DIAGNOSING AND MANAGING DEPRESSION THROUGH TEXT AND FACIAL IMAGES

Final Report

Perera K.K.S – IT21178368

Bachelor of Science in Information Technology

Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology

Sri Lanka

April 2025

DIAGNOSING AND MANAGING DEPRESSION THROUGH TEXT AND FACIAL IMAGES

Final Report

Perera K.K.S – IT21178368

Supervisor: Ms. Wishalya Tissera

Co – Supervisor: Dr. Kapila Dissanayaka

Bachelor of Science in Information Technology

Specializing in Data Science

Department of Computer Science

Sri Lanka Institute of Information Technology

Sri Lanka

April 2025

DECLARATION

I declare that this is my own work, and this report does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 11/04/2025

The above candidate has carried out research for the bachelor's degree Dissertation under my supervision.

Signature of the Supervisor: Date:

Signature of the Co-Supervisor: Date:

Abstract

This research presents a novel multi-modal approach to diagnosing depression by contextually analyzing both user-generated text and real-time emotional states captured through facial expressions. The system consists of two integrated components: a text analysis pipeline and an image-based emotion recognition module. The text component employs a Bidirectional Long Short-Term Memory (BiLSTM) model enhanced with an attention mechanism to identify linguistic cues related to depressive symptoms, achieving a classification accuracy of approximately 96%.

In parallel, facial expressions are analyzed using a pre-trained VGG19 model to classify the user's current emotional state into one of seven categories: sad, surprised, neutral, happy, fearful, disgust, and angry. Explainable AI techniques such as Grad-CAM are applied to the VGG19 model to highlight facial regions contributing to emotion classification, ensuring model interpretability and transparency for clinical use.

Rather than using conventional ensemble techniques, this research proposes a context-aware fusion strategy. The detected emotion is first translated into a corresponding textual representation and then appended to the patient's original text input. This combined text sequence is processed by the BiLSTM-attention model, which learns to interpret the relationship between the patient's expressed thoughts and emotional state to estimate the probability of depression. This unified text-based processing allows for deeper contextual understanding, ultimately supporting more accurate and interpretable mental health diagnostics.

Keywords: Multi-modal fusion, Depression detection, VGG19, BiLSTM, Attention mechanism, Explainable AI, Emotion recognition, Context-aware analysis

Acknowledgements

I would like to express my heartfelt gratitude to all those who supported and guided me throughout the course of this research project. Completing this study would not have been possible without the invaluable contributions, encouragement, and support I received from numerous individuals.

First and foremost, I wish to extend my deepest appreciation to my academic supervisors, Mrs. Wishalya Tissera and Dr. Kapila Dissanayaka. I am profoundly grateful for their consistent guidance, insightful feedback, and unwavering support throughout every stage of this project. In particular, I would like to acknowledge Dr. Kapila for his dedicated mentorship, critical insights, and thoughtful suggestions that significantly shaped the direction and development of this research. His leadership and encouragement served as a constant source of motivation and clarity.

I would also like to express my sincere thanks to all the lecturers and staff members at SLIIT for providing the necessary facilities, academic support, and a nurturing environment that enabled me to carry out this research successfully. The continual encouragement and willingness to assist, both academically and administratively, played an important role in the completion of this work.

To my friends, family, and everyone who offered words of encouragement and support during this journey, your presence and belief in me meant more than words can convey.

Table of Contents

List of Figures

List Of Tables

List of Abbreviations

Abbreviation	**Description**
GP	General Practitioner
CNN	Convolutional Neural Network
ML	Machine Learning
DL	Deep Learning
NLP	Natural Language Processing
TF-IDF	Term Frequency - Inverse Document Frequency
LSTM	Long Short -Term Memory

1. INTRODUCTION

1.1 Background

Mental health disorders, including depression, have become a growing concern globally. As per the World Health Organizations' estimation, 3.8% of the population suffer from Depression. [1] Millions of individuals are impacted globally, and it has a big influence on their everyday life, interpersonal connections, and general well-being. Many people don't receive the proper diagnosis or treatment, even though there are plenty of options available. These obstacles include the stigma attached to mental health issues and the shortcomings of the diagnostic techniques used today.

Conventional techniques for diagnosing depression mostly depend on clinical assessments and self-reported questionnaires, both of which have limitations due to the patient's limited capacity to describe their symptoms and can be subjective. Moreover, these techniques are frequently needed for in-person meetings with mental health specialists, which can be expensive, time-consuming, and unavailable to many. As a result, more effective and scalable techniques for diagnosis and treatment of depression are required.

New opportunities in the field of mental health have been made possible by recent developments in ML and AI. Artificial intelligence (AI)-driven solutions have demonstrated potential in comprehending people's emotional and mental states, especially in the analysis of text and visual data. Through the use of computer vision techniques for picture analysis and natural language processing (NLP) for text analysis, machine learning models are able to identify precise patterns and traits that may be indicative of depression. By enabling real-time input and analysis through mobile applications, these AI-driven models can improve the accessibility and individualization of mental health care.

With this study, I hope to investigate how integrating textual and visual data sources could improve the precision of depression diagnosis. In order to discover indicators of depression, I will use multimodal methodologies to analyze textual inputs, such as patients' feelings in texts and their photographed facial expressions. I'm going to specifically implement two approaches: an intermediate fusion method that uses a multi-modal neural network to integrate both text and image data types within a single architecture, and a late fusion method that combines the outputs of separate text and image models using a weighted average model. My goal is to find the best method for diagnosing depression and its level by analyzing these models' performances and contrasting their results.

Additionally, this research will incorporate attention mechanisms to enhance the focus of the models on relevant features within the text and images, leading to more accurate predictions. The final output of the system will be a comprehensive analysis provided to mental health professionals, assisting them in diagnosing depression and its levels. By utilizing XAI, I aim to improve the transparency of the model's decisions, ensuring that clinicians can trust and understand the AI's recommendations. The proposed system will be evaluated with input from clinical experts to ensure its practical applicability and effectiveness. With the integration of AI into depression diagnosis, I hope to make mental health care more accessible, personalized, and efficient, ultimately improving outcomes for individuals suffering from depression.

2. LITERATURE REVIEW

This paper explores the application of machine learning models to enhance the detection of depression using a tabular dataset. The dataset consists of 10 columns, including a target variable, and captures various patient responses to general questions aimed at identifying depressive symptoms. The study focuses on predicting whether a patient is experiencing depression based on their answers to these questions. By leveraging machine learning techniques, the authors demonstrate the potential for automating depression detection, which could lead to faster and more efficient assessments. However, the research is limited to the detection of depression through tabular dataset and does not extend to further analysis or treatment suggestions. The findings emphasize the importance of using structured data for improving diagnostic accuracy in mental health. [2]

In their study on depressive and non-depressive tweet classification, the authors proposed a sequential deep learning model for depression detection using only text-based data from Twitter. The model comprises three layers: an embedded layer, a 1D-convolutional layer, and an LSTM layer. The study compared the performance of this deep learning model with traditional machine learning models, including Naïve Bayes, KNN, and Random Forest. The sequential deep learning model demonstrated superior performance, achieving an accuracy of 98.47%, surpassing the other models tested. The results highlight the efficacy of the deep learning approach in effectively distinguishing between depressive and non-depressive tweets. [3]

This study addresses the challenge of identifying depression from social media data through machine learning, focusing on improving classification efficiency by integrating feature selection techniques. Depression, particularly major depression, manifests through symptoms such as prolonged low mood and diminished interest in previously enjoyable activities. The research highlights the severe societal impact of depression and explores two machine learning classifiers that predict depressive states based on a combination of demographic and psychological factors. A key aspect of the study is the use of Recursive Feature Elimination (RFE) with linear regression to optimize the feature set, thereby enhancing the model's performance by identifying the most relevant characteristics within the dataset. This approach aims to strike a balance between model accuracy and computational efficiency, offering a novel contribution to the domain of mental health analysis through social media data. [4]

The article demonstrates the development of a system designed to identify and quantify depression levels using visual data, specifically facial expressions. The paper focuses on utilizing machine learning models, particularly Convolutional Neural Networks (CNNs), to analyze visual inputs and detect depressive symptoms. The research is grounded in the premise that non-verbal cues, such as facial expressions, can serve as reliable indicators of an individual's mental health status. The system's architecture leverages pre-trained models, such as VGG19, to capture features from facial images, which are then classified into various depression levels. The paper emphasizes the importance of early detection of depression, highlighting the potential for automated systems to assist clinicians in diagnosing mental health conditions more efficiently. Moreover, the study also discusses the integration of such systems into real-world clinical settings, considering factors like accuracy, reliability, and interpretability of the results. Through experimental evaluations, the researchers demonstrate the efficacy of their approach, showing promising results in the automatic detection of depression levels. [5]

The article delves into the utilization of visual facial cues for identifying and diagnosing depression. The research underscores the importance of facial expressions as non-verbal indicators of emotional and mental health. The study reviews various machine learning techniques, focusing primarily on deep learning approaches like Convolutional Neural Networks (CNNs), which are employed to analyze facial features and classify depressive symptoms. The paper also explores the integration of pre-trained models to extract relevant features from facial images, enhancing the accuracy and reliability of depression detection systems. Furthermore, the research highlights the challenges faced in real-world applications, such as balancing detection accuracy and computational efficiency, as well as addressing issues related to data variability across different demographic groups. The study concludes by emphasizing the potential of integrating these visual based models into clinical settings to assist healthcare professionals in diagnosing depression, making the process more efficient and scalable. [6]

The paper consists of a novel approach to detecting depression levels by leveraging both spatial and temporal features from multimodal data sources. The authors utilize a combination of facial expressions and textual data to create a robust model that captures the nuanced indicators of depression over time. By integrating multimodal inputs, including video frames and speech transcripts, the proposed system enhances the accuracy of depression detection. The research employs a hybrid architecture combining Convolutional Neural Networks (CNNs) for spatial data and Long Short-Term Memory (LSTM) networks for temporal data, effectively addressing the dynamic nature of emotional expressions. Additionally, the study explores the role of feature fusion techniques to merge information from different modalities, thereby improving prediction outcomes. The paper concludes by highlighting the potential of this multimodal approach for real-world applications in mental health diagnostics, offering a promising avenue for early intervention and personalized treatment plans. [7]

