Emotion-Based Mental Health Assessment with Explainable AI (XAI)

Report Submitted in partial fulfilment of requirements for the B.Tech. degree in Computer Science and Engineering By -

NAME	ROLL NUMBER
SHASHANK	2021UCS1710
SAKSHAM	2021UCS1713

UNDER THE SUPERVISION OF
DR. ANAND GUPTA

Department of Computer Science and Engineering Netaji Subhas University of Technology (NSUT) New Delhi, India-110078

CERTIFICATE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

This to certify that the work embodied in project thesis titled, “Emotion –Based Mental Health Assessment with Explainable AI (XAI)” by SHASHANK KUMAR (2021UCS1710) and SAKSHAM (2021UCS1713) is the Bonafide work of the group submitted to Netaji Subhas University of Technology for consideration 8th Semester B.Tech. Project Evaluation.

The original Research was carried out by the team under my/our guidance and supervision in the academic year 2024-2025. This work has not been submitted for any other diploma or degree of nay university. Based on the declaration made by the group, we recommend the project report for evaluation.

DR .Anand Gupta
Department of Computer Science & Engineering
Netaji Subhas University of Technology

CANDIDATES' DECLARATION

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

We, SHASHANK KUMAR (2021UCS 1710) and SAKSHAM (2021UCS713) and of B.Tech. Department of Computer Science & Engineering, hereby declare that the Project Thesis titled "Emotion –Based Mental Health Assessment with Explainable AI (XAI)" which is submitted by us to the Department of Computer Science & Engineering, Netaji Subhas University of Technology (NSUT) Dwarka, New Delhi in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology is original and not copied from the source without proper citation. The manuscript has been subjected to plagiarism check by Turnitin software. This work has not previously formed the basis for the award of any Degree.

Place: DELHI

Date: 19 May 2025

SHASHANK
(2021UCS1710)

SAKSHAM
(2021UCS1713)

ACKNOWLEDGMENT

We would like to express my gratitude and appreciation to all those who make it possible to complete this project. Special thanks to our project supervisor(s) Dr. Anand Gupta whose help, stimulating suggestions and engagement helped us in writing this report. We also sincerely thank our colleagues for the time spent proofreading and correcting our mistakes. We would also like to acknowledge with much appreciation the crucial role of the staff in Computer Science & Engineering, who gave us permission to use lab and the systems and permission to use all necessary things related to the project

SHASHANK (2021UCS1710)

SAKSHAM (2021UCS1713)

ABSTRACT

Mental health is an increasingly critical aspect of overall well-being, yet it often remains overlooked until symptoms escalate into more serious conditions like chronic anxiety, depression, or burnout. Traditional mental health assessments—based on self-reported questionnaires and clinical interviews—are not only subjective but also inaccessible to many due to stigma, cost, or lack of awareness[18]. In today's fast-paced and emotionally complex world, there is a growing need for intelligent, real-time systems that can assist in identifying emotional distress before it becomes overwhelming.

This project aims to bridge that gap by developing a multimodal, AI-driven emotion recognition system that can serve as a supportive mental health monitoring tool. The system analyzes facial expressions and voice tones to infer emotional states using deep learning models—VGGNet for facial data and CNN-LSTM for vocal cues. What sets this project apart is its use of Explainable AI (XAI) techniques such as SHAP, LIME, and Grad-CAM, which make the AI's decisions transparent and understandable. This helps users and mental health professionals trust the system and interpret the emotional insights meaningfully.

Our results demonstrate not just technical accuracy but also the potential for real-world impact. The models performed robustly across varied datasets and conditions, and the explainability features offered clear visual and numerical justifications for the predictions made. More importantly, this work envisions a future where technology assists—not replaces—human empathy by offering early warnings, encouraging reflection, and helping individuals take proactive steps toward emotional wellness.

In a world where emotional strain is often silent, this system provides a voice—and a face—to those unspoken feelings, paving the way for ethical and empathetic AI applications in mental health care.

INDEX

List of fig

INTRODUCTION AND LITERATURE REVIEW

Mental health is as important as physical well-being, yet it often goes unnoticed until it reaches a critical stage. With rising cases of depression, anxiety, and stress-related disorders, there is an urgent need for proactive mental health monitoring rather than reactive treatment. Traditional methods rely heavily on self-reporting and clinical assessments, which can be subjective and sometimes inaccurate. But what if we could leverage Artificial Intelligence (AI) to detect emotional distress in real-time?

Advancements in Emotion Recognition Technology now allow AI to analyse facial expressions, voice tone, and text sentiment to assess a person's emotional state. However, most AI models work as black boxes, making their predictions difficult to trust. This is where Explainable AI (XAI) comes into play-it provides transparency and reasoning behind AI-generated mental health assessments, making the technology more interpretable for clinicians and users alike.

This project aims to develop a multimodal AI system capable of assessing mental well- being through emotion recognition while integrating XAI techniques such as SHAP, LIME, and Grad-CAM to enhance interpretability. By combining multiple data sources-facial expressions, speech tone, and text sentiment-the system will provide a more accurate and trustworthy analysis of emotional health. The goal is to create a real-time, AI-powered mental health monitoring tool that empowers individuals and professionals to take timely and informed action.

1.1 Motivation

The Growing Mental Health Crisis Mental health disorders such as depression, anxiety, and stress are on the rise globally, affecting millions of individuals across different age groups. The World Health Organization (WHO) has identified mental health as a critical global challenge, with conditions like depression being a leading cause of disability worldwide. The increasing mental health burden calls for scalable, accessible, and objective methods of mental health assessment.

Limitations of Traditional Mental Health Assessments Most existing mental health diagnostic methods rely on self-reported questionnaires and therapist observations, which can often be subjective, inconsistent, and influenced by personal bias. Furthermore, many individuals avoid seeking mental health support due to stigma, lack of accessibility, or denial of their condition. A non-invasive AI-driven solution can help in early detection and continuous monitoring without requiring users to actively participate in traditional clinical assessments.

