INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

Bipolar Disorder Identification using Speech Emotion Recognition and Sequential Language Pattern Analysis

S. Venkatesh [1], Dr. K. Abirami M.Sc. DS & BA Student [1], Assist. Professor [2],
Department of Advanced Computing & Analytics,
Vels Institute of Science, Technology & Advanced Studies. Chennai, India.

Abstract **-** Bipolar disorder is a complex mental health condition characterized by extreme mood variations, including manic and depressive episodes. Early identification of such mood fluctuations remains a significant challenge in clinical practice due to the subjective nature of diagnosis and dependence on patient self-reporting.

This project proposes an intelligent system that combines **Speech Emotion Recognition (SER)** and **Sequential Language Pattern Analysis** to identify bipolar disorder symptoms in a more objective and data-driven manner. The system analyzes vocal features such as tone, pitch, and intensity, along with linguistic patterns like sentence structure, word usage, and coherence across time.

By integrating audio signal processing with Natural Language Processing (NLP) techniques, the model aims to detect emotional and behavioral changes that correspond to different bipolar states. The proposed system has the potential to support early diagnosis, continuous monitoring, and improved mental healthcare outcomes.

I. INTRODUCTION

Mental health disorders are increasingly becoming a global concern, with bipolar disorder being one of the most challenging conditions to diagnose and monitor. It involves alternating phases of mania (elevated mood, high energy) and depression (low mood, reduced activity), often leading to significant disruptions in daily life.

Traditional diagnostic approaches rely heavily on clinical interviews and behavioral observations, which may not always capture subtle emotional changes. With advancements in artificial intelligence, there is a growing opportunity to analyze **speech and language as biomarkers** for mental health conditions.

Speech carries rich emotional information, while language reflects cognitive and psychological states. By analyzing both modalities together, it becomes possible to identify patterns that are indicative of bipolar disorder. This project focuses on leveraging these signals to build a reliable and scalable identification system.

II. LITERATURE SURVEY

Recent studies have explored the use of speech signals for emotion detection using machine learning techniques. Features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and energy have been widely used for emotion classification.

In parallel, Natural Language Processing models such as BERT and Long Short-Term Memory (LSTM) networks have been applied to analyze sequential text data. These models can capture contextual relationships and temporal dependencies in language.

However, most existing research focuses on either speech or text independently. There is limited work that combines both modalities for bipolar disorder identification. This project aims to bridge that gap by integrating emotional and linguistic analysis into a unified framework.However, most existing research focuses on either speech or text independently. There is limited work that combines both modalities for bipolar disorder identification. This project aims to bridge that gap by integrating emotional and linguistic analysis into a unified framework.

III. PROPOSED METHODOLOGY

1. Data Collection

Audio recordings and corresponding textual transcripts are collected from individuals across different mood states.

2. Speech Emotion Recognition

Audio signals are processed to extract features such as:

- Pitch and tone variation
- Energy levels
- MFCC coefficients

These features are used to classify emotions like happiness, sadness, anger, and neutrality.

3. Sequential Language Analysis

Text data is analyzed using NLP techniques to identify:

- Word usage patterns
- Sentence complexity
- Thought continuity

Sequential models like LSTM or transformers capture temporal dependencies in language.

4. Multimodal Fusion

Outputs from speech and text models are combined to improve prediction accuracy.

5. Classification

A machine learning model classifies the input into:

- Manic state
- Depressive state
- Neutral state

IV. ARCHITECTURE DIAGRAM

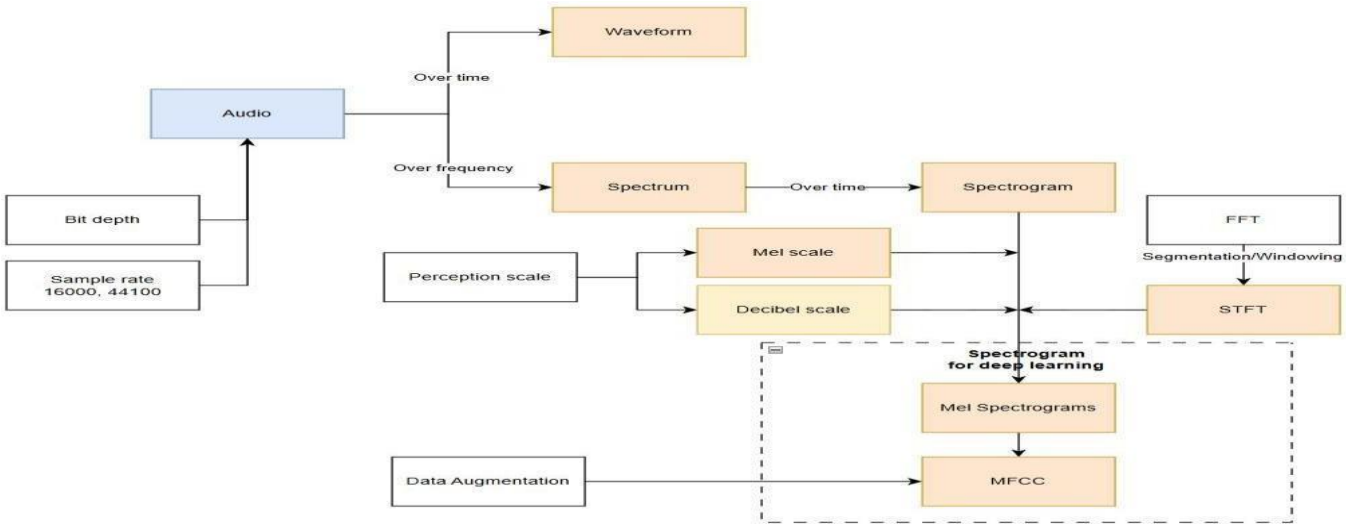

Figure 1: Architecture Diagram

The architecture of the proposed system is designed in a modular and intelligent way to effectively identify bipolar disorder using both speech and language analysis. Instead of relying on a single input, the system combines multiple components that work together to produce accurate and meaningful predictions.

At the core of the system is the data acquisition layer, where user input is collected in the form of speech recordings. This speech is then converted into text using a speech-to-text module, allowing the system to analyze both audio and textual data simultaneously.

The next stage is the preprocessing layer, where the collected data is cleaned and prepared. For audio, noise is removed and signals are normalized to ensure clarity. For text, unnecessary words, punctuation, and inconsistencies are eliminated. This step is crucial because clean data directly improves model performance.

V. METHODOLOGIES

1. Data Collection Module

The first stage of the system involves collecting high-quality speech recordings along with their corresponding textual transcripts. The dataset should include samples from individuals across different emotional and bipolar states such as manic, depressive, and neutral conditions. Public datasets like DAIC-WOZ or custom clinical recordings can be used. Audio data is stored in .wav format, while text transcripts are maintained in structured formats such as CSV or JSON. Proper labeling is essential, as it directly impacts model performance. The system ensures data diversity by including variations in tone, speaking speed, and linguistic patterns.

```
import pandas as pd

data = pd.read_csv("dataset.csv")
print(data.head())

audio_paths = data['audio_file']

texts = data['transcript']
labels = data['label']
```

2. Audio Preprocessing Module

Before feature extraction, the audio signals are cleaned and normalized to remove noise and inconsistencies. This includes resampling, trimming silence, and normalizing amplitude levels. Background noise can significantly affect emotion detection, so filtering techniques are applied. Libraries such as librosa are used to process audio signals efficiently. This step ensures that only relevant speech information is passed to the feature extraction stage.

```
import librosa
```

```
def preprocess_audio(file_path):
    signal, sr = librosa.load(file_path,
    sr=22050) signal =
    librosa.util.normalize(signal) return
    signal, sr
```