The paper introduces a novel approach for estimating depression severity by leveraging multimodal data, including audio, visual, and textual inputs. The proposed method, FNBOT, utilizes optimal transport theory to align the feature distributions of different modalities, thereby capturing complex correlations essential for accurate depression assessment. The architecture features a fusion encoder-decoder structure, where the encoder aligns cross-modal semantics, and the decoder reconstructs modality-specific information. This design enables the model to learn a joint multimodal representation that preserves individual modality characteristics while integrating global semantic cues. Evaluated on the AVEC 2019 dataset, FNBOT outperformed several state-of-the-art models in predicting PHQ-8 depression scores, demonstrating superior correlation and reduced prediction error. This study is particularly relevant to multimodal depression detection research, as it highlights the effectiveness of advanced feature fusion and alignment techniques, providing valuable insights and benchmarks for future developments in the field. [8]

This paper introduced a novel approach for depression detection using non-verbal features from vlog data through their proposed Time-Aware Attention Multimodal Fusion Network (TAMFN). The model effectively fuses acoustic and visual modalities using three core components: a Global Temporal Convolutional Network (GTCN) for capturing both local and global temporal dependencies, an Intermodal Feature Extraction (IFE) module to extract interaction-level features between modalities, and a Time-Aware Attention Multimodal Fusion (TAMF) module that guides the fusion process by modeling the temporal importance of each modality. The TAMFN model demonstrated superior performance over existing traditional machine learning and deep learning models on the D-Vlog dataset, with improved precision, recall, and F1 scores. Additional evaluations on the EATD-Corpus further affirmed the model's robustness across different environments. This work highlights the importance of fusing temporal dynamics across modalities and offers an effective and generalizable framework for depression detection using real-world social media video content. [9]

3. RESEARCH GAP

Millions of individuals worldwide suffer from depression, a common mental health illness that makes early diagnosis and treatment extremely difficult. There is still a study gap in the thorough application of multi-modal techniques to diagnose depression by merging various data inputs, such as text, facial expressions, and speech, despite advances in machine learning and artificial intelligence for mental health applications. A large number of previous research works have concentrated on single-modal methods that employ facial, audio, or textual data to diagnose depression. Although these techniques show promise, they frequently fall short of the depth and accuracy that come from multi-modal analysis. Moreover, rather than offering a comprehensive examination of depression severity levels, the majority of recent research has focused on categorization tasks for determining if an individual is depressed. This restricts these models' practical application in actual clinical settings. The suggested remedy, "Mentcare AI," seeks to bridge this research gap by employing multi-modal neural network techniques to incorporate text and visual data for more accurate depression diagnosis. When XAI is combined with these models, it offers a clearer grasp of the reasoning behind diagnosis, giving medical professionals more insight into the prediction process. This improves the system's accuracy while also boosting healthcare specialists' confidence in it. By comparing the proposed solution with existing models, "Mentcare AI" shows promise in advancing the field of AI-driven mental health applications, offering a more holistic and explainable approach to depression diagnosis.

TABLE 1. COMPARISON BETWEEN EXISTING METHODS AND PROPOSED TOOLS

Device/ Application	Depression Detection	Multi-modal Approach	Model Explainability	Attention Mechanism	Depression Level Detection	Real-time Data Integration
Research [2]	Yes	No	No	No	No	No
Research [3]	Yes	No	No	No	No	No
Research [4]	Yes	No	No	Limited	No	No
Research [5]	Yes	Limited	No	Yes	No	No
Research [6]	Yes	No	No	No	No	Yes
Research [7]	Yes	Yes	No	Yes	No	Yes
Research [8]	Yes	Yes	No	No	No	No
Proposed Solution	Yes	Yes	Yes	Yes	Yes	Yes

4. RESEARCH PROBLEM

Depression is a prevalent mental health disorder that affects millions of people worldwide, contributing to substantial emotional, social, and economic burdens. It is characterized by persistent sadness, loss of interest, and a range of cognitive and physical symptoms, which can severely impair an individual's ability to function in daily life. Despite its widespread impact, depression often goes undiagnosed or is misdiagnosed, leading to inadequate treatment and prolonged suffering.

Conventional approaches to depression diagnosis mostly rely on self-reported questionnaires and clinical interviews, which are highly subjective by nature and can differ greatly depending on the experience of the therapist. In situations where patients are unable to describe their symptoms, these approaches may fall short of fully capturing the complexity of the illness. Furthermore, long appointment wait times are a common consequence of the rising demand for mental health care, which postpones necessary interventions.

Recent technological developments, especially in the area of AI, offer a chance to improve the precision and effectiveness of depression diagnosis. But the majority of current AI-driven methods concentrate on single-modal data, including text-based analysis or facial emotion detection, which might not adequately represent the multifaceted character of depression. It is also challenging for physicians to trust and use AI technologies in practice because of the "black-box" character of many of these models, which raises questions about the transparency and interpretability of the diagnostic process.

This research seeks to address these challenges by developing and evaluating two advanced multi modal approaches: a Weighted Average method and a Multi-modal Neural Network. These approaches integrate text and facial emotion data to improve the identification and classification of depression levels in patients. By leveraging a mobile application for real-time data collection, the model will provide personalized and timely insights into a patient’s mental health. With the integration of data from many sources, the suggested model seeks to offer a more thorough and precise diagnosis. Additionally, the model's predictions will be visible and comprehensible thanks to the integration of explainable AI approaches, enabling doctors to comprehend the underlying logic and make defensible conclusions. By addressing the shortcomings of single-modal AI models and conventional diagnostic techniques, this strategy hopes to improve patient outcomes for the identification and treatment of depression.

5. OBJECTIVES

5.1 Main Objectives

The primary objective of this research is to develop and evaluate a comprehensive system designed to assist general practitioners (GPs) in the diagnosis and management of depression by integrating patient-generated textual inputs and facial imagery. Traditionally, GPs diagnose depression through clinical interviews and standardized questionnaires, such as the Patient Health Questionnaire-9 (PHQ-9), which, while validated, can be influenced by subjective patient responses and may not capture the full spectrum of depressive indicators and it’s time consuming.

To augment this traditional diagnostic process, the proposed system will collect and analyze data from patients who engage with it regularly, providing textual inputs and facial images. This consistent interaction allows for continuous monitoring of the patient's mental state, facilitating the detection of subtle changes over time that might indicate the onset or progression of depression. By employing advanced natural language processing techniques, the system will analyze textual data to identify linguistic markers associated with depression. Concurrently, facial emotion recognition algorithms will assess facial imagery to detect emotional expressions indicative of depressive states. The fusion of these modalities aims to provide a more objective, comprehensive, and nuanced understanding of a patient's mental health status.

A significant aspect of this research is the emphasis on patient engagement with the system. Regular interaction is encouraged to ensure a continuous and rich data stream, which is crucial for accurate monitoring and timely intervention. This approach not only empowers patients to be active participants in their mental health management but also enables GPs to make more informed decisions based on a holistic view of the patient's emotional and linguistic data.

Furthermore, the system incorporates Explainable Artificial Intelligence (XAI) techniques to enhance the transparency and interpretability of its analyses. By elucidating the reasoning behind its assessments, the system builds trust with both clinicians and patients, facilitating its integration into clinical practice. Ultimately, this research seeks to bridge the gap between traditional diagnostic methods and modern technological advancements, aiming to improve the accuracy of depression diagnoses and support GPs in delivering personalized and effective mental health care.

5.2 Specific Objectives

To achieve the main objective of developing a multi-modal neural network for depression level identification, the following sub-objectives have been outlined:

Sub Objective 1: Develop a Weighted Average Model for Depression Diagnosis

The first sub-objective is to develop a Weighted Average model that combines the outputs from text-based analysis and facial emotion recognition. This model will assign appropriate weights to each modality, calculating a weighted average that reflects the relative importance of verbal and non-verbal cues in diagnosing depression. This approach aims to enhance the overall diagnostic accuracy by leveraging the strengths of both data types.

Sub Objective 2: Design and Implement a Multi-modal Neural Network

The second sub-objective is to design and implement a Multi-modal Neural Network that integrates both text and image data within a unified architecture. This neural network will be trained to learn the combined features of verbal and non-verbal cues, allowing for a more strongly matches analysis of the patient's mental state. The network will be capable of not only detecting depression but also categorizing the severity of the condition.

Sub Objective 3: Compare the Performance of the Two Models

The third sub-objective involves a comprehensive comparison of the Weighted Average model and the Multi-modal Neural Network. The performance of each model will be evaluated based on various metrics, such as accuracy, precision, recall, and F1-score. This comparison will identify which approach provides the most reliable and accurate diagnosis of depression, considering both the detection and severity classification aspects.

Sub Objective 4: Integrate Explainable AI Techniques

The fourth sub-objective is to incorporate Explainable AI techniques into both models. This will involve developing methods that make the decision-making process of the models transparent and interpretable. By highlighting the key features and data points that influence the depression diagnosis, these techniques will provide clinicians with insights into how the models arrive at their conclusions, thereby increasing trust and facilitating the adoption of the system in clinical practice.

Sub Objective 5: Develop a Mobile Application for User Input Collection

The fifth sub-objective is to develop a mobile application that will engage users to gather real-time data on their current moods and emotions through text and facial expressions. By asking users how they feel at the moment and so on, the facial images will capture each and every response of patient. The application will collect essential input data, which will be processed by the Weighted Average and Multi-modal Neural Network models for depression diagnosis.

Sub Objective 6: Provide Progress and Current Situation Reports to Doctors

The sixth sub-objective is to develop a system that regularly updates doctors on the progress and current situation of patients based on the analysis performed by the models. This system will generate detailed reports that include both the current depression severity and trends over time, offering clinicians a comprehensive view of the patient's mental health status. These reports will assist doctors in making informed decisions about treatment adjustments and interventions.

Sub Objective 7: Determine the Optimal Model for Clinical Use

The final sub-objective is to determine which of the two models is most suitable for clinical implementation. Based on comparative analysis and clinical evaluation, the study will conclude by recommending the optimal approach for diagnosing depression. This recommendation will consider both diagnostic performance and ease of interpretation, ensuring that the chosen model can be effectively integrated into routine mental health care.

6. REQUIREMENT GATHERING AND FEASIBILITY STUDY

6.1 Requirement Gathering and Analysis

In the Requirement Gathering and Analysis phase of developing a depression diagnosis and management system, a systematic approach is employed to ensure the final product effectively meets both clinical and patient needs. This process begins with identifying and engaging key stakeholders, including general practitioners (GPs), mental health professionals, and patients, to understand their specific requirements and challenges. Techniques such as one-on-one interviews, focus groups, and surveys are utilized to collect comprehensive insights into the functionalities and features desired in the system. For instance, GPs may emphasize the need for an intuitive interface that integrates seamlessly with existing electronic health records, while patients might prioritize user-friendly interactions that encourage regular engagement. Observational studies are also conducted to analyze current workflows in clinical settings, identifying areas where the proposed system can enhance efficiency and accuracy in diagnosing depression. The gathered information is meticulously documented and analyzed to define clear, actionable requirements, distinguishing between functional requirements.

