Multimodal Depression Detection: Using Audio and Video Analysis with FastAPI

MADHU SRI DOLE 1005-21-729-020 AIML , SHIVALIKA MAREDDY 1005-21-729-036 AIML

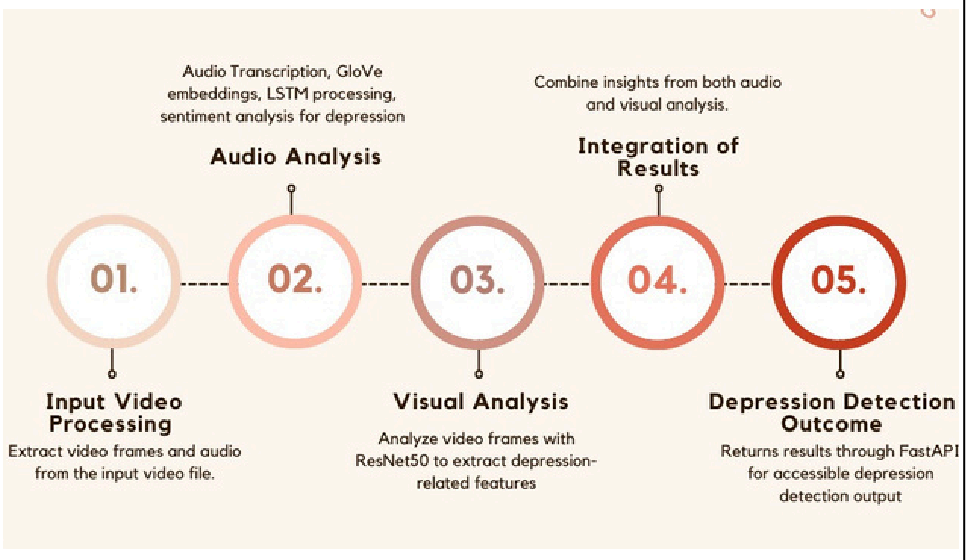
INTRODUCTION

As the title suggests, this project focuses on developing a backend application that operates seamlessly in the background to detect and analyze emotions from video inputs. Utilizing advanced machine learning and deep learning techniques, our application integrates audio and visual cues to provide a comprehensive analysis of emotional states. By converting video data into actionable insights, we aim to enhance the accuracy and efficiency of depression detection. Depression is a prevalent mental health condition that often goes undetected. Traditional assessment methods rely heavily on self-reported symptoms, which can lead to delayed treatment. Advancements in machine learning and multimodal analysis present new opportunities for improving depression detection. This work integrates audio and video analysis to create a more comprehensive detection system