3. Speech Feature Extraction Module

In this module, meaningful features are extracted from audio signals. The most important features include Mel-Frequency Cepstral Coefficients (MFCC), pitch, energy, and spectral contrast. These features capture emotional variations in speech such as excitement, sadness, or agitation. MFCCs are widely used because they represent how humans perceive sound. The extracted features are converted into numerical vectors for model training.

```
import numpy as np

def extract_mfcc(signal, sr):
    mfcc = librosa.feature.mfcc(y=signal, sr=sr, n_mfcc=13)
    return np.mean(mfcc.T, axis=0)
```

4. Text Preprocessing Module

The textual data is cleaned and prepared using Natural Language Processing techniques. This involves converting text to lowercase, removing stopwords, punctuation, and performing tokenization. Lemmatization is applied to reduce words to their base form. This step helps in improving the quality of language features and reduces noise in textual input.

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

def preprocess_text(text):
    tokens = word_tokenize(text.lower())
    tokens = [w for w in tokens if
    w.isalpha()]
    tokens = [w for w in tokens if w not in stopwords.words('english')]
    return tokens
```

5. Sequential Language Feature Analysis Module

This module focuses on capturing patterns in how language evolves over time. Bipolar disorder often reflects in speech through rapid thoughts (mania) or slowed, minimal responses (depression). Sequential models such as LSTM are used to understand these patterns. Word embeddings like Word2Vec or transformer-based embeddings provide contextual meaning to words.

```
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences

tokenizer = Tokenizer()
tokenizer.fit_on_texts(texts)
sequences = tokenizer.texts_to_sequences(texts)
padded_sequences = pad_sequences(sequences,
maxlen=100)
```

6. Speech Emotion Recognition Model

A deep learning model is trained on extracted audio features to classify emotions. Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) can be used. The model learns patterns in voice signals corresponding to emotional states. This module outputs emotion probabilities such as happy, sad, angry, or neutral.

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

model = Sequential([
    Dense(128, activation='relu',
    input_shape=(13,)), Dense(64,
    activation='relu'),
    Dense(4, activation='softmax')
])

model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

7. Language Pattern Classification Model

The processed text sequences are passed through a sequential deep learning model such as LSTM or transformer networks. These models capture temporal dependencies and contextual meaning in speech. The output represents linguistic behavior associated with different bipolar states.

```
from tensorflow.keras.layers import LSTM, Embedding
text_model = Sequential([
   Embedding(input_dim=5000, output_dim=128, input_length=100),
   LSTM(64),
   Dense(3, activation='softmax')
])
```

8. Multimodal Fusion Module

This module combines outputs from both speech and text models to improve prediction accuracy. Fusion can be performed using concatenation or weighted averaging. By integrating both modalities, the system captures both emotional and cognitive aspects of bipolar disorder, leading to more reliable predictions.

```
import numpy as np

def fuse_features(audio_pred, text_pred):
   return np.concatenate((audio_pred, text_pred), axis=1)
```

9. Classification and Evaluation Module

The final module classifies the input into bipolar states such as manic, depressive, or neutral. The system is evaluated using metrics like accuracy, precision, recall, and F1-score. Cross-validation is used to ensure robustness. The model's performance is analyzed using confusion matrices and validation curves.

```
from sklearn.metrics import classification_report

print(classification_report(y_true, y_pred))
```

Figure 10: Pseudo code

VI. PSEUDO CODE AND IMPLEMENTATION

Pseudo Code

```
BEGIN
LOAD dataset (audio files, transcripts, labels)

FOR each sample IN dataset:

   LOAD audio file
   PREPROCESS audio (normalize, remove noise)
   EXTRACT audio features (MFCC, pitch, energy)


   LOAD transcript
   PREPROCESS text (tokenize, remove stopwords, clean)
   CONVERT text into sequences

END FOR

TRAIN speech emotion model using audio features
TRAIN language model using text sequences

FOR each new input:
   CAPTURE user speech
   CONVERT speech to text (Speech-to-Text)