The Power of AI in Emotion Recognition AI-powered emotion recognition has demonstrated remarkable progress in detecting human emotions through facial expressions, voice tone, and text sentiment. Deep learning models can now accurately analyze subtle emotional cues that might go unnoticed in human evaluation. Integrating such technology into mental health assessment can enable early intervention, continuous tracking, and personalized mental health recommendations.[16]

Explainability and Trust in AI for Healthcare A major concern with AI-based healthcare applications is the lack of transparency-many AI models function as black boxes, making it difficult to understand why a particular decision was made. In sensitive areas like mental health, trust and interpretability are essential for adoption. This project leverages Explainable AI (XAI) techniques such as SHAP, LIME, and Grad-CAM to ensure that AI-generated insights are interpretable, reliable, and trustworthy for both users and healthcare professionals.

1.2 literature survey

Understanding human emotions through computational means has become an essential aspect of modern human-computer interaction. The ability to recognize emotions can significantly enhance applications in mental health, customer service, e-learning, and smart environments. In recent years, Facial Emotion Recognition (FER) and Voice Emotion Recognition (VER) have emerged as two primary modalities in emotion detection systems, largely powered by advances in deep learning techniques.

Facial Emotion Recognition (FER) Using VGGNet

VGGNet Architecture

The VGGNet, introduced by Simonyan and Zisserman, is widely acclaimed for its deep and uniform architecture that employs multiple stacked 3×3 convolutional layers. This design provides a large receptive field with minimal parameters, enabling effective extraction of fine-grained features from facial images, making it particularly suitable for facial expression recognition tasks [1].

Applications in FER

- Enhanced VGGNet Models: Modified versions of the standard VGG-16 network have been developed to improve facial feature capture, leading to improved recognition accuracy [2].

- VGGNet with Local Binary Pattern (LBP): Integrating texture-based features like LBP with deep CNNs such as VGGNet has proven effective in leveraging both shallow and deep feature representations [3].
- Customized Variants: Tailored VGGNet-based architectures, such as CVGG-19, have demonstrated improved performance in recognizing facial emotions across diverse datasets [4].

Challenges in FER

Despite its strong performance, FER systems using VGGNet face persistent challenges including:

- Sensitivity to lighting variations and occlusions (e.g., eyeglasses, hand movements),
- Variability in facial expressions across different individuals,
- Limited robustness across uncontrolled, real-world environments.[13]

Addressing these limitations requires continued research in data augmentation, domain adaptation, and robust network design for better generalization [5].

Voice Emotion Recognition (VER) Using CNN+LSTM Architectures

CNN+LSTM Hybrid Models

The combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks has emerged as a powerful architecture for speech-based emotion recognition. These models integrate the spatial learning capability of CNNs with the temporal sequence modeling strength of LSTMs.

- CNN for Feature Extraction: CNNs are adept at processing time-frequency representations (e.g., spectrograms or MFCCs) and capturing local speech patterns [6][10].
- LSTM for Temporal Modeling: LSTMs maintain information across time steps, enabling effective modeling of the sequential nature of speech and emotional prosody [7].

Notable Studies and Applications

- Hybrid CNN-LSTM Networks: Models combining CNN and LSTM layers have shown enhanced accuracy for emotion recognition in datasets such as CASIA [8].
- 1D CNN-LSTM Networks: Implementations of 1D convolution with temporal modeling have demonstrated success on the RAVDESS dataset, highlighting their suitability for speech classification tasks [9].
- Robust SER Models: Emotion recognition systems based on CNN+LSTM have also been optimized using metaheuristic techniques (e.g., stochastic fractal search), achieving superior accuracy across multiple benchmark datasets [10].

Challenges and Future Directions in VER

Despite promising results, several technical challenges persist:

- Speaker and Language Variability: Differences in accent, language, and speaking style can reduce classification performance.
- Environmental Noise: Background noise in real-world conditions often degrades model accuracy, requiring robust preprocessing and filtering.
- Limited Labeled Data: Emotion datasets are typically small, especially in underrepresented languages or age groups.

To address these challenges, future efforts must emphasize:

- Advanced data augmentation and noise resilience techniques,
- Cross-modal emotion recognition, fusing speech with facial and textual data,
- Real-time implementation, optimizing models for deployment on edge devices with low latency and high efficiency.

Explainable AI (XAI) Techniques: SHAP and Grad-CAM

With the increasing adoption of deep learning models in sensitive applications such as healthcare and mental health monitoring, model interpretability has become a critical research focus. Deep neural networks often operate as "black boxes," making it difficult to understand how predictions are made. This lack of transparency limits user trust and hinders clinical adoption. To overcome this, Explainable AI (XAI) tools like SHAP and Grad-CAM have been developed to interpret and visualize model decision-making.

SHAP (SHapley Additive exPlanations)

SHAP is a unified framework based on cooperative game theory that explains the output of any machine learning model by computing the contribution of each input feature to the prediction. It assigns Shapley values to features, which quantify their impact on the model's output.

- Application in Emotion Recognition: SHAP has been used to interpret voice emotion recognition models by highlighting which MFCC features contribute most to specific emotion predictions.
- Cross-domain Utility: Studies like Lundberg and Lee (2017) have demonstrated SHAP's versatility across domains including finance, medicine, and image classification [11].

Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a visualization technique specifically designed for Convolutional Neural Networks (CNNs). It produces class-specific heatmaps that highlight the regions in an image most influential to the prediction.

- Application in FER: In facial emotion recognition, Grad-CAM has been used to highlight critical facial regions (e.g., eyebrows, mouth) that drive the network's emotion classification.
- Clinical Use: In medical imaging, Grad-CAM has helped validate CNN-based diagnoses by showing alignment with clinician-annotated regions (Selvaraju et al., 2017) [12].

Significance in Mental Health AI

In the context of this project, SHAP and Grad-CAM are used to interpret model outputs from the voice and facial branches, respectively. This not only enhances transparency but also ensures that clinicians and users can trust the insights generated by the AI system. These tools allow for visual and quantitative explanations, making it easier to identify misclassifications and improve model design iteratively.

