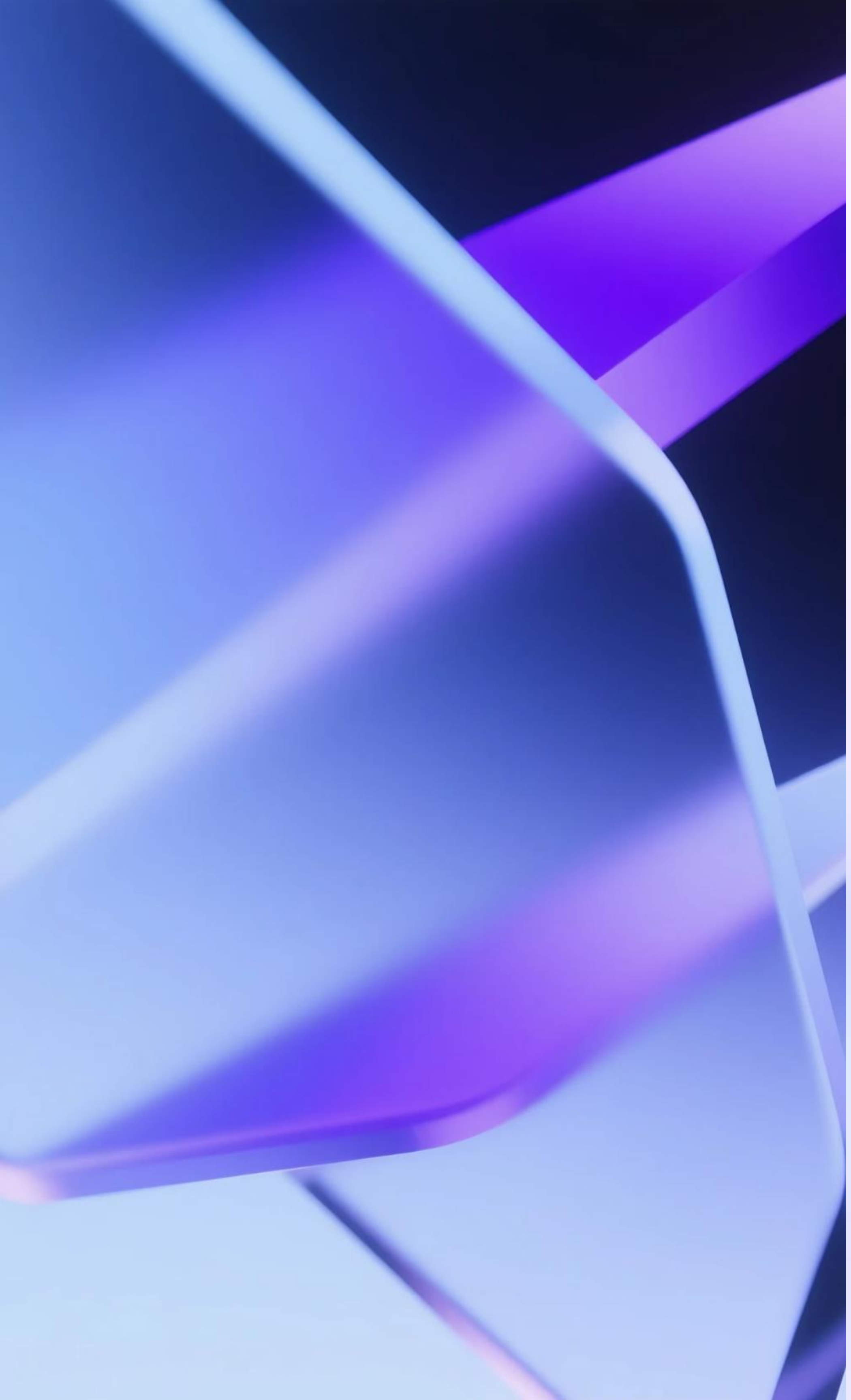


UE23AM342AA1 : Advanced Foundations for ML

Multimodal Generative Augmentation and Cross-Modal Learning for Bipolar Disorder

Project Guide : Dr. Arti Arya

Mentor

Princia D'souza

Abhishek P	PES2UG23AM002
Harsha	PES2UG23AM042
Lohit J	PES2UG23AM054

OUTLINE

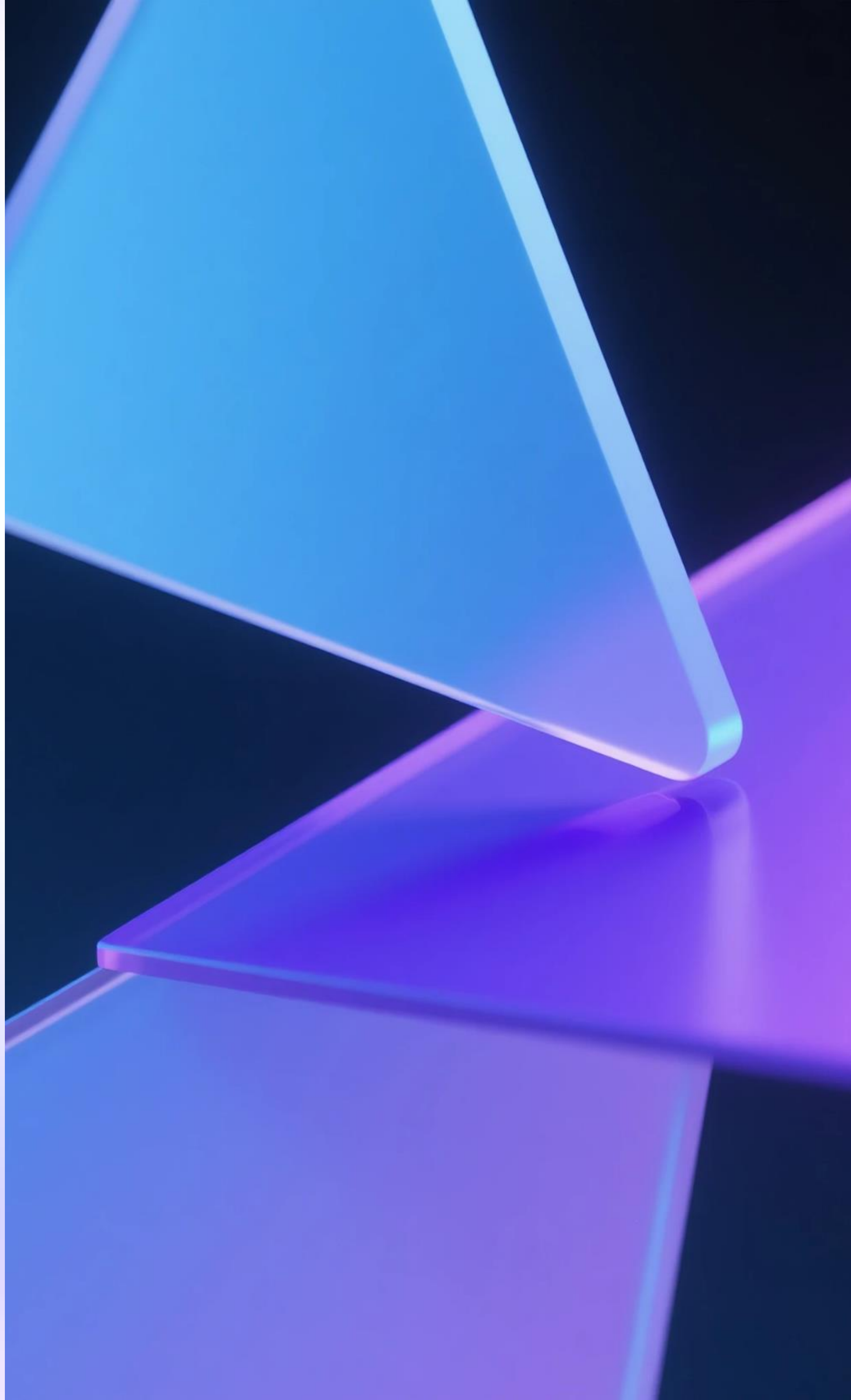
- **Problem Statement**
- **Background Study**
- **Research Gaps**

Problem Statement

- Bipolar Disorder (BD) is characterized by extreme mood swings (mania, hypomania, depression) that significantly impair life functioning. Early, accurate diagnosis and monitoring are challenging due to
(a) overlapping symptoms with other disorders (e.g., depression)
(b) fluctuating mood states
(c) limited labelled data for certain states (mania especially)
(d) missing or weak modalities (e.g., audio, video, clinical notes). Existing ML systems often use single modalities and can’t handle cases when one modality is missing.