6.1.1 Functional requirements

- The application must enable patients to input textual data and capture facial images through an intuitive mobile interface.
- The system must analyze textual inputs using a Bidirectional Long Short-Term Memory (BiLSTM) network to detect linguistic markers indicative of depression.
- The system must process facial images using a Convolutional Neural Network (CNN) model, such as VGG19 or ResNet, to classify emotional states relevant to depression assessment.
- The application must integrate the outputs from text and image analyses to provide a comprehensive evaluation of the patient's mental state.
- The system must deliver a depression level assessment based on the combined analysis of text and image inputs.
- The application must present analysis results to healthcare professionals via a user-friendly dashboard, summarizing key insights and depression severity levels.
- The system must incorporate Explainable AI (XAI) features to elucidate the reasoning behind its predictions, aiding clinicians in understanding the basis of the assessments.
- The application must allow healthcare providers to monitor patient progress over time by securely storing and comparing historical assessments.

- The system must support real-time processing and provide immediate feedback to patient inputs for timely depression analysis.

6.1.2 Non-functional requirements

• **Reliability**: The system must be highly reliable as it assists in diagnosing mental health conditions. To ensure reliability, the models' outputs will be validated through extensive testing and expert review by clinical professionals.

• **Security**: The system will handle sensitive patient data, including text input and facial images. Therefore, strong security measures must be implemented, such as encryption, secure user authentication.

• **Availability**: The system should be available 24/7 to allow doctors to access patient data and provide timely diagnoses without disruptions. The system's availability will be verified through rigorous uptime testing before deployment.

• **Usability**: The application must be designed to be intuitive and easy to navigate, allowing doctors to access patient data and analysis results efficiently. The user interface should include clear navigation, a well-organized dashboard, and responsive design to enhance usability.

• **Scalability**: The system must be scalable to handle an increasing number of users, patients, and data inputs as it grows. Cloud infrastructure, such as AWS services, should be utilized to ensure that the system can scale efficiently with demand.

• **Performance**: The system must deliver fast and responsive performance, ensuring minimal latency in data processing, model inference, and dashboard loading times. Regular performance testing will be conducted to ensure the system meets these standards.

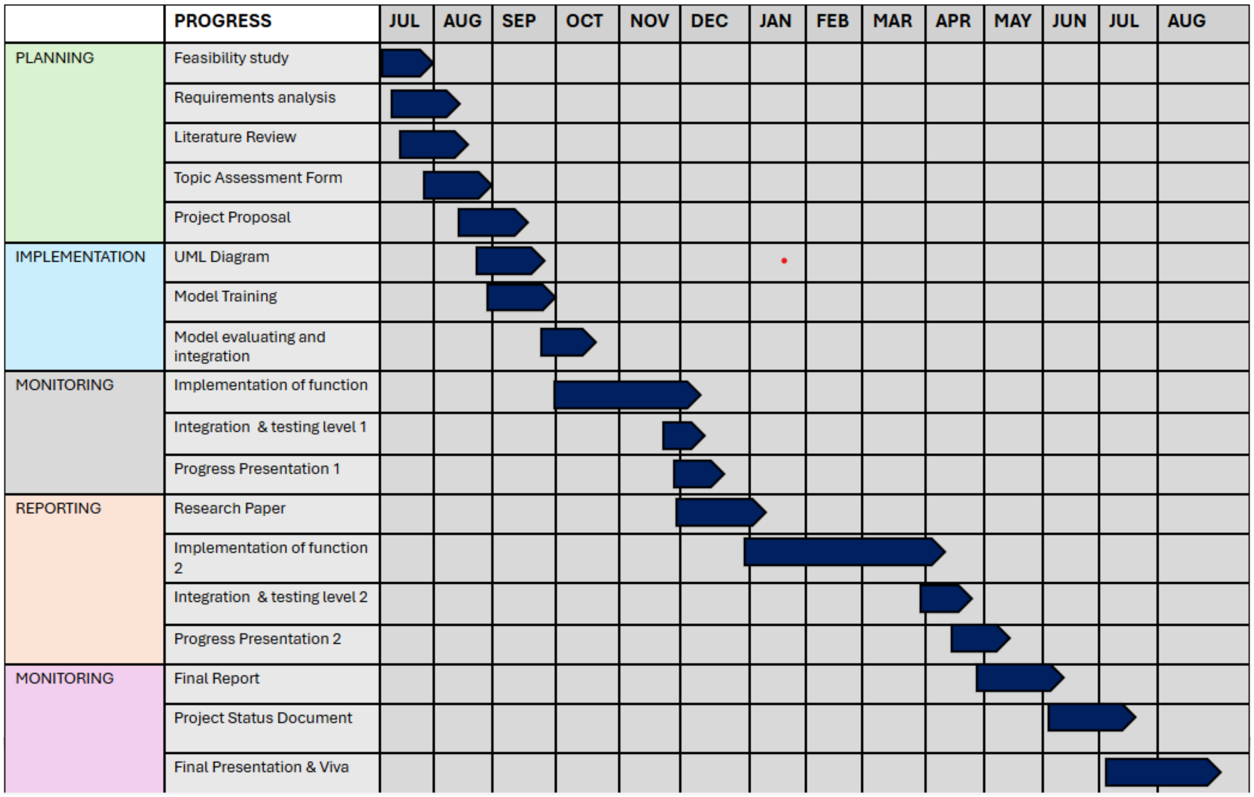
• **Maintainability**: The system should be modular and well-documented to ensure ease of maintenance and updates. Developers should be able to easily modify and upgrade components of the system without disrupting overall functionality.

6.2 Feasibility Study

The proposed system aims to assist general practitioners (GPs) in diagnosing and managing depression by integrating patient-generated textual inputs and facial imagery. This feasibility study evaluates the technical, operational, and economic viability of developing and implementing such a system.

6.3 Schedule feasibility

Schedule feasibility assesses whether a proposed project can be completed within a defined timeframe, considering factors such as project scope, resource availability, and potential constraints. For the development of the Depression Diagnosis and Management System, this evaluation is crucial to ensure timely delivery and integration into clinical settings.

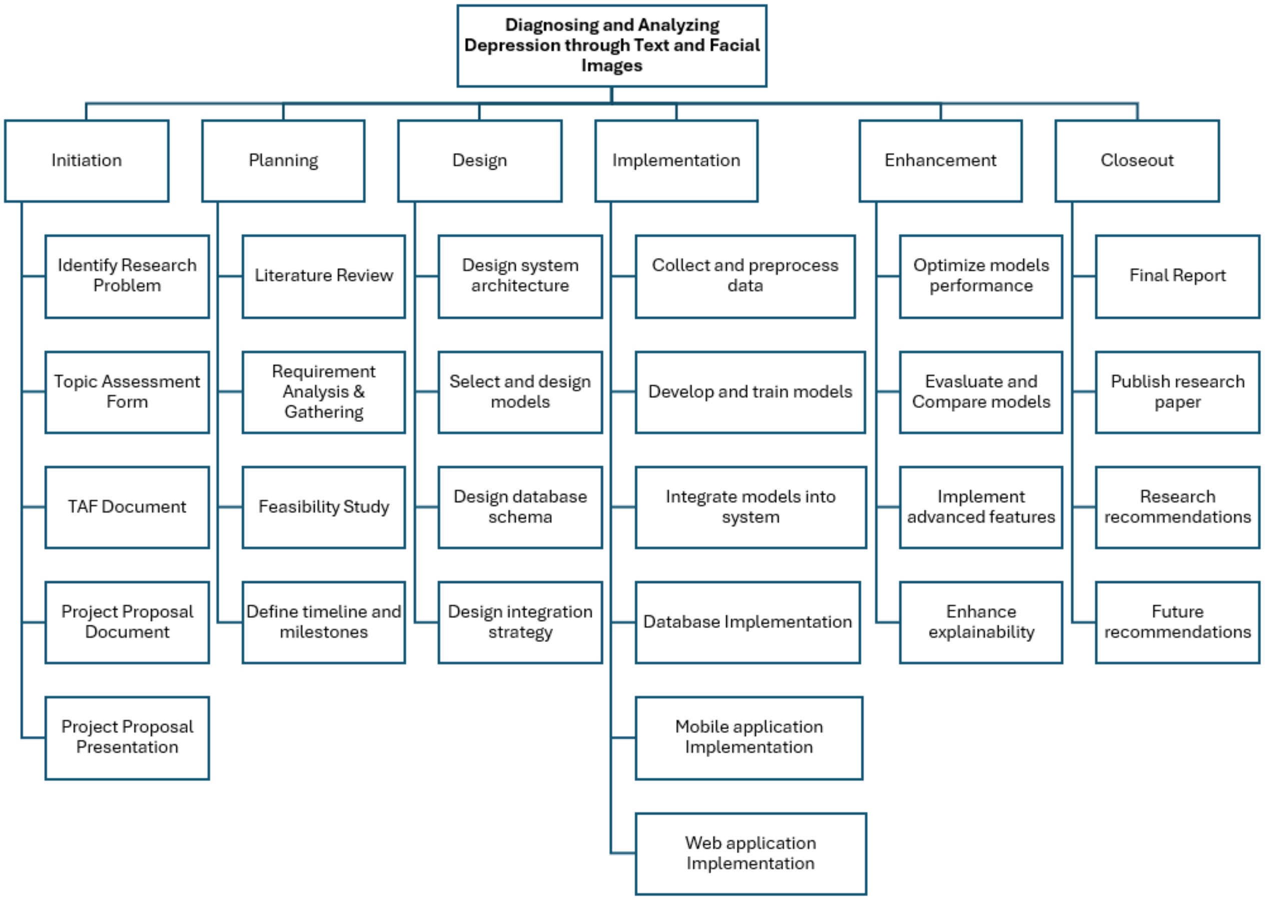


	PROGRESS	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG
PLANNING	Feasibility study	■													
	Requirements analysis	■	■												
	Literature Review	■	■												
	Topic Assessment Form	■	■												
	Project Proposal		■	■											
IMPLEMENTATION	UML Diagram		■	■											
	Model Training			■											
	Model evaluating and integration			■	■										
MONITORING	Implementation of function				■	■	■								
	Integration & testing level 1					■	■								
	Progress Presentation 1						■								
REPORTING	Research Paper					■	■	■							
	Implementation of function 2							■	■	■	■				
	Integration & testing level 2										■				
	Progress Presentation 2										■	■			
MONITORING	Final Report										■	■	■		
	Project Status Document												■	■	
	Final Presentation & Viva													■	■

Fig 1. Gannt Chart

6.4 Technical feasibility

The technical feasibility of the proposed depression detection system was evaluated by examining the necessary technologies, including natural language processing (NLP) algorithms, convolutional neural networks (CNNs), and image processing techniques. The system integrates machine learning models for text analysis and deep learning models for facial expression recognition to ensure accurate and efficient depression detection. Furthermore, the software and hardware requirements, such as computational resources, camera specifications, and data processing capabilities, were meticulously assessed to guarantee smooth implementation and scalability.



