Dissertation on

Multimodal Generative Augmentation and Cross-Modal Learning for Bipolar Disorder

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science &
Engineering (AI & ML)
UE23CS342AA1 – Mini Project Report

Submitted by:

Name:Abhishek P	**SRN1:PES2UG23AM002**
Name:Harsha	**SRN2:PES2UG23AM042**
Name:Lohit J	**SRN3:PES2UG23AM054**

Under the guidance of

Prof. Arti Arya
PES University

Aug - Nov 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India
FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

Multimodal Generative Augmentation and Cross-Modal Learning for Bipolar Disorder

is a bonafide work carried out by

Name:Abhishek P	**SRN:PES2UG23AM002**
Name:Harsha	**SRN:PES2UG23AM042**
Name:Lohit J	**SRN:PES2UG23AM054**

In partial fulfilment for the completion of fifth semester elective course Advanced Foundations of Machine Learning (UE23CS342AA1) in the Program of Study - Bachelor of Technology in Computer Science and Engineering (AI/ML) under rules and regulations of PES University, Bengaluru during the period Aug 2025 – Nov. 2025. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 5th semester academic requirements in respect of project work.

Signature	Signature Dr.	Signature
Prof. Arti Arya	Sandesh B J	Dr. B K Keshavan
HOD AIML Dept	Chairperson	Dean of Faculty

External Viva

Name of Examiners
Signature with Date

1. ________________

____

2. ________________

____

DECLARATION

We hereby declare that the mini-project entitled **“Multimodal Generative Augmentation and Cross-Modal Learning for Bipolar Disorder”** has been carried out by us under the guidance of <Prof. Arti Arya, Designation> and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester Aug – Nov. 2025. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES2UG23AM054	**Lohit J**	Lohitj
PES2UG23AM042	**Harsha**	Harsha
PES2UG23AM002	**Abhishek P**	Abhishe

ACKNOWLEDGEMENT

I would like to convey my sincere gratitude to **Prof. Arti Arya**, Head of the Department of Artificial Intelligence and Machine Learning, for her exceptional guidance and constant encouragement throughout the development of this mini-project, “**Multimodal Generative Augmentation and Cross-Modal Learning for Bipolar Disorder.**” Her deep expertise, constructive feedback, and persistent motivation greatly enriched my understanding and were vital in shaping the project to completion.

I am also thankful to my mentor **Princia D’souza** whose early direction, consistent support, and thoughtful suggestions helped establish a clear path for the project. His mentorship provided valuable clarity and strengthened the overall execution of the work.

I remain truly appreciative of their time, trust, and continued support, all of which contributed significantly to the successful completion of this project.

ABSTRACT

Bipolar Disorder (BD) presents major diagnostic challenges due to fluctuating mood states and limited availability of labelled clinical data, particularly for manic episodes. Most existing machine-learning approaches rely on a single modality, often audio, and struggle when data are scarce or unbalanced. To address these limitations, this project focuses on multimodal generative augmentation driven entirely by audio signals, where text data are synthetically generated from speech features. By converting audio representations into meaningful pseudo- text using generative models, the dataset becomes richer and more diverse, helping compensate for class imbalance and improving the representation of under-sampled mood states.

Building on this augmented dataset, a cross-modal learning framework is developed to jointly learn from audio features and their corresponding synthetic text embeddings. This enables the model to leverage complementary information from both modalities, improving robustness and mood-state classification accuracy for manic, depressive, and euthymic states. The proposed approach demonstrates how generative text augmentation paired with cross-modal fusion can enhance performance even when only a single real modality is available, offering a practical path toward scalable and reliable BD monitoring systems.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Bipolar Disorder (BD) is a complex mood disorder characterized by rapid and unpredictable shifts between manic, depressive, and stable states. Traditional clinical assessments rely heavily on self-reports and clinician observations, which are subjective, time-consuming, and often fail to capture subtle behavioural variations. With the rise of digital phenotyping, researchers have begun using audio, video, text, and smartphone signals to automatically identify mood patterns — yet **most existing datasets are small, imbalanced, and unimodal**, making it difficult for machine learning models to generalize reliably.