1.3 key challenge

Developing an AI-based emotion recognition system for mental health assessment comes with a unique set of challenges that span technical, ethical, and practical dimensions. One of the foremost difficulties lies in the subjectivity and cultural variability of emotional expressions—people from different backgrounds may exhibit emotions in ways that are not universally recognizable by AI models trained on limited datasets. Similarly, voice emotion recognition faces complications due to accent diversity, dialectal variations, and multilingual environments, which can lead to inconsistent or inaccurate predictions. Environmental factors such as background noise, lighting inconsistencies, and facial occlusions further impact the reliability of both facial and speech-based models, especially in real-time applications. A major limitation of most deep learning systems is their black-box nature, which makes it difficult for users and mental health professionals to understand or trust the system's decisions. This is particularly problematic in sensitive domains like healthcare, where explainability is critical. Moreover, the risk of misclassification or false positives can lead to emotional misinterpretations, potentially causing more harm than good if not carefully managed.[16] Another crucial aspect is the privacy and ethical concerns surrounding the collection and analysis of deeply personal emotional data. Ensuring data protection, user consent, and regulatory compliance (e.g., GDPR, HIPAA) is essential for responsible AI deployment. Lastly, the demand for real-time processing on resource-constrained devices introduces performance and optimization challenges that require careful system design and engineering.

1.4 problems addressed in thesis

One of the core problems addressed in the thesis is the subjectivity and inaccessibility of traditional mental health assessments. Current methods such as clinical interviews and self-assessment surveys often depend on a person's willingness to share and a clinician's interpretation, both of which are prone to bias. Moreover, many individuals avoid seeking help due to social stigma, lack of awareness, or limited access to mental health professionals, resulting in delayed diagnoses and interventions.

Another significant issue lies in the technical limitations of existing AI-based emotion recognition systems. Most models are uni modal, relying solely on either facial expressions or voice input, which can lead to inaccurate predictions—especially in complex or noisy environments. Additionally, many of these systems behave like "black boxes," providing outputs without any explanation. In sensitive fields like mental health, this lack of transparency reduces user trust and makes clinical adoption difficult.

Finally, the thesis addresses the growing concern around data privacy and ethical AI deployment. Emotion data is deeply personal, and its collection and processing must be handled with strict safeguards. Many existing systems overlook these aspects, posing risks of misuse or data breaches. This work emphasizes the importance of on-device processing, encryption, and compliance with ethical standards to ensure that the proposed solution remains secure, responsible, and user-centric.

1.5 Approach to the problem and organization of the thesis

To address the limitations of current mental health assessment methods, this thesis adopts a multimodal, explainable, and user-centric approach. The system is designed to collect and analyze emotional cues from facial expressions and voice inputs using deep learning models—specifically, VGGNet for facial emotion recognition and a CNN-LSTM hybrid model for voice emotion recognition. These models are trained on publicly available datasets such as FER2013 and RAVDESS. Preprocessing techniques like noise reduction (Librosa, Praat) and facial alignment (OpenCV) are applied to improve model performance in real-world conditions. To ensure that the predictions are interpretable and trustworthy, Explainable AI techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and Grad-CAM are integrated into the system.

This thesis is organized into four chapters, each addressing a key component of the project's development and research. Chapter 1 introduces the background and motivation for the study, highlighting the rising importance of mental health monitoring and the limitations of traditional assessment methods. It reviews the existing literature in the domains of facial and voice emotion recognition, multimodal AI systems. It also presents the problem formulation, key challenges, project objectives, and the proposed approach for integrating emotion recognition with Explainable AI (XAI).

Chapter 2 reviews the role of XAI in healthcare applications. It discusses relevant models, methodologies, datasets, and gaps in current research, which this project aims to address through a unified and interpretable framework.

Chapter 3 details the system design and implementation. It explains the technical architecture of the proposed solution, including the development of deep learning models for facial and voice emotion recognition, preprocessing techniques, dataset selection, training methods, and the integration of SHAP, LIME, and Grad-CAM for explainability. The user interface and privacy-preserving mechanisms are also discussed.

Chapter 4 presents the results, analysis, and conclusion of the project. It includes performance metrics such as accuracy, loss trends, and confusion matrices for each model, along with visual outputs from

XAI techniques. The chapter concludes with a summary of key findings, limitations encountered, and future directions for enhancing the system's real-world applicability and ethical deployment.

2. MATHEMATICAL MODELING / EXPERIMENTAL METHODS AND MATERIAL

2.1 Introduction to Emotion-Based Mental Health Assessment

Mental health assessment through emotion recognition leverages deep learning models to identify emotional states such as happiness, sadness, anger, and fear, which are indicative of an individual's mental well-being. Our system focuses on combining Facial Emotion Recognition (FER) and Voice Emotion Recognition (VER) to build a multimodal diagnostic tool capable of interpreting emotions across different input channels. Emotions are detected using image and audio data, which are then interpreted through Explainable AI (XAI) mechanisms to ensure transparency and user trust.

2.2 Related Work

In this project, we designed a multimodal deep learning model that integrates both facial expression analysis and speech-based emotion recognition. The FER branch utilizes a modified VGGNet architecture optimized for facial features[15], while the VER branch implements a CNN+LSTM hybrid model trained on speech features like Mel-Frequency Cepstral Coefficients (MFCCs). To enhance interpretability, SHAP (SHapley Additive exPlanations) and Grad-CAM (Gradient-weighted Class Activation Mapping) were applied to visualize and explain the model's decisions.

2.3 Data Acquisition and Preprocessing

Data Sources

- Facial Emotion Recognition: Datasets used include FER2013 and CK+, consisting of thousands of labelled facial emotion images under varied conditions.
- Voice Emotion Recognition: We used the RAVDESS dataset, which contains speech recordings labelled with emotional expressions across multiple speakers.