Diagnosing and Analyzing Depression through Text and Facial Images
Initiation
Identify Research Problem
Topic Assessment Form
TAF Document
Project Proposal Document
Project Proposal Presentation
Planning
Literature Review
Requirement Analysis & Gathering
Feasibility Study
Define timeline and milestones
Design
Design system architecture
Select and design models
Design database schema
Design integration strategy
Implementation
Collect and preprocess data
Develop and train models
Integrate models into system
Database Implementation
Mobile application Implementation
Web application Implementation
Enhancement
Optimize models performance
Evasluate and Compare models
Implement advanced features
Enhance explainability
Closeout
Final Report
Publish research paper
Research recommendations
Future recommendations

Fig 2. Work Breakdown Structure

6.5 Economic feasibility

The economic feasibility of the proposed Depression Detection System has been evaluated by analyzing the anticipated costs associated with its development and implementation. The budget allocation is detailed below:

TABLE 2. BUDGET ALLOCATION

Component	**Est. Amount in USD**	**Est. Amount in LKR**
Data Collection through open sources	6.77	2000.00
Charges for Tools Used for Research (Cloud Services, Grammarly, etc.)	31.37	10,000.00
Cloud Platforms (AWS, GCP, OpenAI key)	40.00	13,000.00
Total	96.14	**25,000.00**

The total estimated expenditure for the project is $78.14 (LKR 25,000.00). This budget encompasses data collection, research tools, and cloud platform services, ensuring the project's cost-effectiveness and financial viability. The modest investment required suggests that the project is economically feasible and can be undertaken without imposing significant financial strain on the organization.

6.6 Operational feasibility

6.6.1 User Engagement

Encouraging regular patient interaction with the system is crucial. User-friendly mobile interfaces and timely feedback can promote engagement, leading to more accurate monitoring. Continuous remote monitoring has been shown to be feasible in digital health applications.

6.6.2 Clinical Integration

The system's output should seamlessly integrate into existing clinical workflows, providing GPs with actionable insights without disrupting their routines. Dashboards summarizing key findings can facilitate this integration.

7. METHODOLOGY

To develop an effective system for diagnosing and managing depression through text and facial images, I will follow a systematic methodology encompassing data collection, model development, and evaluation phases.

7.1 Text Processing Section

In this section, a series of text preprocessing steps were carried out to ensure the quality and consistency of the textual data used for training and evaluation. The preprocessing began with tokenization, which involves breaking down raw text into individual tokens or words. This was followed by stopword removal to eliminate common but insignificant words such as "is", "the", and "in", which do not contribute meaningful information for classification tasks. Punctuation and special character removal was also applied to clean the text and reduce noise. Finally, TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was used to convert the cleaned text into numerical representations, capturing the importance of each word relative to the document and the corpus.

These steps are crucial for normalizing the text and minimizing irrelevant variations. By removing noise and focusing on significant features, the models were trained on more informative input, ultimately leading to better performance. Preprocessing plays a vital role in ensuring that the models can generalize well and make accurate predictions.

For the implementation, different approaches were employed based on the type of model. Tokenizer and pad_sequences() from TensorFlow/Keras were used for deep learning models to ensure uniform input lengths. Traditional machine learning models utilized TF-IDF Vectorizer, which is suitable for sparse, high-dimensional text data. Additionally, GloVe (Global Vectors for Word Representation) embeddings were used to obtain dense vector representations of words, capturing semantic relationships. This conversion of raw text into numerical form is essential, as machine learning models require numerical input to perform computations during training and inference.

7.2 Models Performance

Four different deep learning models were evaluated to identify the most effective architecture for text classification. The Bidirectional LSTM model achieved the highest accuracy of 0.96, indicating its strong ability to capture contextual dependencies in both forward and backward directions. This bidirectional approach allows the model to understand the full context of a sentence more effectively than a unidirectional LSTM.

The Convolutional Neural Network (CNN) model followed closely with an accuracy of 0.95. CNNs are particularly effective in extracting local patterns and features from the text, making them well-suited for classification tasks even though they are traditionally used in image processing.

The LSTM and GRU models, however, both achieved an accuracy of 0.50. These results suggest that without additional context awareness (as in bidirectional LSTMs) or advanced tuning, simple sequential models may struggle to capture the necessary patterns for accurate classification in this dataset. This comparison highlights the importance of model architecture and the advantages of using richer contextual models like Bidirectional LSTM and CNN for text-based depression detection.

As illustrated in Fig 3, the Bidirectional LSTM outperformed other models with a validation accuracy of 0.96, followed closely by the CNN with 0.95. In contrast, both the LSTM and GRU models achieved an accuracy of only 0.50, indicating limited performance in capturing the contextual patterns required for accurate classification.

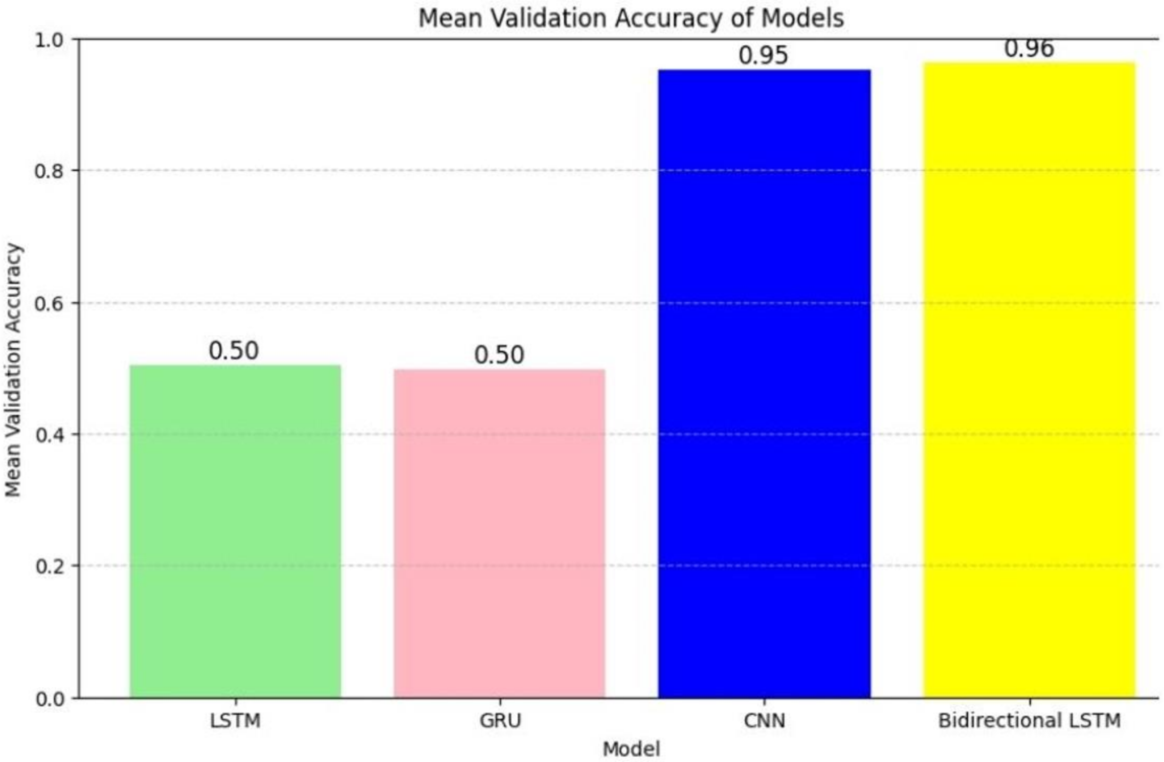


Fig 3. Text Model Comparison

The proposed text classification model is designed to detect depression from user-generated content. The model begins with integer-encoded input text, which is passed through a pre-trained embedding layer that transforms the input into dense vector representations. These embeddings are then fed into a Bidirectional Long Short-Term Memory (BiLSTM) layer, which captures contextual dependencies in both forward and backward directions. The output of the BiLSTM is processed by an attention layer that assigns varying levels of importance to different hidden states, generating a context vector that highlights the most relevant features for classification. This context vector passes through a dropout layer to prevent overfitting, followed by a dense layer that performs binary classification using a sigmoid activation function. The final output is a predicted probability score indicating whether the input text suggests depressive content (1) or not (0).

7.3 Image processing Section

To ensure consistent and high-quality inputs for the classification task, a series of image preprocessing steps were applied. All images were resized to 224x224 pixels, aligning with the input requirements of most standard CNN architectures. To enhance the clarity of facial features, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied, which is particularly useful in enhancing low-light and low-contrast images. Next, Gaussian Blur was used to suppress unwanted noise while preserving essential features such as facial outlines. Finally, the images were converted to grayscale, reducing the dimensionality of the data and computational cost, while retaining vital intensity-based information necessary for emotion detection.

7.4 Feature Extraction

To enhance the model’s ability to accurately detect and classify facial expressions, multiple feature extraction techniques were applied. These methods aim to highlight structural, textural, and edge-based characteristics of the facial images, ensuring that the deep learning models focus on relevant patterns.

The Canny edge detection technique was employed to identify and emphasize prominent edges and contours within facial images. This is particularly useful in emotion recognition, as facial expressions often involve changes in the outlines of the eyes, mouth, and eyebrows. Highlighting these regions helps the model learn structural patterns that are significant in differentiating emotions.

HOG was used to extract shape and texture-based features by computing gradient orientation histograms across the image. This technique captures the distribution of edge directions, which is crucial for representing subtle variations in facial expressions such as furrowing of brows or smiling. These features enhance the model’s ability to generalize over varying facial structures and lighting conditions.