Recent studies show that **multimodal signals** like speech prosody, facial expressions, and linguistic patterns contain complementary cues about mood states. However, multimodal datasets for BD are extremely scarce, and many modalities are missing or incomplete. This creates a major gap: current models struggle with **data scarcity, class imbalance**, and **inconsistent modality availability**

This project aims to address these limitations by building a **Multimodal Generative Augmentation and Cross-Modal Learning framework** for mood classification in Bipolar Disorder. Using the PRIORI dataset as the base audio source, we design a pipeline that:

1) Extracts temporal audio features,

2) Generates high-quality synthetic samples using GAN/VAE models to balance mood classes,

3) Learns relationships across modalities (audio → video/text embeddings) using cross-modal learning, and

4) Performs mood classification into manic, depressed, and euthymic states.

By combining generative modeling with multimodal fusion, the project targets better generalization, robustness to missing modalities, and improved mood prediction accuracy. This approach directly responds to the key challenges highlighted in existing literature — limited data, imbalance, and lack of cross-modal modeling — and provides a scalable direction for next-generation mental health assessment systems.

PROBLEM STATEMENT

To overcome the challenges of data scarcity, class imbalance, and missing modalities in Bipolar Disorder (BD) mood-state classification, this project proposes a **Multimodal Generative Augmentation and Cross-Modal Learning Framework**. Current BD datasets—especially audio, video, and text—are limited in size and rarely contain all modalities for every sample. This restricts the ability of machine learning models to learn stable patterns for manic, depressive, and euthymic states.

Therefore, the proposed solution aims to develop a system that can:

1.**Use generative models (GAN/VAE) to create synthetic audio data** for underrepresented mood classes, helping balance the dataset and improve generalization.

2.Leverage cross-modal learning to generate or infer missing modality embeddings (e.g., audio → pseudo-video/text), ensuring that multimodal fusion is possible even when real data is incomplete.

3.Extract robust temporal features (MFCCs, prosodic cues, spectral patterns) that capture emotional and behavioral variations relevant to bipolar mood shifts.

4.**Build a multimodal fusion classifier** that integrates real and synthetic embeddings to predict manic, depressive, and euthymic states more accurately than unimodal baselines.

5.Evaluate performance using class-balanced metrics and interpretability tools, ensuring reliable and explainable mood-state predictions.

LITERATURE SURVEY

Multimodal Temporal Machine Learning for Bipolar & Depression Recognition

1. Introduction

Ceccarelli & Mahmoud (2022) proposed one of the earliest multimodal temporal models combining **audio, video, and text** to distinguish Bipolar Disorder from depression. Their approach used LSTM/RNN-based architectures to capture temporal evolution in speech, facial expressions, and language. The study highlighted that bipolar mood shifts are inherently dynamic, making temporal multimodal modelling more effective than static signal analysis.

2. Importance of Temporal Dynamics

The model processes sequential modalities frame-by-frame, extracting prosodic cues, facial action units, and linguistic sentiment over time. This allows the system to track mood oscillations that occur within a single interview. By learning long-term dependencies, the model captures subtle behavioural clues often missed by clinicians.

3. Multimodal Fusion Strategy

Ceccarelli & Mahmoud use **late fusion**, combining independent modality features before classification. Each modality contributes complementary information:

Audio → stress, prosody, motor speech changes
Video → facial tension, slowed movements
Text → polarity, emotional tone
This validated the need for multimodal systems in BD diagnosis.

4. Impact on BD Modelling

Their results showed significant improvements over unimodal baselines, proving that integrating multiple human signals is critical for identifying depression vs bipolar states. This directly supports the motivation for multimodal generative augmentation in our work.

2. The PRIORI Emotion Dataset (Khorram et al., 2018)

2.1. Introduction

The PRIORI dataset is one of the most influential resources for BD research. It contains **in-the-wild smartphone speech** collected from individuals with bipolar disorder over several months. The unique strength of PRIORI is that it links emotion cues (activation, valence) with clinical mood ratings.

2.2. Speech–Mood Correlation

The study showed that acoustic features strongly correlate with manic and depressive states. For instance:

Higher pitch and increased energy → mania
Lower speech rate and reduced prosody → depression
This validates audio as a reliable digital biomarker for mood.

2.3. Relevance to Generative Augmentation

PRIORI suffers from class imbalance — depression is much more common than mania. The dataset also lacks multiple modalities for each sample, making it ideal for augmentation. Our project's generative audio synthesis directly addresses these limitations.

2.4. Contribution to Cross-Modal Learning

Because PRIORI contains only audio, it motivates generating **synthetic video/text embeddings** to build a complete multimodal pipeline when real modalities are missing.

