Chhattisgarh Swami Vivekanand Technical University

University Teaching Department

Department of Computer Science and Engineering

Presented By

Sejal Sahu

300012722011

CSE-AI

Ashutosh Roy

300012722035

CSE-AI

Minor Project Presentation On

Multimodal Voice Based Emotion and Stress Detection Using Deep Learning for Mental Health Monitoring

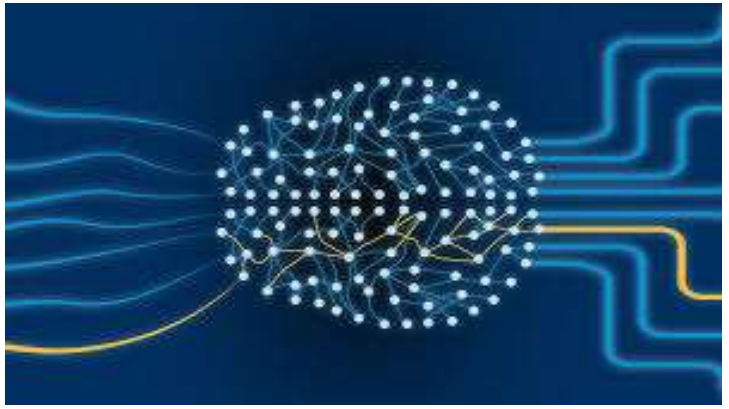
Dr. Toran Verma

Associate Professor, Department of CSE

Content

Introduction

- **Mental Health Disorders**
  Stress, anxiety, and depression are increasingly prevalent in modern society, affecting overall well-being and productivity.
- **Early Detection**
  Early identification of emotional and stress-related conditions is crucial for timely mental health monitoring and effective intervention.
- **Speech as an Information Source**
  Human speech carries rich emotional cues—such as **pitch, tone, intensity, and speaking rate**—which vary significantly with emotional and stress states.
- **Deep Learning Evolution**
  Speech analysis has evolved from **handcrafted features (MFCCs)** to **self-supervised deep learning models like Wav2Vec2**, enabling more robust and generalized emotion recognition directly from raw audio.



Problem Statement

- The problem is the lack of robust and generalized systems for early detection of emotion and stress from speech, especially across diverse speakers and environments.

- **Rising Mental Health Issues** - Stress, anxiety, and depression are increasingly prevalent in modern society.

- **Need for Early Detection** - Early identification is crucial for effective mental health monitoring and timely intervention.

- **Limitations of Existing Methods** - Traditional approaches using handcrafted features (e.g., MFCCs) fail to generalize across speakers and environments.

- **Research Challenge** - Develop a robust, automated deep learning system to detect emotion and stress from speech signals.

Research Objectives

- Design deep learning system for automatic emotion & stress detection from speech signals
- Utilize self-supervised Wav2Vec2 for high-level acoustic feature extraction
- Model temporal speech dynamics using Bi-directional LSTM network
- Implement multi-task learning for joint emotion classification & stress regression
- Integrate multiple open-source emotional speech datasets under unified taxonomy
- Evaluate performance using standard metrics (accuracy, RMSE)

Literature Survey

Study / Paper	Method / Approach	Features / Models Used	Dataset(s)	Key Findings	Limitation / Gap
Swain et al., 2018 (Survey)	Review of traditional SER	ML & early Deep Learning	Various	Comprehensive overview of features & classifiers in speech emotion recognition.	Focuses mostly on **feature-based ML**, not modern self-supervised methods. (ResearchGate)
Pepino et al., 2021	Wav2Vec2.0 embeddings for SER	Self-supervised Wav2Vec2 + NN	RAVDESS, IEMOCAP	Using Wav2Vec2 features improves performance over traditional features.	Only **emotion classification**, not stress prediction. (ResearchGate)
Wang et al., 2025	Fine-tuned Wav2Vec2 + NCDE classifier	Transformer + differential equations	IEMOCAP	New architecture for emotion recognition with stability & fast convergence.	Complex model, not widely tested on multiple corpora. (PMC)
Madanian, 2023	Survey on SER ML methods	Traditional + deep models	Multiple	Identifies challenges of speaker-independent accuracy & evaluation.	Doesn't use modern transformers like Wav2Vec2. (ScienceDirect)