LBP is a texture descriptor that labels the pixels of an image by thresholding the neighborhood of each pixel. This method captures fine-grained texture patterns that are essential for detecting micro-expressions and skin texture changes. It’s particularly effective in enhancing the model's sensitivity to small and localized changes in the face, which can be key indicators of emotion.

All the extracted features were visualized and validated to ensure they preserved meaningful facial structures. These feature-enhanced images were used as additional input references during training to support better pattern learning by the deep learning models.

Figure 4 showcases sample outputs from CLAHE, Canny Edge, HOG, and LBP transformations.

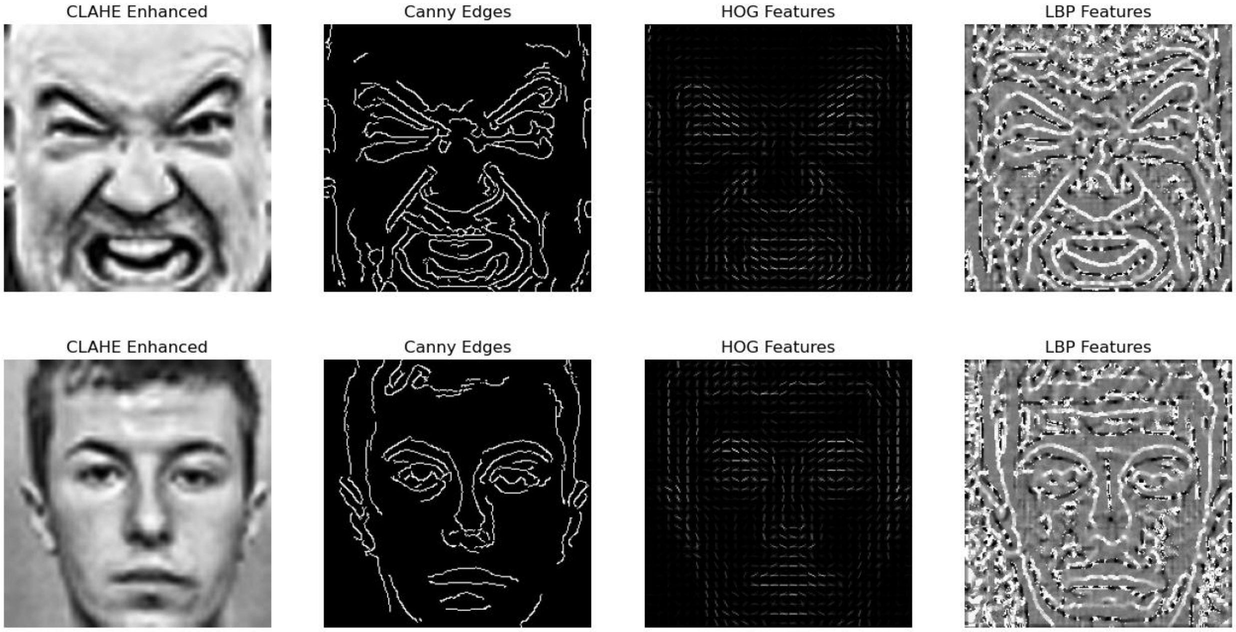


Fig 4. Feature Extractions Methods

7.5 Data Augmentation

To improve model generalization and prevent overfitting, data augmentation techniques were applied during training. These included rescaling, rotation, zooming, and horizontal flipping. By artificially increasing the variability of the training dataset, the model was exposed to different perspectives and distortions of the same image, enabling it to learn more robust and generalized features across a diverse set of inputs.

7.6 Model Architectures Evaluated

Multiple deep learning models were evaluated to identify the most effective architecture for emotion classification. These included VGG19 (pretrained and fine-tuned), a lightweight custom variant named VGG16, a basic Convolutional Neural Network (CNN) built from scratch, ResNet50 which leverages residual learning to handle deeper networks effectively, and EfficientNet, known for its balance between accuracy and computational efficiency through compound scaling. Each model was trained on the same preprocessed dataset and evaluated under identical conditions. This comprehensive multi-model evaluation enabled a robust comparison, ultimately helping to identify the architecture that achieved the best performance in terms of both accuracy and generalization across diverse facial expressions.

7.7 Final Model Architecture (VGG19-based)

Among the models tested, the VGG19-based architecture [10] demonstrated superior performance and was selected as the final model for emotion classification. This model leverages transfer learning, utilizing a pretrained VGG19 network as a fixed feature extractor, with its weights frozen to retain learned representations. Input facial images resized to 224x224 were passed through the

VGG19 layers, and the resulting feature maps were flattened before entering a series of custom fully connected layers. The first dense layer consisted of 256 neurons with ReLU activation, followed by Batch Normalization to stabilize learning and speed up convergence. A Dropout layer with a 50% rate was applied to reduce overfitting. This was followed by a second dense layer with 128 neurons and ReLU activation, again accompanied by Batch Normalization and a 30% Dropout layer. The final output layer included 7 neurons with Softmax activation, enabling classification across seven emotion categories. The complete architecture is visually represented in Figure 5.

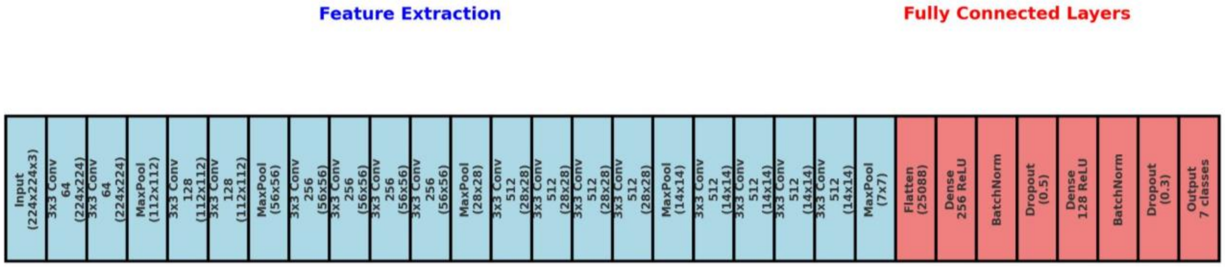
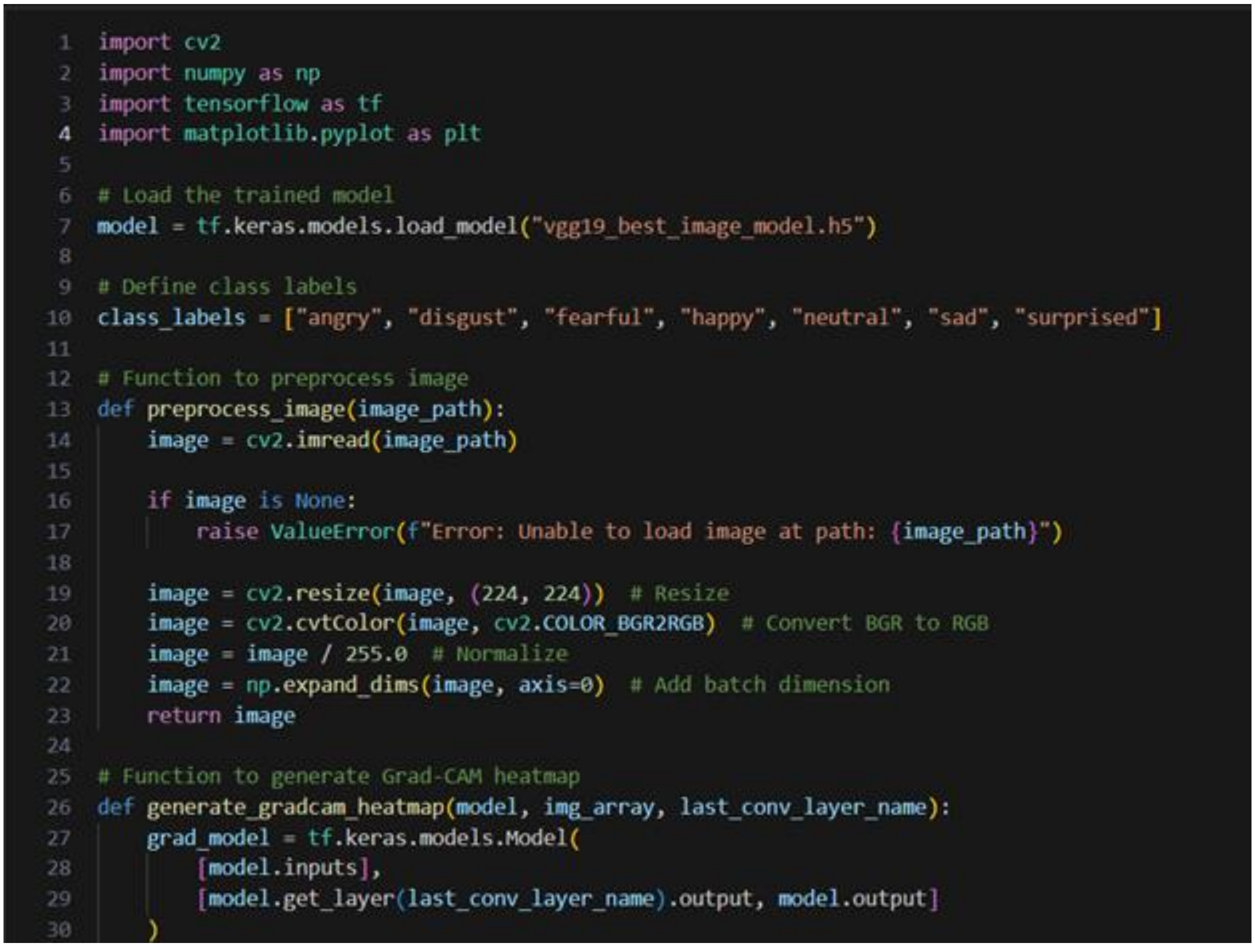