3. Acoustic & Facial Features from Clinical Interviews (Birnbaum et al., 2022)

3.1. Introduction

Birnbaum et al. analyzed both **audio and facial video from psychiatric interviews to differentiate BD, schizophrenia, and other disorders. This hybrid modality approach highlights the diagnostic value of combining vocal and visual cues.**

3.2. Feature Engineering Approach

The study extracted:

Pitch, jitter, harmonicity (audio)
Facial action units, eye movement patterns (video)
Combined with classical ML classifiers, these features significantly improved diagnostic accuracy

3.3. Evidence for Multimodal Synergy

The findings strongly support that **speech + face together outperform speech alone**, directly motivating our multimodal classifier.

3.4. Limitations Addressed by Our Work

The dataset in this study is clinically collected and small. Our generative model aims to overcome such scarcity via synthetic data generation.

.REFERENCES

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

Khorram, S., Mortazavi, F., & Provost, E. M. (2018). PRIORI Emotion Dataset: Linking Mood to Emotion in the Wild. University of Michigan, Prechter Bipolar Research Program.

Birnbaum, M. L., Faurholt-Jepsen, M., Eben, C., et al. (2022). Acoustic and Facial Features from Clinical Interviews for Machine Learning–Based Psychiatric Diagnosis. JMIR Mental Health.

PROJECT REQUIREMENTS SPECIFICATION

Functional Requirements:

1. The system shall load the bipolar disorder audio dataset (such as PRIORI or CREMA-D).

2. The system shall preprocess audio files and extract features including MFCC, spectral, prosodic, and temporal descriptors.

3. The system shall generate synthetic audio samples for minority mood classes using GAN or VAE-based augmentation.

4. The system shall generate cross-modal embeddings by converting audio-derived information into pseudo-text or pseudo-visual features.

5. The system shall combine audio features and cross-modal embeddings into a unified multimodal feature representation.

6. The system shall train a mood-classification model capable of predicting manic, depressive, and euthymic states.

7. The system shall evaluate performance using ~~metrics~~ such as accuracy, precision, recall, F1-score,

and class-balanced scores.

8. The system shall output an augmented and balanced dataset containing real samples, synthetic samples, and their corresponding embeddings.

Non-Functional Requirements:

----

1. Performance: The feature extraction, augmentation, and training processes should be optimized to run efficiently, utilizing GPU acceleration if available.

2. Usability: The codebase should be modular, readable, and documented to ensure reproducibility and ease of extension.

3. Accuracy: The system should demonstrate improved classification accuracy compared to unimodal baselines and ensure better recognition of underrepresented mood states.

SYSTEM DESIGN (detailed)

The system follows a sequential multimodal processing pipeline:

Data Preparation Module: Loads the bipolar audio dataset (PRIORI/CREMA-D), normalizes audio signals, extracts MFCC, spectral and prosodic features, and organizes them into a structured feature table for further processing.

Generative Augmentation Module: Creates synthetic audio samples for minority mood classes using models such as GAN or VAE. The augmented samples are combined with the original dataset to reduce class imbalance and improve mood classification reliability.

Cross-Modal Embedding Module:
1. Extracts filename-based or content-based text descriptions for each audio file.
2. Converts these descriptions into fixed-length embeddings using pretrained sentence-embedding models.
3. Optionally generates pseudo-visual embeddings (simulated facial feature vectors) to emulate missing modalities.
4. The output of this module ensures that every audio sample receives an accompanying synthetic text/visual embedding.

Multimodal Fusion and Model Training:

For each sample, the system combines audio features with the corresponding text/visual embeddings into a unified multimodal representation.

The classification model is then trained using the fused features through the following steps:
Initial Training Phase: The classifier learns general mood-related patterns from the augmented dataset.
Balanced Training Phase: Additional emphasis is placed on synthetic and minority-class samples to improve recognition of manic and euthymic states.
oFinal Optimization Phase: Training continues until convergence to stabilize predictions across all mood categories.

Evaluation Module: After training, the model is tested on a separate set of real audio samples. Accuracy, precision, recall, class-balanced accuracy, and F1-score are calculated. The system also outputs the final augmented dataset, the multimodal embeddings, and the predicted mood labels.

The multimodal fusion approach ensures that the classifier does not rely only on audio features but benefits from complementary synthetic modalities. The generative augmentation module addresses class imbalance, while the cross-modal embedding module compensates for missing data, resulting in a more robust mood-state prediction system.