Literature Survey

Author / Year	Method	Features / Models	Datasets	Key Findings	Limitations
Swain et al., 2018	Survey	ML & Deep Learning	Multiple	Reviewed classifiers & features.	Focus on traditional features only. (ResearchGate)
Pepino et al., 2021	Wav2Vec2 Embedding	Self-supervised features	RAVDESS, IEMOCAP	Wav2Vec2 improves emotion recognition.	Only emotion classification. (ResearchGate)
Wang et al., 2025	Wav2Vec2 + NCDE	Transformer + Differential Eq	IEMOCAP	Stable, accurate emotion recognition.	Only one dataset. (PMC)
Nguyen et al., 2026	Feature Fusion	Transfer + Prosody	Multi	Fusion enhances generalization.	Doesn't include stress regression. (ScienceDirect)

Research Methodology

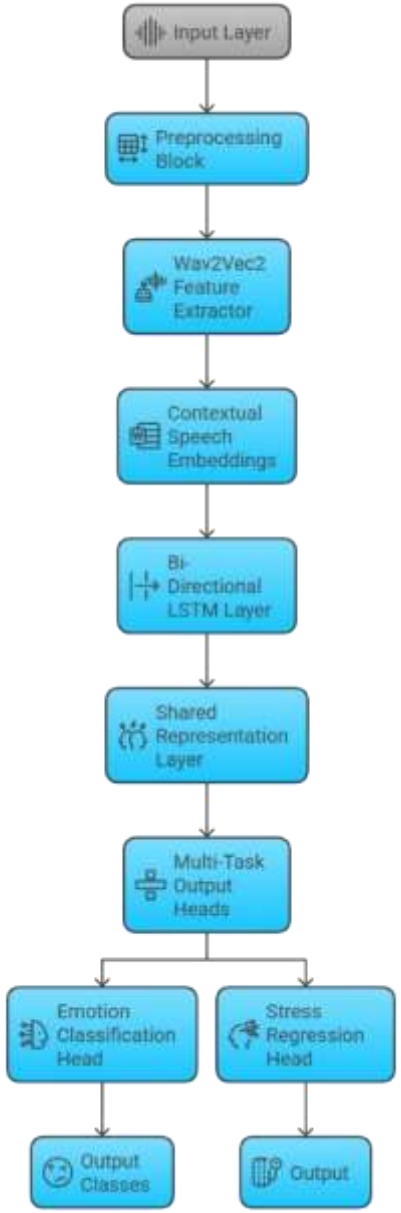
1. **Input Speech**
   Raw audio waveform

2. **Data Preprocessing**
   Resampling to 16 kHz, normalization

3. **Feature Extraction (Wav2Vec2)**
   Self-supervised speech representations, frozen pretrained layers

4. **Temporal Modeling (Bi-LSTM)**
   Captures bidirectional temporal speech patterns

5. **Multi-Task Output**
   a. **Emotion Classification:** 5 emotion classes
   b. **Stress Regression:** Continuous stress values



Result/Discussions

- **Angry emotion achieves the highest F1-score (0.96)**
- **Fear shows the highest recall (0.80)**, indicating strong detection of high-arousal emotions.

•Achieved **85.9% overall emotion classification accuracy**

• **5 emotion classes**: Neutral, Happy, Sad, Angry, Fear

•**Stress prediction performance**:

a. **MSE:** 0.0826
b. **RMSE:** 0.1268 (low error, good regression accuracy)

Green = Best performance
Blue = Strong / Second-best

Emotion Class	Precision (%)	Recall (%)	F1-Score (%)
Neutral	0.71	0.03	**0.83**
Happy	**0.94**	0.68	0.79
Sad	**0.94**	**0.74**	**0.83**
Angry	**0.96**	0.70	**0.96**
Fear	0.87	**0.80**	0.76