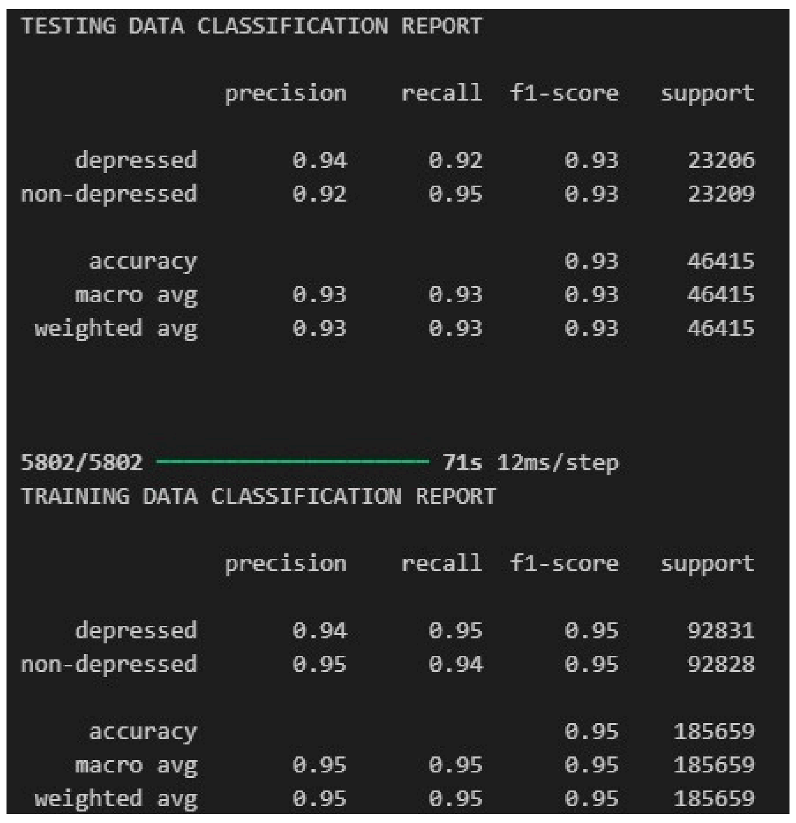
DISCUSSION

Our approach utilizes advanced machine learning techniques, including Whisper for audio transcription, GloVe for word embedding, LSTM for sequence modeling, and a Custom CNN for visual feature extraction. By combining these methods, we aim to improve the sensitivity and specificity of depression detection. The FastAPI framework facilitates real-time data processing and model deployment, allowing for user-friendly interaction.

FLOW-CHART


Audio Transcription, GloVe embeddings, LSTM processing, sentiment analysis for depression
Audio Analysis
Combine insights from both audio and visual analysis.
Integration of Results
01.
02.
03.
04.
05.
Input Video Processing
Extract video frames and audio from the input video file.
Visual Analysis
Analyze video frames with ResNet50 to extract depression-related features
Depression Detection Outcome
Returns results through FastAPI for accessible depression detection output

CLASSIFICATION REPORT



```
TESTING DATA CLASSIFICATION REPORT

              precision    recall  f1-score   support

    depressed       0.94      0.92      0.93     23206
non-depressed       0.92      0.95      0.93     23209

     accuracy                           0.93     46415
    macro avg       0.93      0.93      0.93     46415
 weighted avg       0.93      0.93      0.93     46415



5802/5802 ━━━━━━━━━━━━━━━━━━━━ 71s 12ms/step
TRAINING DATA CLASSIFICATION REPORT

              precision    recall  f1-score   support

    depressed       0.94      0.95      0.95     92831
non-depressed       0.95      0.94      0.95     92828

     accuracy                           0.95    185659
    macro avg       0.95      0.95      0.95    185659
 weighted avg       0.95      0.95      0.95    185659
```

CONCLUSION:

This project presents a robust, multimodal approach to depression detection by leveraging both audio and visual cues through advanced machine learning techniques. Integrating Whisper for audio transcription, GloVe embeddings, LSTM for temporal analysis, and a custom CNN for facial expression recognition, the model offers a more comprehensive assessment of emotional states linked to depression. The FastAPI framework ensures efficient, real-time processing, making the system practical for clinical and non-clinical applications. By enhancing detection accuracy, this application has the potential to facilitate timely interventions and support mental health professionals in monitoring and addressing depressive symptoms effectively.