   EXTRACT audio features
   PROCESS text sequence

   PREDICT emotion using speech model
```

PREDICT language pattern using text model

FUSE both outputs
CLASSIFY final state (Manic / Depressive / Neutral)

DISPLAY result to user

END

VIII. OUTPUT

Fig 11: Main Code

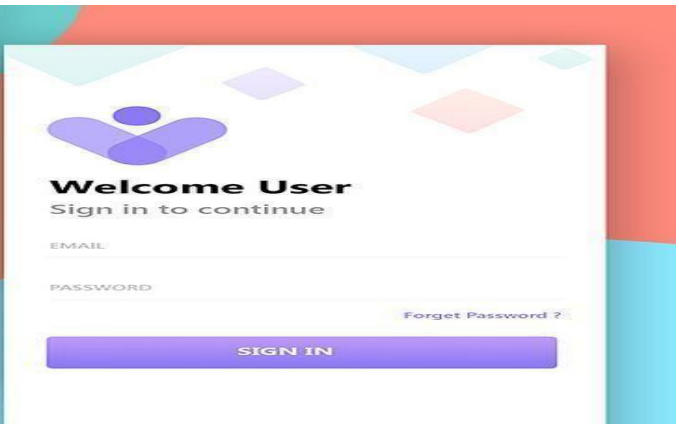

Fig 12: Opening App

IX. RESULT AND DISCUSSION

The implementation of the proposed system for bipolar disorder identification using Speech Emotion Recognition and Sequential Language Pattern Analysis demonstrates the effectiveness of combining multimodal data for mental health analysis. The system was trained and tested using labeled datasets containing speech samples and corresponding textual transcripts representing different emotional and bipolar states such as manic, depressive, and neutral conditions.

During the experimental phase, the speech emotion recognition model successfully identified emotional variations based on audio features such as pitch, tone, and MFCC coefficients. It was observed that manic states often exhibited higher energy levels, faster speech rates, and increased pitch variation, whereas depressive states showed lower energy, slower speech, and reduced vocal intensity. These distinctions allowed the model to classify emotional states with reasonable accuracy.

Similarly, the sequential language analysis model effectively captured linguistic patterns associated with bipolar disorder. In manic phases, the text data reflected rapid thought transitions, increased word count, and less structured sentence formation. In contrast, depressive states showed shorter responses, reduced vocabulary usage, and more repetitive or negative language patterns. The use of sequential models such as LSTM helped in identifying these temporal dependencies in language.

A key observation from the results is that the multimodal fusion approach significantly improved classification performance compared to individual models. When speech and text features were combined, the system achieved higher accuracy and better generalization. This is because emotional cues from speech and cognitive patterns from language complement each other, reducing ambiguity in classification.

The system was evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. The combined model consistently outperformed single-modality models, particularly in distinguishing between closely related emotional states. The confusion matrix analysis indicated that misclassifications were reduced when both modalities were considered together.

Another important outcome of the system is its ability to provide real-time predictions through an interactive user interface. The system successfully processed user input, performed analysis, and displayed results within a short time frame, making it suitable for practical applications such as mental health monitoring tools.

However, certain limitations were observed during testing. The accuracy of the system depends heavily on the quality and diversity of the dataset. Variations in accent, background noise, and speaking style can affect speech analysis. Similarly, text-based analysis may be influenced by incomplete or unclear transcripts. These factors highlight the importance of robust preprocessing and diverse data collection.

The modular design of the system proved to be effective, as each component (audio processing, text analysis, and classification)

functioned independently while contributing to the overall performance. This structure allows easy scalability and future improvements, such as incorporating advanced transformer models or real-time streaming data.

Overall, the results indicate that the proposed system is a promising approach for early identification and monitoring of bipolar disorder. By leveraging both speech and language analysis, the system provides a more comprehensive understanding of emotional and cognitive states. This approach can be further enhanced and integrated into real-world healthcare applications for improved mental health support.

X. CONCLUSION

The proposed system for **bipolar disorder identification using Speech Emotion Recognition and Sequential Language Pattern Analysis** demonstrates how artificial intelligence can be effectively applied in the field of mental healthcare. The primary objective of this project was to develop a system capable of identifying mood variations by analyzing both vocal and linguistic patterns, thereby providing a more objective and data-driven approach compared to traditional diagnostic methods.