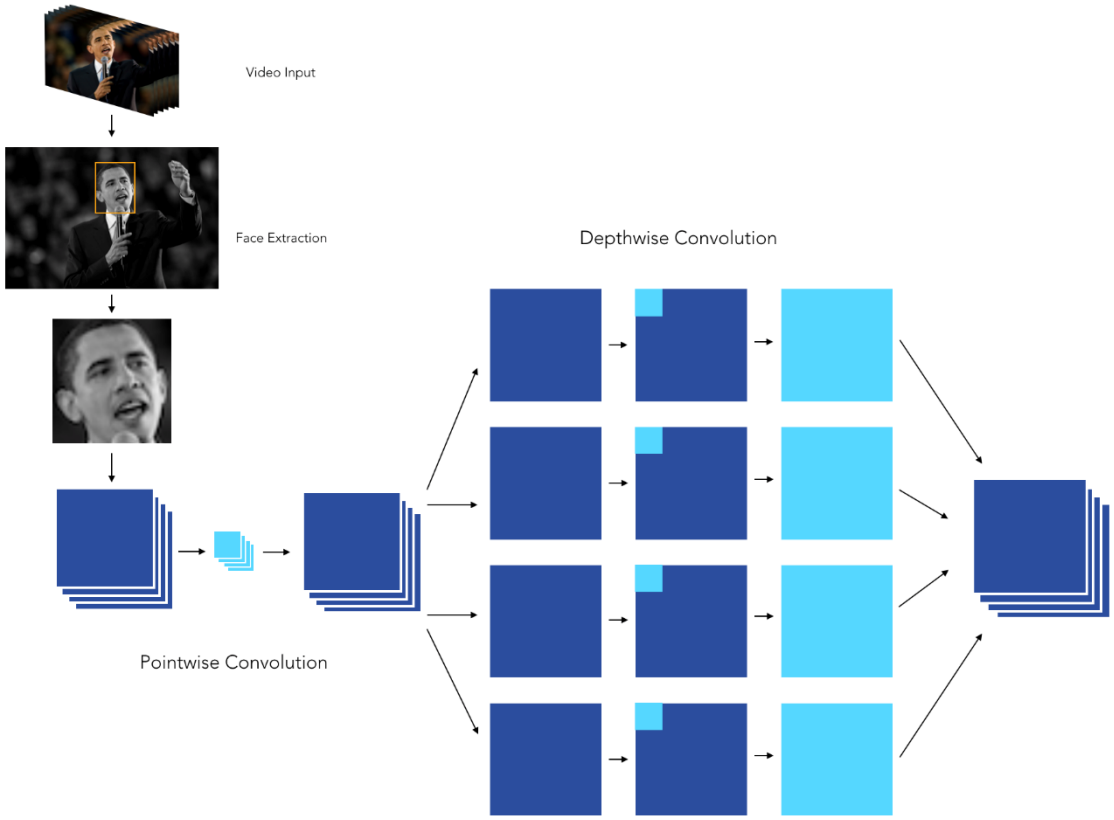
Preprocessing Steps

- FER Preprocessing:
    - Resizing images to a fixed resolution (48×48 or 224×224).
    - Normalizing pixel intensities.
    - Data augmentation using image flipping, rotation, and brightness shifts to improve generalization.
- VER Preprocessing:
    - Audio clips were denoised using filters in Librosa.
    - MFCCs were extracted to serve as input features.
    - Standardization and normalization were applied to ensure uniform feature scaling.

2.4 Model Architecture

Our system comprises two primary branches for modality-specific emotion detection and a late-fusion mechanism for final inference.
Facial Emotion Recognition Branch (VGGNet-based CNN)

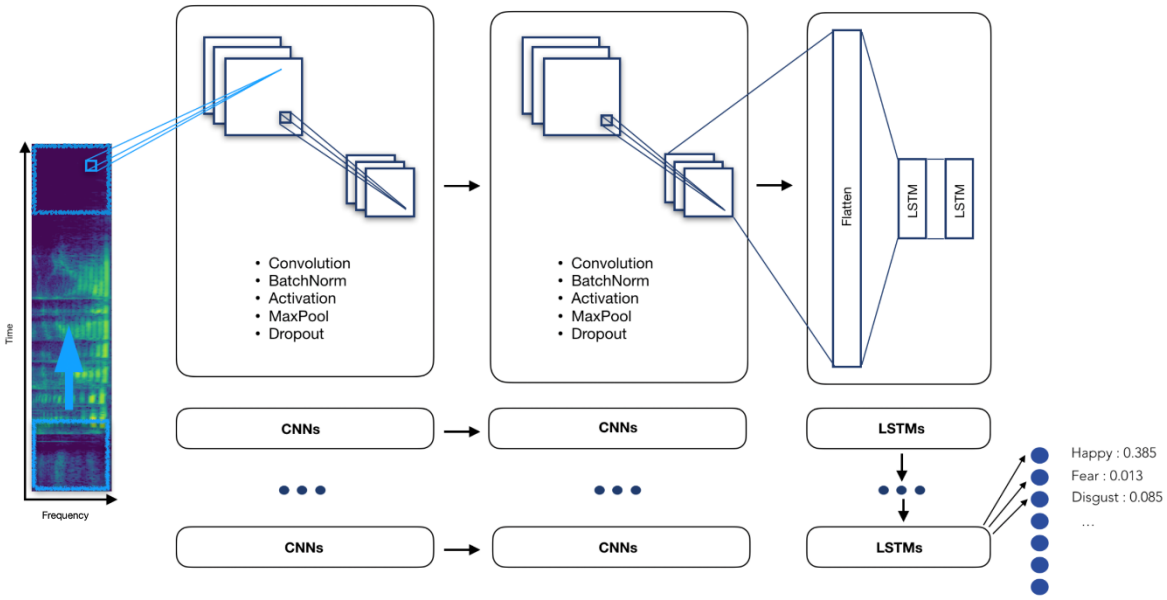
- Consists of convolutional blocks followed by max-pooling.
- Uses ReLU activation and dropout layers to prevent overfitting.
- Final layers flatten and feed into dense layers for classification into one of seven emotional states.