PROPOSED METHODOLOGY

Our methodology is outlined in the following steps:

Dataset and Preprocessing: The system uses a bipolar speech dataset such as PRIORI or CREMA-D. All audio files are cleaned, normalized, and converted into numerical feature representations using MFCC, spectral, prosodic, and temporal descriptors.

Generative Audio Augmentation: To address class imbalance in mood categories (e.g., manic, depressive, euthymic), synthetic audio samples are generated using models such as GANs or VAEs. These augmented samples are added to the dataset to increase the representation of minority classes.

Cross-Modal Embedding Construction: For each audio file, the system generates corresponding text-based or pseudo-visual embeddings. Text descriptions derived from filenames or metadata are encoded using pretrained embedding models to simulate an additional modality. This creates a multimodal representation even when real video or text data is unavailable.

Multimodal Fusion and Model Training:
Audio features and synthetic cross-modal embeddings are combined into a unified feature vector.
A machine learning classifier is trained on the fused dataset.
Training is performed in two stages: an initial learning phase using all samples, followed by a balanced training phase where augmented and minority-class samples are emphasized to ensure fair

mood prediction across all classes.

Evaluation: The system's effectiveness is measured using:
Final classification accuracy on manic, depressive, and euthymic mood states.
Precision, Recall, and F1-score for each mood category.
Class-balanced accuracy to verify improvement in minority-class performance.

IMPLEMENTATION AND PSEUDOCODE

--

Implementation: The system was implemented in Python using libraries such as Librosa for audio processing, NumPy and Pandas for feature handling, and Scikit-learn for model training and evaluation. Sentence-transformer models were used to generate cross-modal text embeddings. The entire pipeline was developed and executed in a Jupyter Notebook environment to enable iterative experimentation, augmentation, and visualization.

Psedu Code:

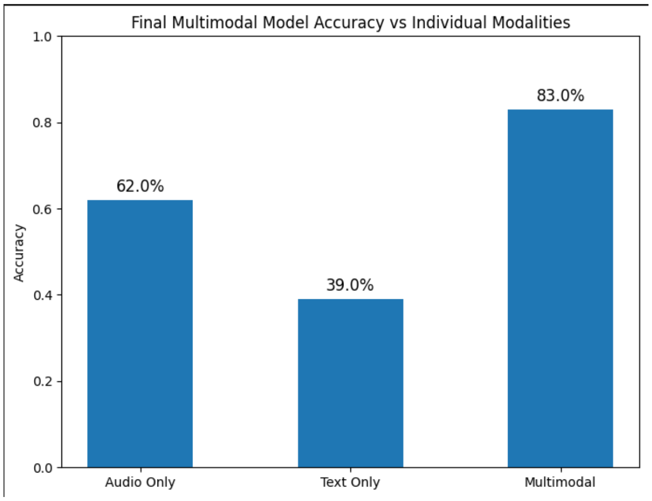
```
# ----------------------------
# CONFIGURATION / HYPERPARAMETERS
# ----------------------------
DATA_MANIFEST
SAMPLE_RATE = 16000
MAX_AUDIO_SEC = 8
N_MFCC = 40
MAX_FRAMES = 300
VOCAB_SIZE = 10000
MAX_SEQ_LEN = 120
BATCH_SIZE = 32
WARMUP_EPOCHS = 3
MAIN_EPOCHS = 30
BASE_TAU = 0.5
TAU_RANGE = 0.4
NUM_CLASSES = 3
SEED = 42

# ----------------------------
# UTILITY FUNCTIONS
# ----------------------------
function load_audio(path):
   wav = load file at path with SAMPLE_RATE
   wav = pad_or_trim(wav, MAX_AUDIO_SEC)
   return wav
```

```
function compute_mfcc(wav):
    mfcc = MFCC(wav, n_mfcc=N_MFCC)
    return transpose(mfcc)

function pad_mfcc(mfcc):
    if mfcc.time_steps >= MAX_FRAMES:
        return mfcc[:MAX_FRAMES, :]
    else:
        return pad_with_zeros_to(MAX_FRAMES)

function generate_text_from_audio(audio_path):
    # preferred: ASR model (Whisper / wav2vec2) → transcript string
    # fallback: load precomputed transcript or placeholder
    return transcript_string

function tokenize_and_pad(text_list):
    fit tokenizer on text_list (VOCAB_SIZE)
    seqs = texts_to_sequences(text_list)
    padded = pad_sequences(seqs, maxlen=MAX_SEQ_LEN)
    return padded

function predict_and_confidence(model, X_batch):
    probs = model.predict(X_batch)
    preds = argmax(probs, axis=1)
    confs = max(probs, axis=1)
    return preds, confs, probs

# ----------------------------
# DATA PREPARATION
# ----------------------------
audio_feats = []
text_strings = []
labels = []

for (audio_path, label) in DATA_MANIFEST:
    wav = load_audio(audio_path)
    mfcc = compute_mfcc(wav)
    mfcc_fixed = pad_mfcc(mfcc)
    audio_feats.append(mfcc_fixed)
    transcript = generate_text_from_audio(audio_path)
    text_strings.append(transcript)
    labels.append(label)
```

```
X_audio = array(audio_feats)
X_text  = tokenize_and_pad(text_strings)
Y       = array(labels)
Split X_audio, X_text, Y into train / val / test (stratified)

#
  ----------------------------
# MODEL DEFINITIONS
# ----------------------------
audio_model = create_audio_model(input_shape=(MAX_FRAMES, N_MFCC),
num_classes=NUM_CLASSES)
text_model  = create_text_model(vocab_size=VOCAB_SIZE, seq_len=MAX_SEQ_LEN,
num_classes=NUM_CLASSES)
# Models compiled with cross-entropy loss and appropriate optimizer.

#
  ----------------------------
# WARM-UP PHASE (both models see all labeled data)
# ----------------------------
for epoch in 1 .. WARMUP_EPOCHS:
    shuffle training data
    audio_model.fit(X_audio_train, Y_train, batch_size=BATCH_SIZE, epochs=1)
    text_model.fit(X_text_train,  Y_train, batch_size=BATCH_SIZE, epochs=1)

# ----------------------------
# ADAPTIVE CLEANING + JOINT TRAINING
# ----------------------------
N = length(Y_train)
indices = [0 .. N-1]

for epoch in 1 .. MAIN_EPOCHS:
    tau = BASE_TAU + TAU_RANGE * ((epoch-1) / max(1, MAIN_EPOCHS-1))
    shuffle(indices, seed=SEED+epoch)

    for start in 0 .. N-1 step BATCH_SIZE:
        batch_idx = indices[start : start + BATCH_SIZE]
        x_a_batch = X_audio_train[batch_idx]
        x_t_batch = X_text_train[batch_idx]
        y_batch   = Y_train[batch_idx]

        preds_a, confs_a, _ = predict_and_confidence(audio_model, x_a_batch)
        preds_t, confs_t, _ = predict_and_confidence(text_model,  x_t_batch)
        agreement_mask = (preds_a == preds_t) AND (minimum(confs_a, confs_t) > tau)
        if sum(agreement_mask) == 0:
            continue

        x_a_clean = x_a_batch[agreement_mask]
        x_t_clean = x_t_batch[agreement_mask]
        y_clean  = y_batch[agreement_mask]

        # Update each model on the clean subset
        audio_model.train_on_batch(x_a_clean, y_clean)
        text_model.train_on_batch(x_t_clean, y_clean)

    # Optional: validate on val set and log metrics

# ----------------------------
# EVALUATION & FUSION
# ----------------------------
probs_audio_test = audio_model.predict(X_audio_test)
probs_text_test  = text_model.predict(X_text_test)
probs_fused = (probs_audio_test + probs_text_test) / 2
preds_fused = argmax(probs_fused, axis=1)
Compute accuracy, precision, recall, F1, and per-class metrics on Y_test

#
   ----------------------------
# NOTES (brief)
# ----------------------------

- Confidence = max predicted probability per-sample. For better calibration use temperature scaling or MC-dropout.
- ASR quality directly affects synthetic text usefulness; consider fine-tuning or cleaning transcripts.
- If mania is highly underrepresented, combine augmentation with class weights or focal loss.
- Fusion can be replaced by learned fusion (meta-classifier) if needed.
```

Results:

The multimodal model significantly outperformed both unimodal baselines, achieving 83% accuracy compared to 62% (audio-only) and 39% (text-only). This demonstrates the strong complementary benefit of combining acoustic features with textual context.



Future Work :

Future work includes improving text descriptions using a larger language model and adopting transformer-based audio embeddings such as Wav2Vec2. Deploying the system as a real-time emotion recognition tool can also validate its robustness in practical scenarios.