Result/Discussions

•High values along the diagonal indicate accurate emotion classification

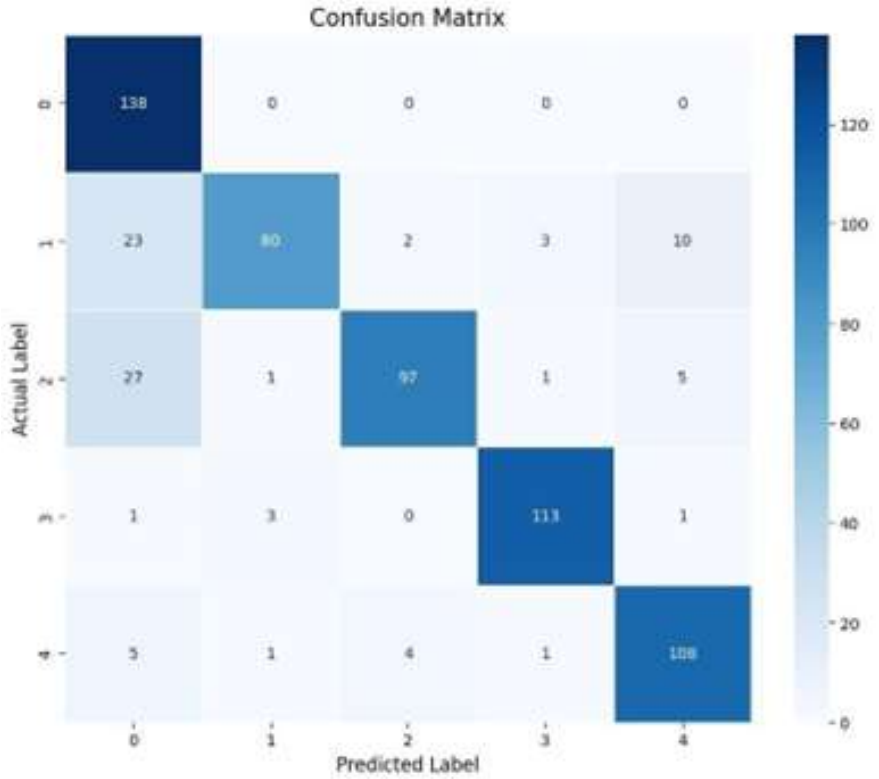
•Angry and Neutral emotions show the highest correct predictions

•Minor confusion observed between similar emotions (Neutral–Happy, Sad–Fear)