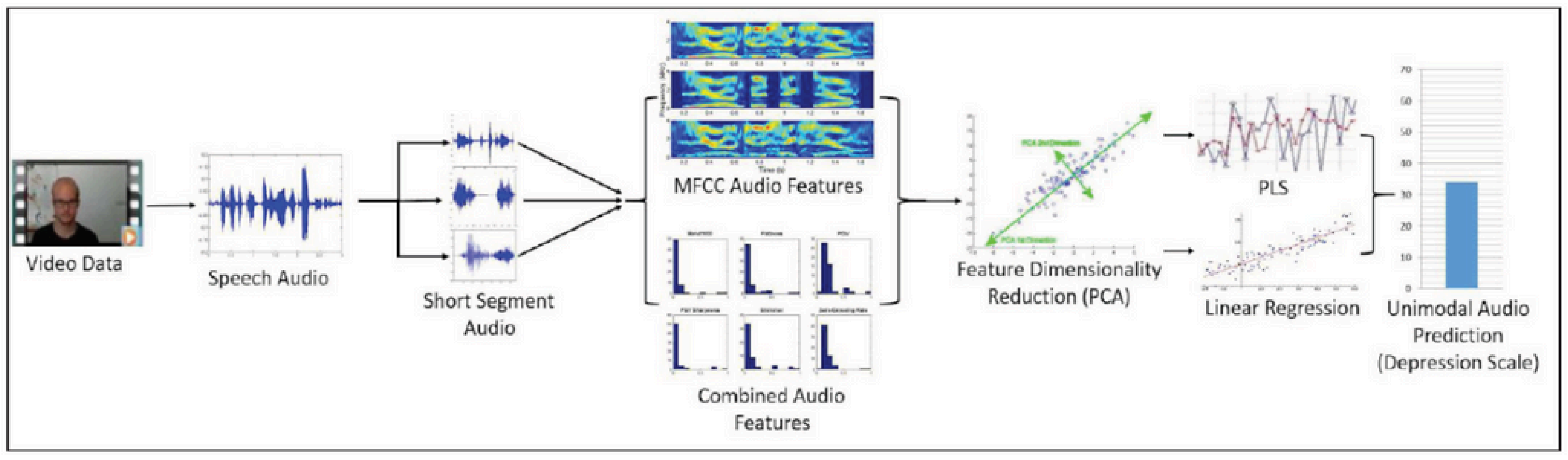
HIGHLIGHTS OF THE WORK

Multimodal Emotion Detection: Combines audio and visual data to capture a fuller spectrum of emotional cues, enabling a more robust depression detection framework.

Audio Processing: Utilizes Whisper for accurate audio transcription, turning speech into text for further analysis.

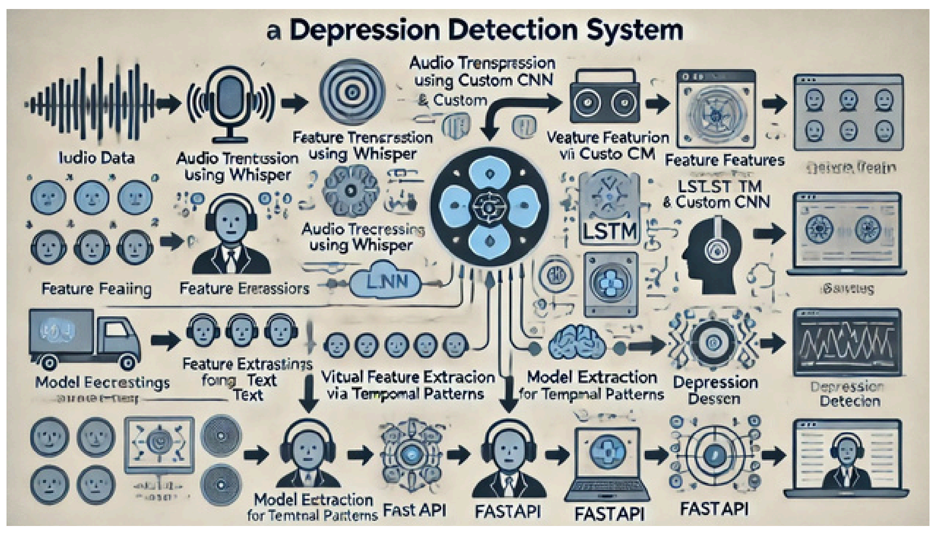
Visual Feature Extraction: Employs a custom CNN to detect facial expressions and visual cues that correlate with emotional states, enriching the model's sensitivity to non-verbal signals.