Multimodal Generative Augmentation + Cross-Modal Learning for Bipolar Disorder aims to address data scarcity and imbalance across mood states by generating synthetic examples in multiple modalities (audio, video, text/clinical notes) and building models that can learn from cross-modal relationships (e.g., using info from video to help audio predictions when audio is noisy or missing).

The goal is to improve detection / classification of different mood states in BD (mania, depression, euthymia), especially for underrepresented classes, and make the system robust to missing or noisy modalities.



Background Study

Multimodal / Cross-modal in Bipolar Disorder

- Multimodal Temporal Machine Learning for Bipolar Disorder and Depression Recognition (Ceccarelli & Mahmoud, 2022) uses video + audio + text to classify BD vs depression.
- “Multimodal Emotion Integration in Bipolar Disorder: facial + prosodic channels” shows BD patients have delayed response time in cross-modal emotional tasks.
- “Combined Processing of Facial and Vocal Emotion in Remitted Patients With Bipolar I Disorder” studying how BD remitted patients match facial and prosodic emotional cues.
- Imaging studies: e.g., combining structural, functional, and diffusion MRI improves classification of BD vs healthy controls.

These works show multimodal approaches are feasible and helpful, but they often use only “observed real data” (no synthetic augmentation) and/or do not address missing modalities or heavy imbalance across mood states. Also, many focus on imaging or emotional perception tasks rather than full diagnosis across mood states

Generative Modeling / Data Augmentation in Mental Health

- “Using a generative model of affect to characterize affective variability” (PNAS, 2022) models mood volatility vs noise in BD + BPD; *but this is on affect ratings (time series), not multimodal raw data (video/audio/text).*
- *Very few works generate synthetic video or audio or text data in BD context, especially cross-modal synthetic data.* - “Auditory-Visual Speech Integration in BD: Preliminary” *studies how BD integrates speech & visuals but does not generate data.*

There is a gap in applying generative augmentation across multiple raw modalities, not just modeling affect or using imaging. Also, cross-modal synthetic data (e.g. generating audio from video) is almost unexplored in BD.

Imbalance & Underrepresentation

Many studies show fewer samples of mania or hypomania vs depressive/euthymic. E.g., audio / smartphone datasets tend to have fewer manic episodes. PRIORI dataset (speech) shows manic state much less frequent. - Imaging datasets also often have more healthy controls or depressed states compared to mania. - Few works report how models perform under missing modalities or when one modality is weak.

Imbalance in mood-state classes is a real problem → models tend to be biased toward more frequent classes. Cross-modal learning + synthetic augmentation can help with this. Also, handling missing modalities is practically useful

Research Gaps

Lack of Generative Augmentation Across All Modalities
Existing BD multimodal studies generally use real observed data but rarely generate synthetic audio, video, and text together.
Few works do augmentation in more than one modality, much less cross-modal generation (e.g., generating audio from video or text).

Handling Missing/Noisy Modalities
Many multimodal models assume all modalities are present and well-recorded. In real clinical or remote settings, one or more modalities may be missing or corrupted (e.g., poor audio, no video). There is little work on models trained to be robust or to fill in missing modalities via cross-modal learning.

Class Imbalance for Mood States
The mania / hypomania classes are underrepresented in datasets, so models often underperform for those states. There is limited work specifically targeted to boost performance on underrepresented mood states via data augmentation or weighting.

Explainability & Clinical Interpretability
Clinicians need not only accurate classification, but which features / modalities are influencing decision, especially when synthetic data are used. Few BD works provide interpretability across modalities (audio, video, text).

Thank You