•High-arousal emotions are classified more accurately

•No major class imbalance or bias observed

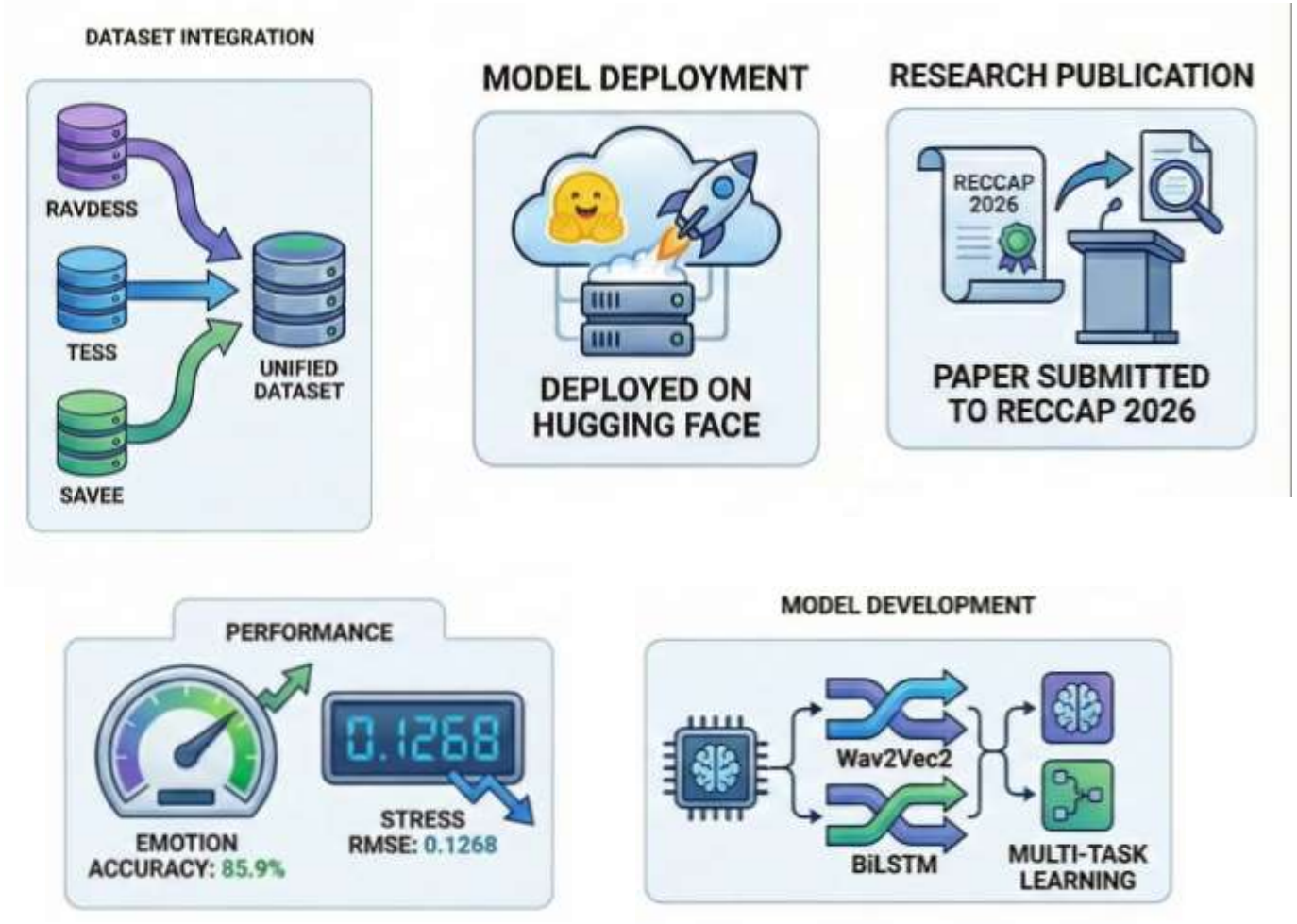
•Confirms robust and stable model performance



Scope of Project

- Emotion classification into five emotions
- Stress regression based on emotional arousal
- Multi-dataset generalization using RAVDESS, TESS, and SAVEE
- End-to-end deep learning pipeline from raw speech input
- Model suitable for mental health monitoring applications
- Integration with **doctor's clinics and hospitals**
- Support for **mental health assessment tools**
- Deployable as **standalone modules** or integrated with existing healthcare systems

Progress so Far



- **Dataset Integration**

Unified RAVDESS, TESS, and SAVEE datasets

- **Model Development**

Hybrid Wav2Vec2–BiLSTM with multi-task learning

- **Performance**

Emotion Accuracy: **85.9%**

Stress RMSE: **0.1268**

- **Model Deployment**

Deployed on Hugging Face for reproducibility

- **Research Publication**

Paper submitted to **RECCAP 2026**

References

- **[1]** A. Baevski et al., *wav2vec 2.0: A framework for self-supervised learning of speech representations*, NeurIPS, 2020.
- **[2]** S. Schneider et al., *wav2vec 2.0*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
- **[3]** S. Hochreiter & J. Schmidhuber, *Long Short-Term Memory*, Neural Computation, 1997.
- **[4]** B. Schuller & A. Batliner, *Speech Emotion Recognition: Two Decades in a Nutshell*, Communications of the ACM, 2018.
- **[5]** S. R. Livingstone & F. A. Russo, *RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song*, PLOS ONE, 2018.
- **[6]** K. Dupuis & M. K. Pichora-Fuller, *Toronto Emotional Speech Set (TESS)*, 2010.
- **[7]** **S. Sahu**, *Hybrid Wav2Vec2–BiLSTM Model for Emotion and Stress Detection from Speech*, Hugging Face Model Hub
- **[8]** Y. LeCun, Y. Bengio, and G. Hinton, *Deep Learning*, Nature, 2015.

Thank you