Fig 5. VGG19 Model Architecture

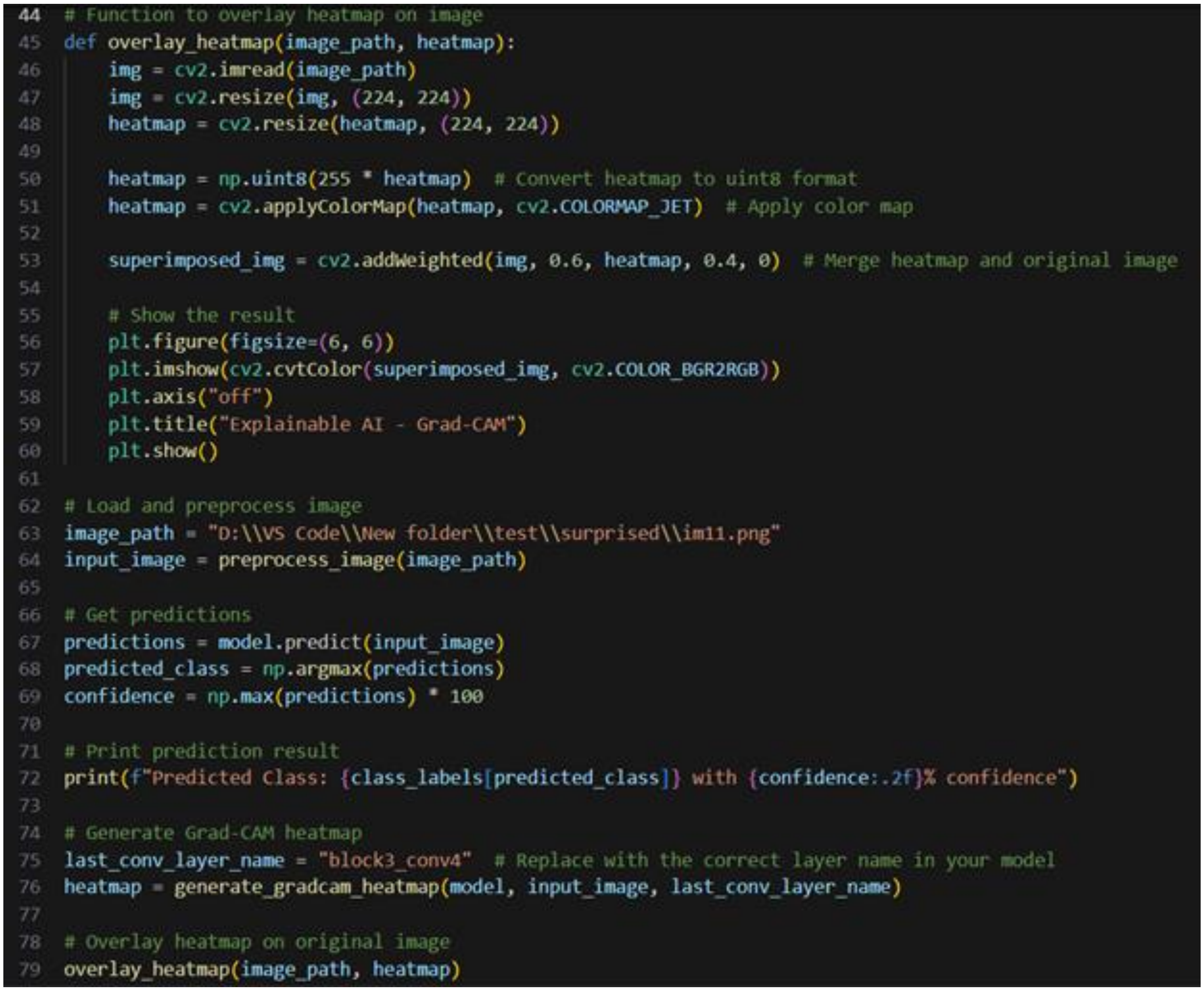
7.8 Explainable AI (XAI)

To ensure transparency and interpretability of the model's predictions, Explainable AI (XAI) techniques were integrated into the system. These methods help visualize which regions of the image the model focused on when predicting a particular emotion, allowing us to validate that the model is learning relevant facial features. Tools such as Grad-CAM (Gradient-weighted Class Activation Mapping) were used to generate heatmaps over facial regions, offering insights into decision-making. This interpretability is particularly crucial in medical or psychological applications, where understanding why a model makes a specific classification can aid in building trust with clinicians and support more informed diagnosis.



```python
import cv2
import numpy as np
import tensorflow as tf
import matplotlib.pyplot as plt

# Load the trained model
model = tf.keras.models.load_model("vgg19_best_image_model.h5")

# Define class labels
class_labels = ["angry", "disgust", "fearful", "happy", "neutral", "sad", "surprised"]

# Function to preprocess image
def preprocess_image(image_path):
    image = cv2.imread(image_path)

    if image is None:
        raise ValueError(f"Error: Unable to load image at path: {image_path}")

    image = cv2.resize(image, (224, 224))  # Resize
    image = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)  # Convert BGR to RGB
    image = image / 255.0  # Normalize
    image = np.expand_dims(image, axis=0)  # Add batch dimension
    return image

# Function to generate Grad-CAM heatmap
def generate_gradcam_heatmap(model, img_array, last_conv_layer_name):
    grad_model = tf.keras.models.Model(
        [model.inputs],
        [model.get_layer(last_conv_layer_name).output, model.output]
    )
```

Fig 6. Code Snippet for XAI



```python
# Function to overlay heatmap on image
def overlay_heatmap(image_path, heatmap):
    img = cv2.imread(image_path)
    img = cv2.resize(img, (224, 224))
    heatmap = cv2.resize(heatmap, (224, 224))

    heatmap = np.uint8(255 * heatmap)  # Convert heatmap to uint8 format
    heatmap = cv2.applyColorMap(heatmap, cv2.COLORMAP_JET)  # Apply color map

    superimposed_img = cv2.addWeighted(img, 0.6, heatmap, 0.4, 0)  # Merge heatmap and original image

    # Show the result
    plt.figure(figsize=(6, 6))
    plt.imshow(cv2.cvtColor(superimposed_img, cv2.COLOR_BGR2RGB))
    plt.axis("off")
    plt.title("Explainable AI - Grad-CAM")
    plt.show()

# Load and preprocess image
image_path = "D:\\VS Code\\New folder\\test\\surprised\\im11.png"
input_image = preprocess_image(image_path)

# Get predictions
predictions = model.predict(input_image)
predicted_class = np.argmax(predictions)
confidence = np.max(predictions) * 100

# Print prediction result
print(f"Predicted Class: {class_labels[predicted_class]} with {confidence:.2f}% confidence")

# Generate Grad-CAM heatmap
last_conv_layer_name = "block3_conv4"  # Replace with the correct layer name in your model
heatmap = generate_gradcam_heatmap(model, input_image, last_conv_layer_name)

# Overlay heatmap on original image
overlay_heatmap(image_path, heatmap)
```

Fig 7. Code Snippet for XAI

7.8 Proposed Models Fusion methods

The core of the system development lies in building two models: the text analysis model and the image analysis model. The text analysis model will employ Natural Language Processing (NLP) techniques, specifically using the Bi-LSTM model, to analyze the text data for signs of depression. In contrast, the image analysis model will utilize VGG19 classify facial expressions into emotional categories. These models will work in collaboration to detect depression by analyzing both textual and visual cues.

To combine the predictions from these models, two multi-modal fusion approaches will be implemented. The first approach, the Weighted Average Method, involves generating predictions from both models and combining them using a weighted average to derive a final depression score. The weights will be determined based on the performance and relevance of each model. (w_{text} and w_{image} are the weights for the text and image predictions, respectively)

$$\textit{Final Score} = (w_{text} \times \textit{Text Prediction}) + (w_{image} \times \textit{Image Prediction})$$

The second, more advanced approach, is a Multi-Modal Neural Network that integrates text and image features into a single model with distinct branches for each modality. These branches will be fused at a later stage to produce a unified prediction, with an attention mechanism incorporated to dynamically focus on the most relevant features from both modalities, enhancing the model's performance.

An attention mechanism will play an important role in the multi-modal neural network by helping the model focus on significant aspects of both text and image data, thereby improving the accuracy of depression detection. This mechanism will assign different attention weights to various parts of the input data, allowing the model to better understand and integrate information from both modalities.

To ensure the system's transparency and trustworthiness, Explainable AI techniques will be implemented. These techniques will generate explanations for the model’s classifications, such as highlighting important features or decision pathways that led to a particular depression level assessment. By providing these explanations, clinicians will gain a better understanding of the rationale behind the model’s predictions, enabling them to make more informed decisions in the diagnosis and treatment of patients.

The models and fusion methods will be evaluated using a combination of performance metrics and clinical expert feedback. Metrics such as accuracy, precision, recall, F1 score, and AUC-ROC will assess the performance of the models, while clinical experts will review the model outputs to validate the accuracy and relevance of the predictions. Feedback from these experts will be used to refine the models and enhance their clinical applicability.

Upon successful evaluation, the models will be integrated into a mobile application designed for real-time mood and emotion capture. This application will allow patients to input text and images, which will be analyzed by the models to provide depression level assessments. The results will be summarized and presented to doctors, assisting them in diagnosing and managing depression more effectively.

In conclusion, this methodology aims to create a comprehensive system for depression diagnosis that leverages both textual and visual data. By combining multi-modal approaches with advanced techniques like attention mechanisms and Explainable AI, the system strives to offer a robust and clinically useful tool for mental health professionals. Through iterative development and evaluation, this approach ensures that the system will be effective and reliable in real-world scenarios.

Figure 8 will represent the proposed multi-modal function approaches.

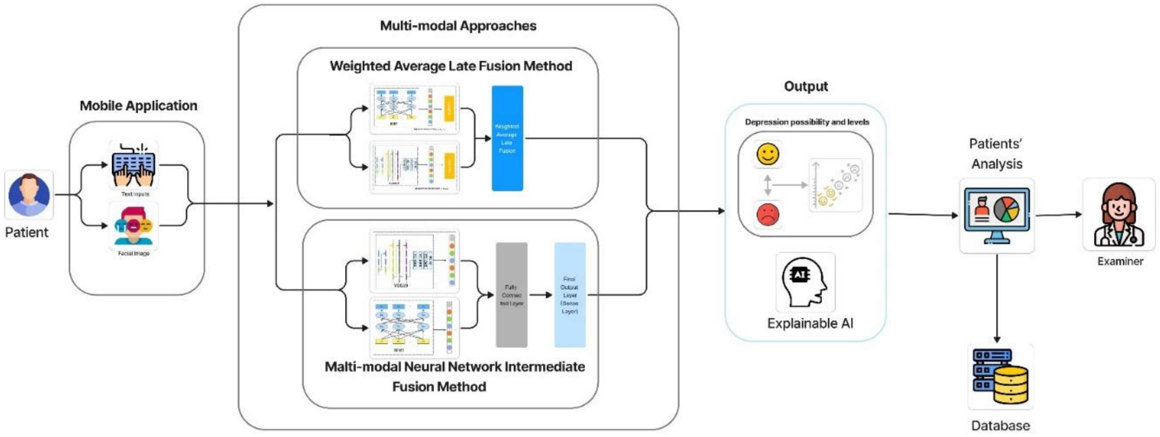

Multi-modal Approaches
Weighted Average Late Fusion Method
BERT
VGG19
Weighted Average Late Fusion
Mobile Application
Text Inputs
Facial Image
Patient
VGG19
BERT
Fully Connected Layer
Final Output Layer (Dense Layer)
Malti-modal Neural Network Intermediate Fusion Method
Output
Depression possibility and levels
Explainable AI
Patients' Analysis
Examiner
Database

Fig 8.Proposed Multi-modal Methods

It is important to note that this methodology is still in active development, and architectural components, fusion strategies, and evaluation methods may evolve as the project progresses. The iterative nature of this research allows for continuous improvement based on model performance, user feedback, and clinical input, ensuring the final system is both robust and clinically viable.