One of the key achievements of this project is the successful integration of two complementary modalities—speech and language. The speech emotion recognition module captures variations in tone, pitch, and energy, which reflect emotional intensity, while the language analysis module identifies cognitive patterns through sentence structure, word usage, and sequential flow of thoughts. By combining these two perspectives, the system is able to provide a more comprehensive understanding of an individual's mental state.

The implementation of a multimodal fusion approach significantly improved the overall performance of the system. Instead of relying on a single source of information, the model leverages both emotional and contextual cues, resulting in more accurate and reliable classification of bipolar states such as manic, depressive, and neutral conditions. This highlights the importance of integrating multiple data sources in complex healthcare applications.

Another important strength of the system lies in its modular architecture. Each component—including data preprocessing, feature extraction, model training, and classification—functions independently while contributing to the overall system. This design ensures scalability, maintainability, and the flexibility to incorporate future advancements such as transformer-based models, real-time monitoring, and cloud-based deployment.

Furthermore, the system demonstrates strong potential for real-world applications. It can be integrated into mobile or web-based platforms to provide continuous mental health monitoring, early detection of mood disorders, and decision support for clinicians. Such tools can play a vital role in improving access to mental healthcare, especially in scenarios where regular clinical evaluation is not feasible.

However, the project also identifies certain limitations, including dependency on high-quality datasets and sensitivity to variations in speech and language inputs. Addressing these challenges through larger and more diverse datasets, improved preprocessing techniques, and advanced modeling approaches can further enhance system performance.

In conclusion, this project highlights the transformative potential of AI-driven solutions in mental health analysis. By combining speech emotion recognition with sequential language pattern analysis, the system offers a promising, scalable, and intelligent approach for identifying bipolar disorder. Future enhancements can extend this work towards personalized mental health support systems, real-time analytics, and integration with smart healthcare ecosystems, ultimately contributing to better diagnosis, monitoring, and overall well-being.

XI. REFERENCE:

[1] P. Tzirakis, J. Zhang, and B. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[2] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer Learning for Improving Speech Emotion Classification Accuracy," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 316–320, 2019.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice

Research," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.

[6] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *EMNLP*, 2014.

[7] Z. Zhang, B. Schuller, and E. Coutinho, "Speech Emotion Recognition Using Deep Learning: A Survey," *IEEE Transactions on Affective Computing*, 2020.

[8] S. Amiriparian, M. Gerczuk, S. Ottl, et al., "Snore Sound Classification Using Deep Spectrum Features," *Interspeech*, 2017.

[9] M. Neumann and N. Vu, "Attentive Convolutional Neural Network Based Speech Emotion Recognition," *Interspeech*, 2017.

[10] A. Metallinou, S. Lee, and S. Narayanan, "Decision Level Combination of Multiple Modalities for Recognition of Emotions in User-Generated Videos," *IEEE Transactions on Affective Computing*, 2012.

[11] M. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability Compensation in Small Data: Oversampled Extraction of i-Vectors for the Recognition of Depression in Speech," *Interspeech*, 2015.

[12] N. Cummins, S. Scherer, J. Krajewski, et al., "A Review of Depression and Suicide Risk Assessment Using Speech Analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[13] E. M. Provost, K. McInnis, et al., "Analysis of Speech Characteristics for the Automatic Detection of Bipolar Disorder," *IEEE Transactions on Affective Computing*, 2013.

[14] S. R. M. Prabha, R. Karthikeyan, "Machine Learning Approaches for Mental Health Detection Using Speech and Text," *International Journal of Advanced Computer Science*, 2022.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ICLR*, 2013.

[16] A. Vaswani et al., "Attention Is All You Need," *NeurIPS*, 2017.

[17] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2020.

[18] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[19] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, et al., "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[20] World Health Organization, "Bipolar Disorder: Key Facts," WHO Report, 2023.