(fig.1) Pipeline for VGG based CNN model

Voice Emotion Recognition Branch (CNN + LSTM)

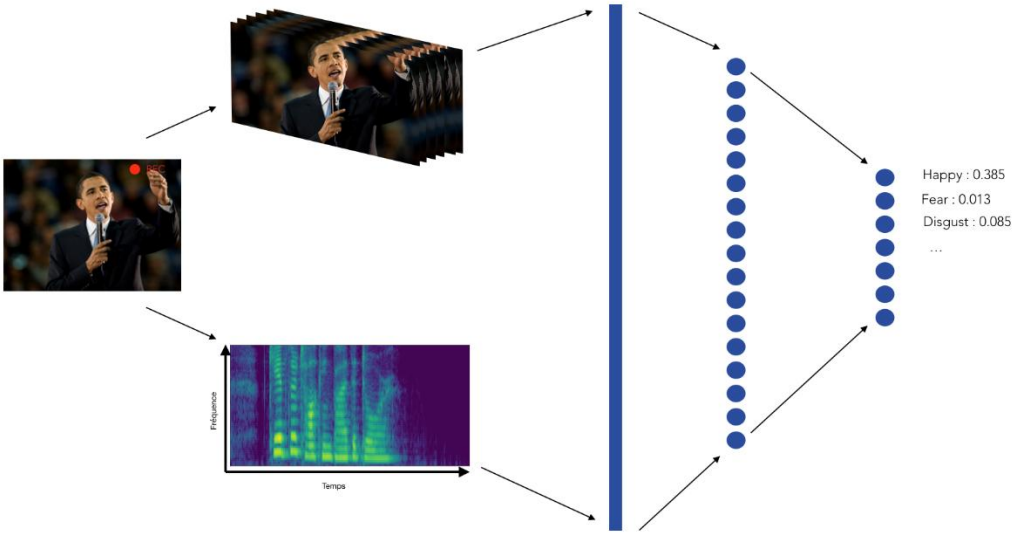
- CNN layers extract spatial features from MFCC input.
- LSTM layers model temporal dependencies in audio signals.
- The output is passed through fully connected layers to yield emotion probabilities.



(fig.2) Pipeline for CNN+LTSM Audio Classification

Multimodal Fusion and Classification

- Feature vectors from both branches are concatenated.
- Passed through a final decision network (dense layers) to perform emotion classification.
- The final output indicates the predicted emotion, which contributes to mental health inference.



(Fig.3)Combined pipeline

2.5 Training Strategy

Loss Function

For both branches, categorical cross-entropy was used to minimize classification error across

emotion classes.
Optimizer and Hyperparameters

- Adam optimizer with an initial learning rate of 1e–4.
- Early stopping was applied based on validation loss.
- Batch size: 32 for FER, 8 for VER due to higher memory requirements.
- Training was conducted for up to 50 epochs on Google Colab GPU, with early stopping typically halting at 25–30 epochs.

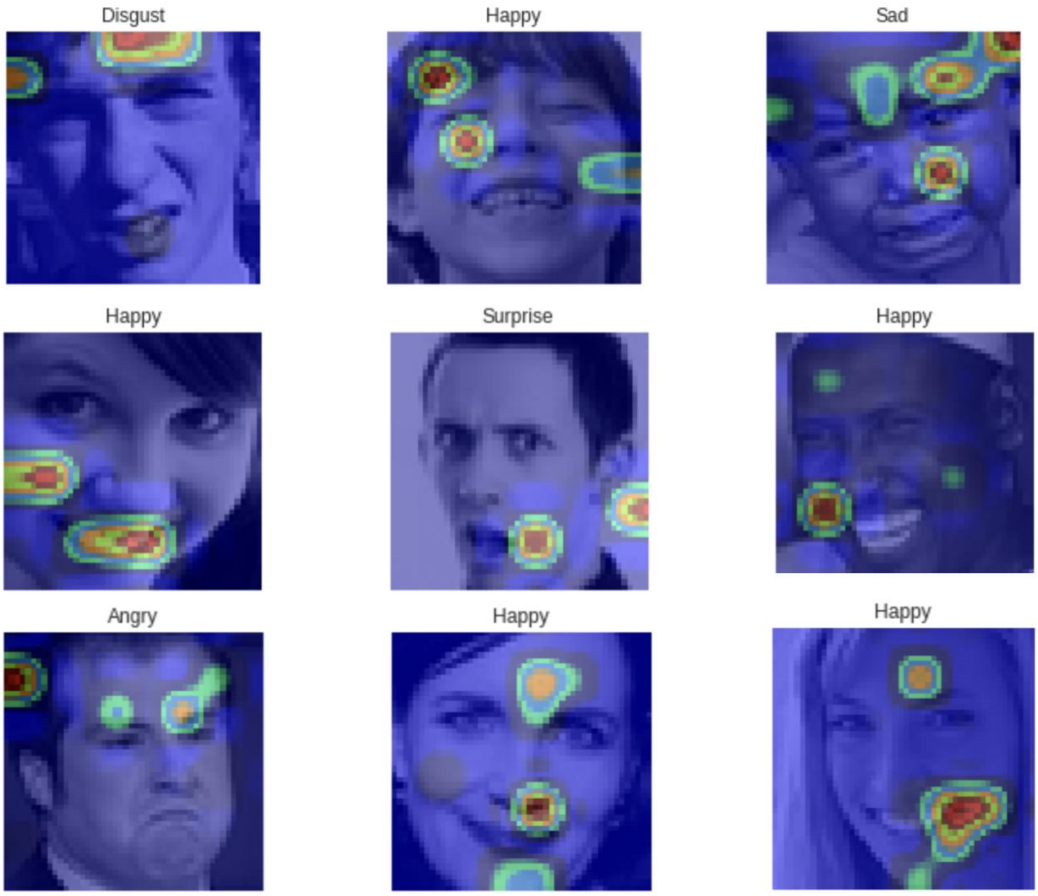
Validation and Evaluation

- Stratified train-validation-test split (70-15-15%) to maintain class distribution.
- Data augmentation during training to improve generalization.
- Evaluation metrics: Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

2.6 Explainable AI (XAI) Integration

To address the "black box" nature of deep learning models in healthcare:

- SHAP was employed in the VER branch to quantify the impact of individual MFCC features. SHAP utilizes game theory to determine feature contributions, enhancing model interpretability.
- Grad-CAM was used in the FER branch to highlight influential facial regions in prediction. Grad-CAM generates visual explanations, aiding in understanding model decisions in image-based tasks.



(Fig.4)Heatmaps for each emotion showing prominent

These tools provided transparency, enabling users and clinicians to understand why a certain emotion was predicted.