8. RESULTS & DISCUSSION

8.1 Data Collection and Preprocessing

For this research, two distinct datasets, one textual and one image-based were obtained from Kaggle to facilitate the development of a multi-modal approach for depression diagnosis. The textual dataset consists of labeled as either "depressed" or "not depressed," while the image dataset consists of facial expressions categorized into seven emotional states which were sad, surprised, neutral, happy, fearful, disgust, and angry. These datasets were carefully selected to ensure diversity and representativeness, enabling a robust analysis of depression indicators across different modalities. The dataset quantities are summarized in Table 1 and Table 2.

For this research, two distinct datasets, one textual and one image-based were obtained from Kaggle to facilitate the development of a multi-modal approach for depression diagnosis. The textual dataset consists of labeled as either "depressed" or "not depressed," while the image dataset consists of facial expressions categorized into seven emotional states which were sad, surprised, neutral, happy, fearful, disgust, and angry. These datasets were carefully selected to ensure diversity and representativeness, enabling a robust analysis of depression indicators across different modalities. The dataset quantities are summarized in Table 1 and Table 2.

TABLE 3. TEXT DATASET DETAILS

Label	No of Tweets
Depressed	3832
Not Depressed	3441

TABLE 4. IMAGE DATASET DETAILS

Emotion	No of Train Images	No of Test Images
Angry	3995	958
Disgusted	436	111
Fearful	4097	1024
Happy	7215	1774
Neutral	4965	1233
Sad	4830	1247
Surprised	3171	831

8.2 Text Processing

The objective of this study was to develop an effective deep learning-based classification system capable of detecting depression from text data. A dataset consisting of 7,731 Reddit posts labeled as either 'depressed' or 'not depressed' was preprocessed and used to train several deep learning architectures, including LSTM, GRU, CNN, and Bidirectional LSTM models. The dataset was relatively balanced, with 3,900 non-depressed and 3,831 depressed samples, ensuring fairness in model training and evaluation.

Initial preprocessing included extensive text cleaning using regular expressions and the removal of stopwords, followed by tokenization and padding to standardize input lengths. A 100-dimensional GloVe embedding was used to represent the text semantically for LSTM-based models, ensuring contextual information was preserved during training. The dataset was split with 80% for training and 20% for testing, while further 20% of the training data was reserved for validation during model fitting.

The performance of the Long Short-Term Memory (LSTM) model was found to be suboptimal, achieving an accuracy of 0.4939 and an AUC of 0.5039, which is close to random guessing. The classification report and confusion matrix indicate that the model failed to detect any non-depressed samples, resulting in zero precision and recall for class 0. This pattern was mirrored in the Gated Recurrent Unit (GRU) model, which also achieved similarly low performance with an accuracy of 0.4926 and an AUC of 0.5013. These results suggest that both LSTM and GRU models overfit the dominant class or failed to generalize patterns from the data effectively, possibly due to inadequate architecture depth, dropout handling, or ineffective embedding integration.

In contrast, the Convolutional Neural Network (CNN) model significantly outperformed the recurrent architecture, achieving an accuracy of 0.9632 and an AUC of 0.9802. The CNN model demonstrated strong generalization capabilities and was effective in learning local patterns in the text, which are critical in sentiment and emotion detection tasks. The confusion matrix revealed high true positive and true negative rates, with the model correctly classifying most samples from both classes.

The Bidirectional LSTM (BiLSTM) model also yielded outstanding results, with an accuracy of 0.9644 and an AUC of 0.9802, slightly outperforming the CNN. The ability of the BiLSTM to consider context from both directions likely contributed to its superior performance, capturing dependencies that unidirectional models may miss. Both precision and recall values for depressed and non-depressed classes were above 0.95, indicating a balanced and reliable classifier. Model performances comparison as mentioned below in the figure.

TABLE 5. MODEL METRICS COMPARISON

Model	Accuracy	AUC	Precision	Recall
LSTM	0.494	0.504	0.240	0.490
GRU	0.493	0.501	0.240	0.490
CNN	0.963	0.980	0.960	0.960
Bi-LSTM	0.964	0.980	0.960	0.960

Sample predictions further support the model's interpretability and practical value. For instance, the sentence *"I feel depressed"* yielded a high probability score of 0.55 and was correctly classified as depressed. Conversely, a positive statement like *"I am happy and excited about the future."* was predicted as not depressed with a probability close to zero, showing the model’s ability to distinguish emotional polarity effectively.

8.3 Bi-LSTM Model Evaluation

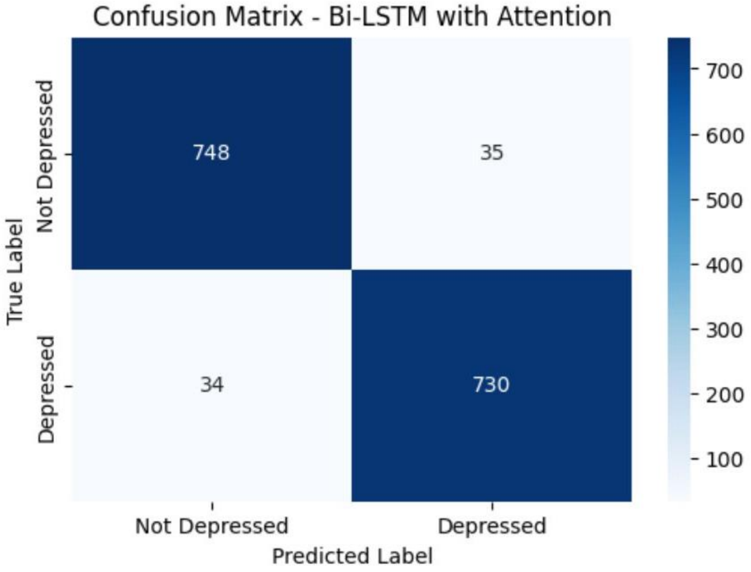


Fig 9. Confusion Matrix of Bi-LSTM with Attention Mechanism

The confusion matrix presented above offers a clear insight into the classification performance of the Bi-LSTM model enhanced with an attention mechanism. The model correctly identified 748 out of 783 instances labeled as *Not Depressed*, and 730 out of 764 instances labeled as *Depressed*. This reflects a high level of accuracy in both positive and negative class predictions, showcasing the model's ability to differentiate between depressed and non-depressed textual inputs effectively.

Only 35 non-depressed instances were misclassified as depressed (false positives), and 34 depressed instances were misclassified as not depressed (false negatives). These low numbers indicate that the model maintains a good balance between sensitivity (recall) and specificity, making it a reliable choice for mental health detection tasks using text. The near-equal misclassification rates in both classes suggest that the model does not exhibit significant bias toward either category, which is a desirable characteristic in psychological or clinical applications where fairness and reliability are critical.