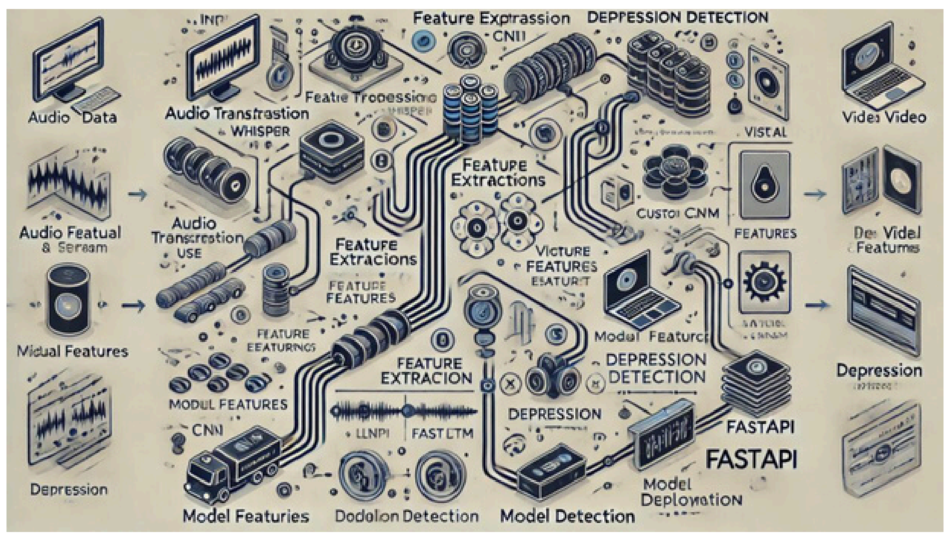
Real-Time Performance: The background processing capability ensures that analysis happens without disrupting user interactions, allowing for real-time or near-real-time detection.



METHODOLOGY

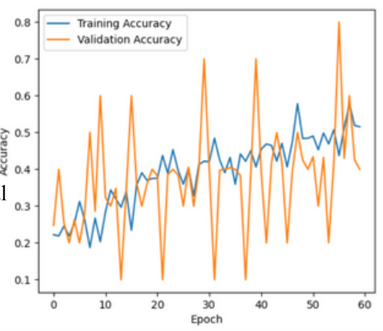
1. **Data Collection**
   - Visual Data (FER Dataset): Used for facial expression recognition with 35,887 images labeled by emotions
   - Audio Data: Emotional speech samples labeled for different emotions to capture vocal cues.
2. **Data Preprocessing**
   - Visual: Images resized, normalized, and augmented (e.g., rotation, flipping) to improve model robustness.
   - Audio: Noise reduction applied; MFCCs extracted for emotional representation & Whisper used to transcribe audio for text analysis.
3. **Feature Extraction**
   - Whisper Transcription & GloVe Embeddings: Converts speech to text & applies GloVe embeddings for meaningful word vectors.
   - LSTM for Temporal Patterns: Identifies patterns in word sequences linked to emotional states.
   - Custom CNN for Facial Expressions: Extracts visual cues from faces to detect emotions.
4. **Model Training and Fusion**
   - Multimodal Fusion: Combines audio (LSTM) and visual (CNN) features in a dense layer to improve emotion and depression detection.
   - Classification: Final model predicts emotional state and depression risk based on fused audio-visual features.
5. **Backend with FastAPI**
   - Real-Time Processing: FastAPI enables asynchronous, background processing of inputs for real-time analysis.
   - API Interface: Provides endpoints for input processing, classification, and result output, facilitating easy frontend integration.