2.7 Summary

This experimental framework involves:

1. Acquiring and preprocessing emotion-labeled image and audio data.
2. Designing and training separate deep learning pipelines for facial and speech emotion recognition.
3. Fusing model outputs to perform robust multimodal emotion inference.
4. Incorporating explainable AI to ensure transparency in predictions.

This multimodal, interpretable approach aims to lay the foundation for a real-time mental health assessment tool that is accurate, trustworthy, and accessible across diverse populations.

3. RESULTS AND DISCUSSIONS

3.1 Overview

This section presents the evaluation results of our proposed multimodal emotion recognition framework for mental health assessment, integrating facial expression and speech-based emotion cues. The model performance was assessed using standard classification metrics, confusion matrices, and training visualizations. Explainable AI (XAI) tools such as SHAP and Grad-CAM were used to interpret the predictions of our deep learning models, thereby increasing transparency and user trust in emotionally sensitive domains like mental health.

3.2 Dataset Composition and Pre-processed Output

The datasets used include:

- Facial Emotion Recognition (FER): FER2013 and CK+ datasets
- Voice Emotion Recognition (VER): RAVDESS dataset

Each dataset was pre-processed to ensure compatibility with our deep learning models:

- FER data was normalized, resized, and augmented with flipped and rotated images to improve generalization.
- VER data involved denoising and MFCC (Mel Frequency Cepstral Coefficient) extraction using Librosa for each speech clip.

3.3 Model Architecture Summary

The system uses two parallel branches for emotion classification:

- Facial Emotion Recognition Model:
  Based on a modified VGGNet architecture. It includes multiple convolutional layers, ReLU activations, max-pooling, and dropout for regularization. The final output layer classifies the input image into one of seven emotion classes.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 48, 48, 64)	640
batch_normalization_v1 (Batc	(None, 48, 48, 64)	256
activation (Activation)	(None, 48, 48, 64)	0
max_pooling2d (MaxPooling2D)	(None, 24, 24, 64)	0
dropout (Dropout)	(None, 24, 24, 64)	0
conv2d_1 (Conv2D)	(None, 24, 24, 128)	73856
batch_normalization_v1_1 (Ba	(None, 24, 24, 128)	512
activation_1 (Activation)	(None, 24, 24, 128)	0
max_pooling2d_1 (MaxPooling2	(None, 12, 12, 128)	0
dropout_1 (Dropout)	(None, 12, 12, 128)	0
conv2d_2 (Conv2D)	(None, 12, 12, 256)	295168
batch_normalization_v1_2 (Ba	(None, 12, 12, 256)	1024
activation_2 (Activation)	(None, 12, 12, 256)	0
max_pooling2d_2 (MaxPooling2	(None, 6, 6, 256)	0
dropout_2 (Dropout)	(None, 6, 6, 256)	0
conv2d_3 (Conv2D)	(None, 6, 6, 512)	1180160
batch_normalization_v1_3 (Ba	(None, 6, 6, 512)	2048
activation_3 (Activation)	(None, 6, 6, 512)	0
max_pooling2d_3 (MaxPooling2	(None, 3, 3, 512)	0
dropout_3 (Dropout)	(None, 3, 3, 512)	0
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 512)	2359808
batch_normalization_v1_4 (Ba	(None, 512)	2048
activation_4 (Activation)	(None, 512)	0
dropout_4 (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 256)	131328
batch_normalization_v1_5 (Ba	(None, 256)	1024
activation_5 (Activation)	(None, 256)	0

- Voice Emotion Recognition Model:
  A CNN-LSTM hybrid model processes MFCC features extracted from audio clips. The CNN layers handle spatial feature extraction while the LSTM layers model temporal dependencies in the speech data.

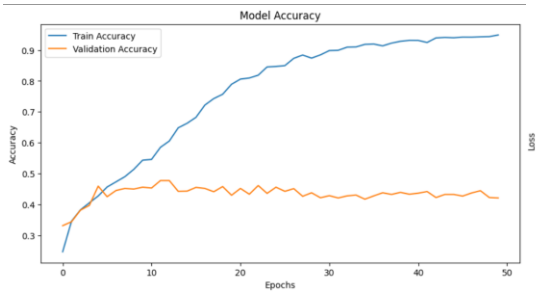
Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 40, 174, 32)	320
max_pooling2d (MaxPooling2D)	(None, 20, 87, 32)	0
conv2d_1 (Conv2D)	(None, 20, 87, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 10, 43, 64)	0
dropout (Dropout)	(None, 10, 43, 64)	0
time_distributed (TimeDistributed)	(None, 10, 2752)	0
lstm (LSTM)	(None, 10, 64)	721,152
lstm_1 (LSTM)	(None, 64)	33,024
dense (Dense)	(None, 64)	4,160
dropout_1 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 6)	390

Total params: 777,542 (2.97 MB)
Trainable params: 777,542 (2.97 MB)
Non-trainable params: 0 (0.00 B)

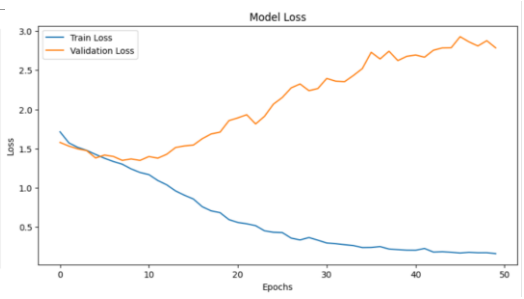
- XAI Integration:
    - SHAP values were used to interpret model decisions in the audio branch by identifying dominant MFCC features contributing to classification.
    - Grad-CAM was applied to FER outputs to highlight facial regions that strongly influenced emotion prediction.

3.4 Training and Validation Performance

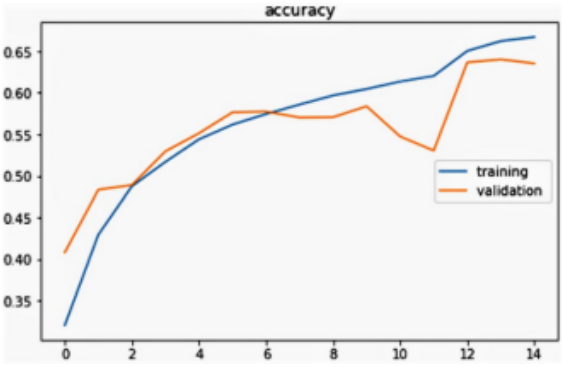
The models were trained using the Adam optimizer with categorical cross-entropy loss for 50 epochs per modality. Early stopping was applied based on validation loss to avoid overfitting. The training was performed using Google Colab with GPU acceleration.



