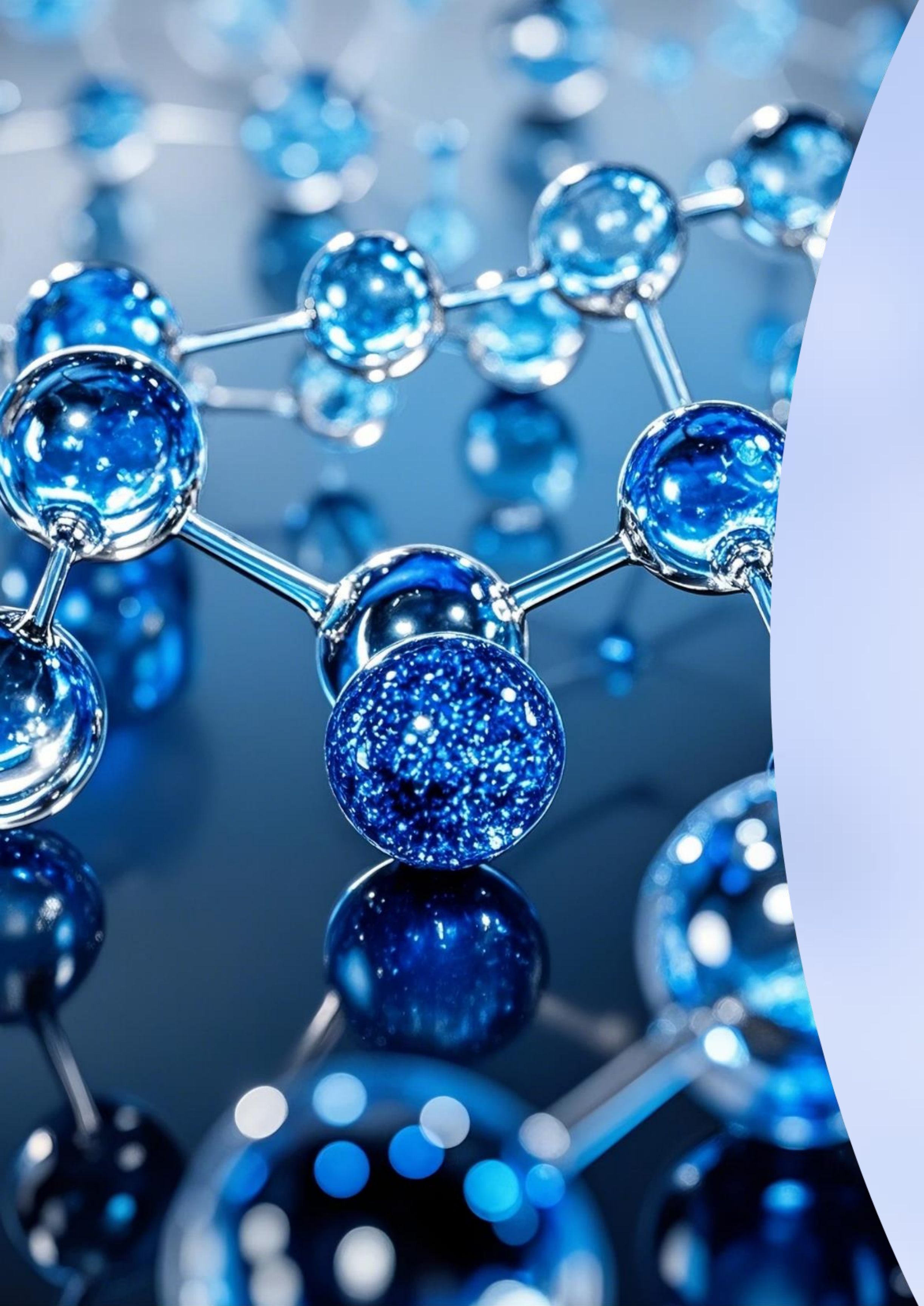
Review/analysis of audio and text markers across mental health disorders; found that multimodal models sometimes outperform unimodal.