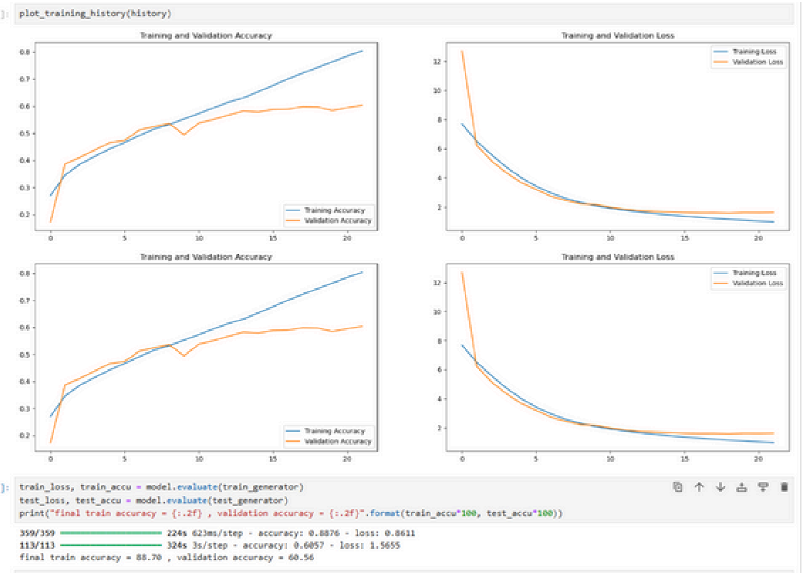

a Depression Detection System
ludio Data
Audio Trentusion using Whisper
Feature Trenscrstion using Whisper
Audio Trenspression using Custom CNN & Custom
Veature Featurion vii Custo CM
Feature Features
LST.ST TM & Custom CNN
LSTM
LNN
Feature Fealing
Feature Exrassions
Audio Trecressing using Whisper
Model Feecrestings
Feature Extrastions fong Text Text
Vitual Feature Extraction via Temporal Patterns
Model Extraction for Temporal Patterns
Depression Dessen
Depression Detecion
Model Ectraction for Temnal Patterns
Fast API
FASTAPI
FASTAPI
FASTAPI



RESULTS:

The developed multimodal system demonstrates significant improvements in accurately detecting emotions and assessing depression. Bycombining both audio and visual cues, the model achieves higher sensitivity and specificity compared to single-modality approaches. Initial testing indicates that the system effectively identifies emotional states and patterns associated with depressive symptoms in realtime, thanks to the FastAPI backend, which processes inputs smoothly in the background.





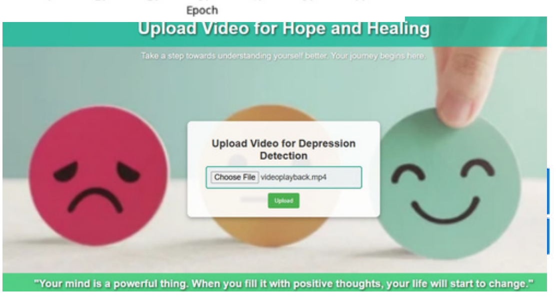
```
[42]: plot_training_history(history)
```

Training and Validation Accuracy
Training and Validation Loss
Training Accuracy
Validation Accuracy
Training Loss
Validation Loss
Training and Validation Accuracy
Training and Validation Loss

```
[45]: train_loss, train_accu = model.evaluate(train_generator)
      test_loss, test_accu = model.evaluate(test_generator)
      print("final train accuracy = {:.2f} , validation accuracy = {:.2f}".format(train_accu*100, test_accu*100))

359/359 ━━━━━━━━━━━━━━━━━━━━ 224s 623ms/step - accuracy: 0.8876 - loss: 0.8611
113/113 ━━━━━━━━━━━━━━━━━━━━ 324s 3s/step - accuracy: 0.6057 - loss: 1.5655
final train accuracy = 88.70 , validation accuracy = 60.56
```

A live demo is available for users to interact with the application and experience the real-time emotion and depression detection features


Upload Video for Hope and Healing
Take a step towards understanding yourself better. Your journey begins here.
Upload Video for Depression Detection
Choose File
videoplayback.mp4
Upload
"Your mind is a powerful thing. When you fill it with positive thoughts, your life will start to change."

DEMO:https://drive.google.com/file/d/1CRd5wZABjzJ1aLpoPmtK_6wdNj5_2EDM/view?usp=drive_link

REFERENCES

- W. C. de Melo, E. Granger, and A. Hadid, "Depression Detection Based on Deep Distribution Learning," presented at the 2019 IEEE International Conference on Image Processing (ICIP), Sep. 2019, doi:10.1109/icip.2019.8803467.
- M. Marcus et al., "Depression: A global public health concern," WHO Dept. Mental Health Substance Abuse, vol. 1, pp. 6–8, 2012.
- M. A. Jazaery and G. Guo, "Video-Based Depression Level Analysis by Encoding Deep Spatiotemporal Features," IEEE Trans. Affective Comput., pp. 1–1, 2018, doi: 10.1109/taffc.2018.2870884