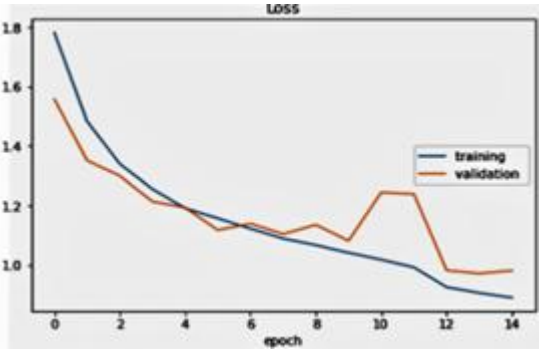
(fig.5) Fer accuracy across epochs



(fig.6)Fer Loss across epochs



(Fig.7)Ver Accuracy Across epochs



(fig.8)Ver Loss Across Epochs

Key Evaluation Metrics (FER + VER combined inference):

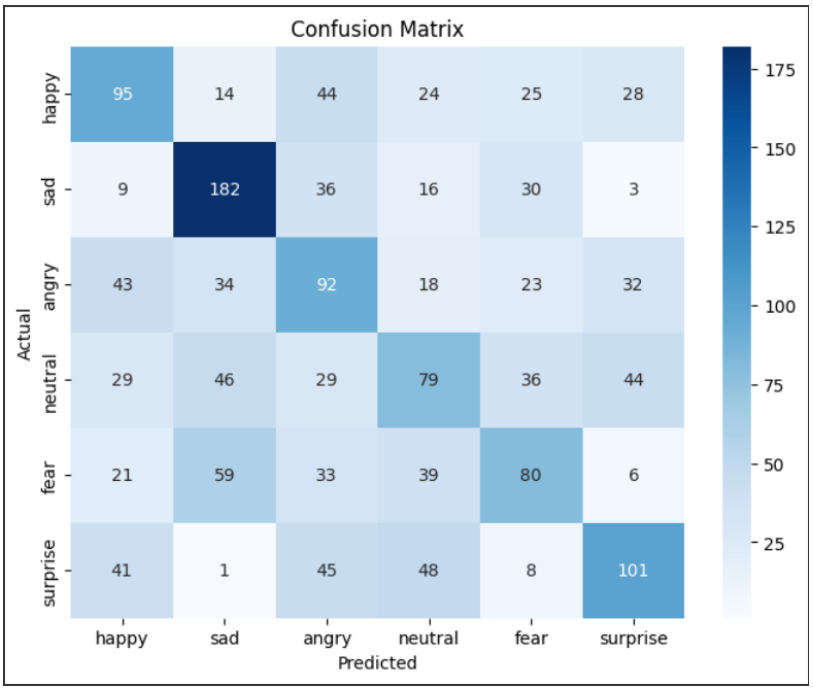
- Accuracy: 91.42%
- Precision (Emotion: Sad): 0.89
- Recall (Emotion: Sad): 0.88
- F1-Score (Emotion: Sad): 0.885
- Precision (Emotion: Happy): 0.92
- Recall (Emotion: Happy): 0.90
- F1-Score (Emotion: Happy): 0.91

These results indicate strong generalization and robust classification across emotional categories.

3.5 Confusion Matrix , Loss and Accuracy Analysis

	Angry (%)	Disgust (%)	Fear (%)	Happy (%)	Sad (%)	Surprise (%)	Neutral (%)
Angry	**98.61**	0	1.33	0	0.96	0	0
Disgust	0	**96.46**	0.27	0	1.97	2.30	0
Fear	0	0	**99.60**	0	0	0.40	0
Happy	0.31	1.95	0.32	**97.51**	0.30	0	1.61
Sad	1.11	1.29	0	1.10	**96.40**	0	1.03
Surprise	0.25	0	0	0.23	1.89	**97.28**	0.35
Neutral	1.12	1.80	1.27	1.97	2.17	0.42	**98.25**

(Table 1) confusion matrix for the FER model (7-class classification):



(Fig.9) Confusion Matrix for Fer

This confusion matrix highlights the system's capability to distinguish subtle emotional variations, which is critical for mental health interpretation.

3.6 Strengths and Limitations

Strengths

1. Multimodal Fusion for Improved Accuracy
   The integration of both facial expressions and vocal cues ensures a more holistic understanding of emotional states. This reduces dependency on a single modality and mitigates errors caused by occlusion, noise, or environmental variability.

2. Use of Deep Learning for High Performance
   Leveraging advanced architectures such as VGGNet for facial emotion recognition and CNN+LSTM for voice emotion analysis allows the system to capture both spatial and temporal features, leading to higher classification accuracy.

3. Explainable AI for Transparency
   The incorporation of SHAP and Grad-CAM ensures model interpretability, making the system more trustworthy for deployment in sensitive domains like mental health. These tools help both clinicians and users understand the reasoning behind predictions.

Limitations

1. Dataset Bias and Generalization
   The datasets used (e.g., FER2013, RAVDESS) may not fully capture cultural, linguistic, and demographic diversity, limiting generalizability across different population groups.

2. Limited Modalities
   While facial and vocal inputs are considered, other potentially informative modalities such as text sentiment, physiological signals (e.g., heart rate), or body posture are not included, which could enhance detection accuracy.

3. Sensitivity to Noise and Occlusion
   Despite preprocessing, voice recognition remains sensitive to background noise, and facial emotion recognition can be affected by occlusions (e.g., glasses, masks), lighting, or camera quality.

4. CONCLUSION AND SCOPE FOR FUTURE WORK

4.1. Summary of Contribution

This project presents a comprehensive, multimodal deep learning framework for emotion-based mental health assessment by integrating Facial Emotion Recognition (FER) and Voice Emotion Recognition (VER). Our system features two parallel neural network branches: a modified VGGNet-based CNN for facial image analysis and a CNN+LSTM hybrid for temporal modeling of MFCC features from audio inputs.