Reinforces your justification that multimodal is better than unimodal; shows the value of text and audio, which fits your plan for cross-modal for cross-modal augmentation.

9. Predicting Mood Disorder Symptoms with Remotely Collected Collected Videos Using an Interpretable Multimodal Dynamic Dynamic Attention Fusion Network (Banerjee et al., 2021)

Collected video, audio, and text from a smartphone app for mood disorder (including depression), used transformer, attention, and SHAP for interpretability.

Demonstrates interpretable multimodal fusion, which you can adapt to BD and incorporate into your explainability novelty.



10. Leveraging Embedding Techniques in Multimodal Machine Learning for Mental Mental Illness Assessment (Hassan et al., 2025)

Used embedding models (audio, video, text) with CNN/BiLSTM and fusion, achieved high balanced accuracy for depression/PTSD; explored chunking and LLM predictions.

Provides state-of-the-art in embedding and fusion; you can borrow embedding embedding techniques and chunking for your generative and cross-modal method. modal method.

Dataset

PRIORI (Predicting Individual Outcomes for Rapid Intervention) Intervention)

Developed at the University of Michigan – Prechter Bipolar Research Program.

Other datasets, request mail sent.

Proposed Approach

Raw Audio Data (PRIORI)

↓

Preprocessing & Feature Extraction (MFCC, Spectral, Prosodic)

↓

Generative Model (GAN/VAE) → Synthetic Data Generation

↓

Augmented & Balanced Dataset

↓

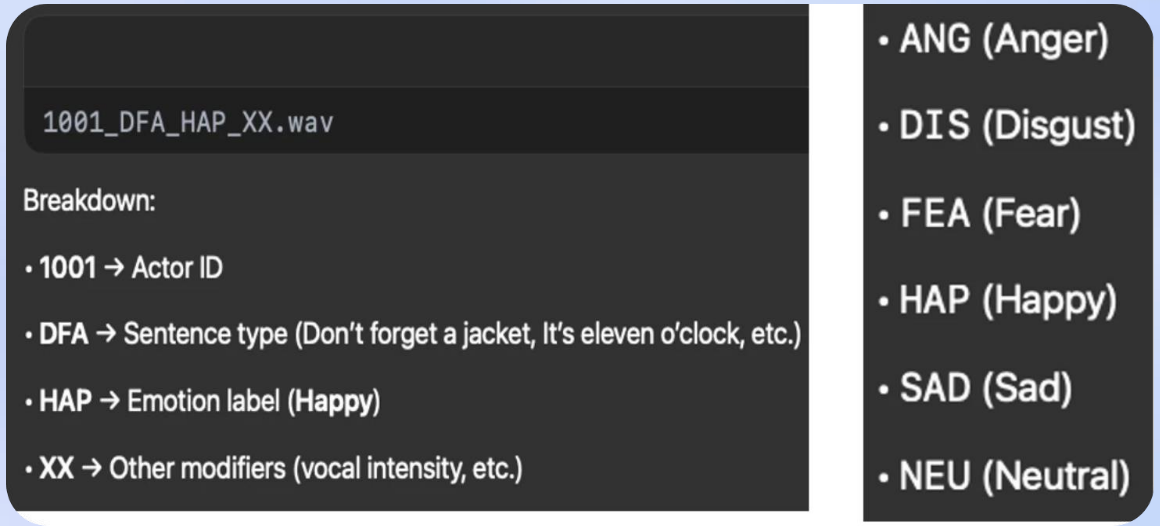
Cross-Modal Learning (Audio + Synthetic Video/Text Embeddings)

↓

Mood Classification (Manic/Depressed/Euthymic)

↓

Performance Evaluation & Explainability



```
1001_DFA_HAP_XX.wav
```

Breakdown:

- **1001** → Actor ID
- **DFA** → Sentence type (Don't forget a jacket, It's eleven o'clock, etc.)
- **HAP** → Emotion label (**Happy**)
- **XX** → Other modifiers (vocal intensity, etc.)

- ANG (Anger)
- DIS (Disgust)
- FEA (Fear)
- HAP (Happy)
- SAD (Sad)
- NEU (Neutral)

Thank You