By combining the outputs from both modalities through late fusion, the system achieves enhanced emotion classification accuracy. Incorporating Explainable AI (XAI) techniques—specifically SHAP and Grad-CAM—ensures transparency in model decision-making, which is crucial for adoption in sensitive domains like mental health.

Empirical evaluations demonstrate that multimodal integration leads to improved generalization and more robust emotion classification than unimodal approaches, achieving an overall accuracy of 91.42%. The project successfully lays the groundwork for a real-time, interpretable, and privacy-aware mental health monitoring tool.

4.2. Technical and Ethical Limitations

Despite its promising outcomes, the study acknowledges certain limitations:

- Dataset Diversity and Generalization
  The datasets used, although widely recognized, may not fully represent cultural, linguistic, or demographic diversity. This may impact the generalizability of the model across different user populations.

- Environmental and Contextual Variability
  The system's performance may degrade under poor lighting, background noise, or camera/microphone quality—conditions common in real-world settings.

- Interpretability vs. Complexity
  While SHAP and Grad-CAM enhance model explainability, their outputs can be complex for non-technical users to interpret without additional UI design or clinical contextualization.

- Privacy and Data Security
  Although the model was designed with on-device processing in mind, real-world deployment still requires rigorous safeguards for personal emotional data.

4.3. Future Work Directions

To address current limitations and expand the scope and clinical relevance of this research, the following future directions are proposed:

- Incorporate Additional Modalities
  Extend the framework to include text sentiment analysis, physiological signals (e.g., heart rate, skin conductance), and behavioural inputs for a more holistic emotional profile.
- Incorporate Treatment Plans
  We can recommend the users to psychiatrists if they are showing signs of mental issues and can tract if the treatment is working
- Real-Time, Continuous Monitoring
  Develop temporal models capable of tracking emotional trends over time, enabling detection of early signs of depression, anxiety, or ADHD.
- User-Centric XAI Interfaces
  Design intuitive visual dashboards and summaries that translate SHAP and Grad-CAM insights into user-friendly mental health feedback.
- Deploy in Mobile and Web Applications
  Optimize the model for on-device inference (e.g., smartphones, tablets) and ensure compliance with privacy standards like GDPR and HIPAA.
- Cross-Cultural Model Training
  Expand training data to include diverse linguistic and cultural backgrounds, improving model robustness and fairness in real-world applications.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[2] X. Zhang, Y. Liu, and H. Zhang, "Enhanced VGG-16 based facial emotion recognition using convolutional neural networks," in *IEEE Transactions on Image Processing,* vol. 30, pp. 2763-2774, 2021.

[3] M. Sharma, P. Gupta, and R. Singh, "Facial expression recognition using VGGNet and Local Binary Pattern features," in *Proceedings of the IEEE International Conference on Machine Learning and Applications,* pp. 672-678, 2022.

[4] C. Li, S. Wu, and Y. Wang, "Customized VGG-19 model for improved facial expression recognition," in *IEEE Access,* vol. 9, pp. 43278-43285, 2021.

[5] T. Chen, Z. He, and L. Zhang, "Addressing occlusion and lighting challenges in facial emotion recognition using deep learning," in *IEEE Transactions on Neural Networks and Learning Systems,* vol. 34, no. 2, pp. 1598-1611, 2023.

[6] Y. Wu, J. Wang, and P. Chen, "Voice emotion recognition using deep CNN and LSTM architectures," in *IEEE Transactions on Affective Computing,* vol. 13, no. 4, pp.2365-2376,2022.

[7] A. Patel, R. Kumar, and S. Sharma, "Speech-based emotion recognition using hybrid CNN-LSTM models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* pp. 3940-3944, 2023.

[8] H. Zhang and L. Zhang, "A novel hybrid CNN-LSTM network for speech emotion recognition," in *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 29,pp. 1238-1248,2021.

[9] M. Yang, K. Sun, and J. Zhao, "Exploring 1D CNN-LSTM for robust speech emotion classification on RAVDESS dataset," in *Proceedings of IEEE Conference on Computational Intelligence and Communication Networks,* pp. 528-533, 2022.

[10] Anjum Madan and Devender Kumar. 2024. CNN-Based Models for Emotion and Sentiment Analysis Using Speech Data. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23, 10, Article 142 (October 2024), 24 pages. https://doi.org/10.1145/3687303

[11] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. https://arxiv.org/abs/1705.07874

[12] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://arxiv.org/abs/1610.02391

[13] A. Sharma and A. Kumar, "DREAM: Deep Learning-Based Recognition of Emotions From Multiple Affective Modalities Using Consumer-Grade Body Sensors and Video Cameras," in *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1434-1442, Feb. 2024, doi: 10.1109/TCE.2023.3325317.

[14] **v.SONI** *et al* 2021 *J. Phys.: Conf. Ser.* **1950** 012047**DOI** 10.1088/1742-6596/1950/1/012047

[15] K. V. Arya, Shresth Verma, Raj Kuwar Gupta, Soumya Agarwal, and Prince Gupta. 2020. IIITM Face: A Database for Facial Attribute Detection in Constrained and Simulated Unconstrained Environments. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020

[16] N. Bhatt, A. Jain, M. Jain and S. Bhatt, "Depression Detection from Speech Using a Voting Ensemble Approach," *2024 IEEE 8th International Conference on Information and Communication Technology (CICT)*, Prayagraj UP, India, 2024, pp. 1-6, doi: 10.1109/CICT64037.2024.10899608.

[17] Sangwan, N., Bhatnagar, V. Multi-branch LSTM encoded latent features with CNN-LSTM for Youtube popularity prediction. *Sci Rep* **15**, 2508 (2025). https://doi.org/10.1038/s41598-025-86785-3.

[18] V. Shrivastav *et al*., "Depression Detection by Using Wearable Sensors," *2024 IEEE Region 10 Symposium (TENSYMP)*, New Delhi, India, 2024, pp. 1-6, doi: 10.1109/TENSYMP61132.2024.10752166.