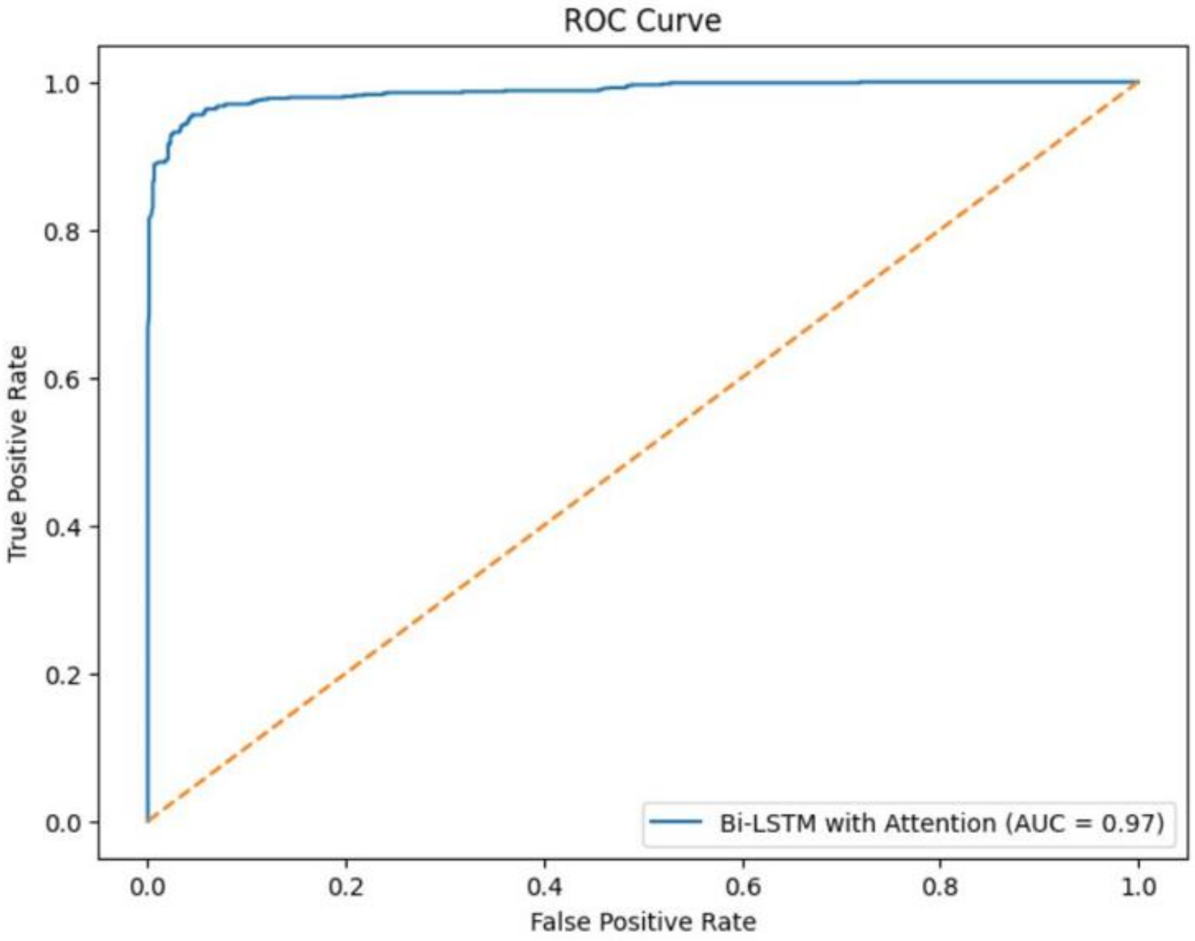
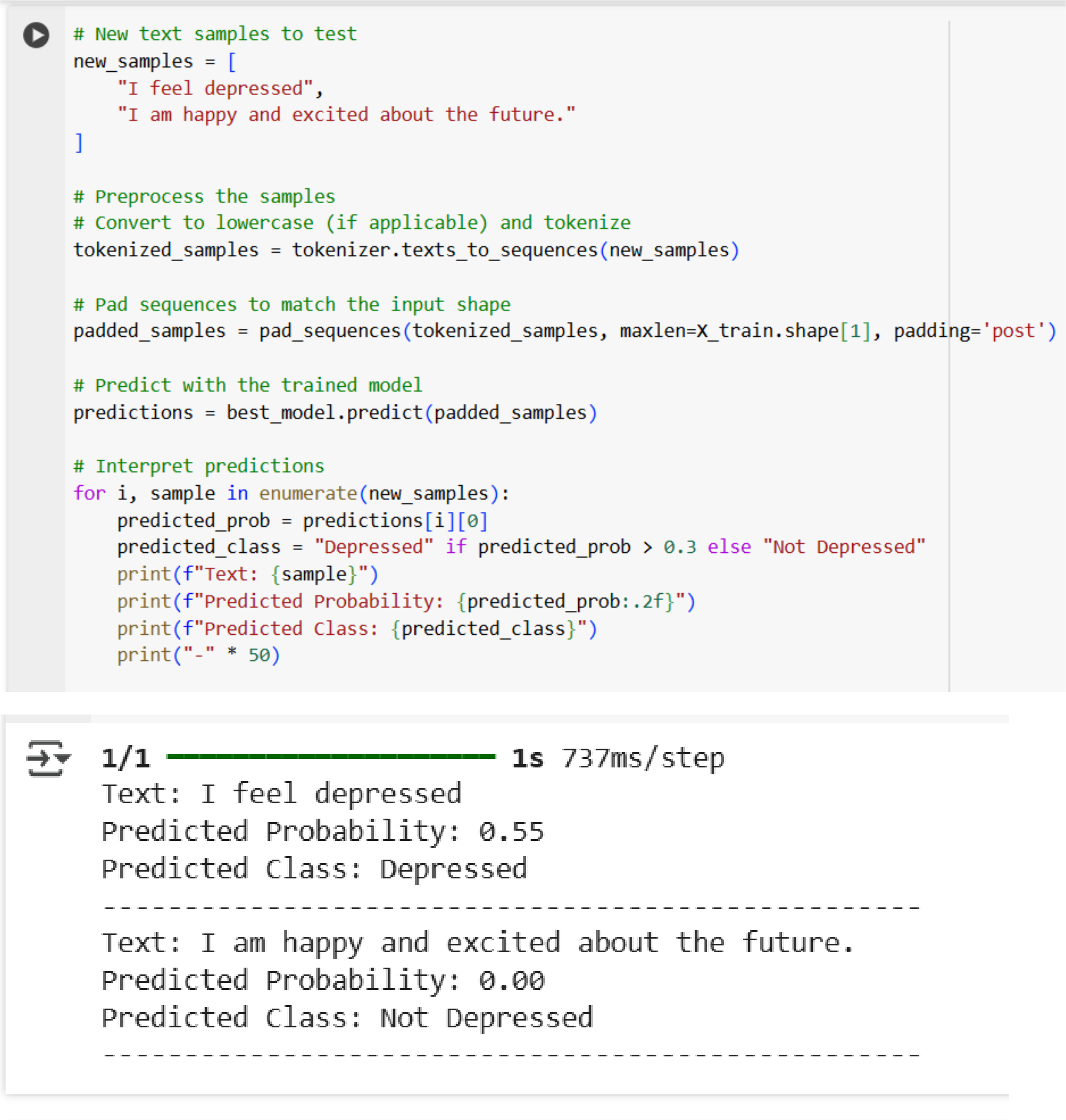


Fig 10. ROC Curve of Bi-LSTM Model Performance

The Receiver Operating Characteristic (ROC) curve provides a visual representation of the Bi-LSTM with Attention model’s classification performance across various threshold levels. The curve shown above rises steeply towards the top-left corner, indicating a high true positive rate (sensitivity) with a low false positive rate, which reflects the model's excellent ability to distinguish between depressed and not depressed individuals. The Area Under the Curve (AUC) is recorded at 0.97, which is remarkably high and signifies a near-perfect classifier.

An AUC value of 0.97 means that there is a 97% chance that the model will correctly differentiate a randomly chosen depressed instance from a non-depressed one. This strong performance metric highlights the robustness of incorporating attention mechanisms within the Bi-LSTM architecture, which helps the model focus on the most emotionally relevant parts of the input text. Additionally, the significant gap between the ROC curve and the diagonal line (representing random guessing) confirms that the model performs far better than chance.

Overall, the findings from the confusion matrix demonstrate that the integration of attention mechanisms into the Bi-LSTM architecture has positively contributed to identifying subtle linguistic cues associated with depression. This reinforces the effectiveness of deep learning approaches in understanding emotional context from text data and highlights the potential of such models in supporting early mental health assessments.



```python
# New text samples to test
new_samples = [
    "I feel depressed",
    "I am happy and excited about the future."
]

# Preprocess the samples
# Convert to lowercase (if applicable) and tokenize
tokenized_samples = tokenizer.texts_to_sequences(new_samples)

# Pad sequences to match the input shape
padded_samples = pad_sequences(tokenized_samples, maxlen=X_train.shape[1], padding='post')

# Predict with the trained model
predictions = best_model.predict(padded_samples)

# Interpret predictions
for i, sample in enumerate(new_samples):
    predicted_prob = predictions[i][0]
    predicted_class = "Depressed" if predicted_prob > 0.3 else "Not Depressed"
    print(f"Text: {sample}")
    print(f"Predicted Probability: {predicted_prob:.2f}")
    print(f"Predicted Class: {predicted_class}")
    print("-" * 50)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 1s 737ms/step
Text: I feel depressed
Predicted Probability: 0.55
Predicted Class: Depressed
--------------------------------------------------
Text: I am happy and excited about the future.
Predicted Probability: 0.00
Predicted Class: Not Depressed
--------------------------------------------------
```

Fig 11. Output Sample of Text Model

In text model building summary, while recurrent models like LSTM and GRU struggled with binary depression classification, the CNN and BiLSTM models showcased impressive results, highlighting their effectiveness in handling linguistic patterns for emotion and sentiment recognition. The findings suggest that for tasks involving short-text classification with emotional cues, models that emphasize local feature extraction (like CNNs) or bidirectional temporal context (like BiLSTMs) may be more appropriate than traditional sequential RNNs. Future work can explore hybrid architectures or attention mechanisms to further improve the interpretability and robustness of the system in real-world applications.

8.1 Image Preprocessing and Feature Extraction

The image processing pipeline involved multiple enhancement and feature extraction techniques to better represent facial expressions. The initial grayscale images were resized to 224×224 and enhanced using CLAHE (Contrast Limited Adaptive Histogram Equalization), improving local contrast and aiding edge detection. Gaussian blur helped reduce noise, followed by Canny edge detection to highlight facial structure. Additionally, advanced feature descriptors such as Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP) were extracted. These descriptors played a crucial role in visualizing texture and edge information, highlighting facial muscle movements that are essential for emotion classification.

8.2 Image processing model creation

In the task of emotion classification from facial images, a crucial component of identifying signs of depression. Three deep learning models were trained and evaluated: VGG19, a custom Convolutional Neural Network (CNN), and ResNet50. Among these, VGG19 emerged as the most effective model, achieving a training accuracy of 77.00% and a validation accuracy of 49.89%. While the noticeable gap between training and validation accuracy suggests a degree of overfitting, VGG19 still managed to generalize better than the other models. Its deeper architecture, consisting of 19 layers, enables it to learn more complex hierarchical features from facial images, particularly subtle patterns in the eyes, mouth, and forehead regions that are often associated with depressive emotions such as sadness, fear, and lack of expression. The model's ability to capture these fine-grained visual cues makes it a promising candidate for emotion recognition tasks in mental health applications, where precision in detecting subtle changes in expression is critical.

The custom CNN model, on the other hand, achieved a training accuracy of 37.94% and a validation accuracy of 39.80%. Although these values are considerably lower than those of VGG19, the close proximity of the training and validation scores indicates a better generalization with minimal overfitting. However, due to its relatively shallow architecture and fewer layers, the model was limited in its ability to capture deeper, more abstract features from the facial images. As a result, its performance in identifying complex emotional states, especially those related to depression was restricted.

The third model, ResNet50, attained the lowest performance, with a training accuracy of 30.81% and a validation accuracy of 29.48%. Despite being a very powerful and widely used architecture in many computer vision tasks, ResNet50's poor performance in this case could be attributed to its complexity and depth, which require significantly more data and training time to converge effectively. The model might have been too deep for the size or nature of the dataset used, leading to underfitting and a failure to capture key emotional features from the facial images.

TABLE 6. IMAGE MODEL PERFORMANCE COMPARISON

Model	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss
VGG19	0.7700	0.4989	0.6251	1.8012
CNN	0.3794	0.3980	1.5682	1.5251
RestNet50	0.3081	0.2948	1.7176	1.6939

Therefore, VGG19's architecture strikes a balance between depth and learnability, making it particularly suitable for detecting depression-related emotions in facial expressions. Its layered feature extraction capability allows it to identify nuanced patterns that simpler models might miss, while being more manageable to train than deeper models like ResNet50. Therefore, among the three tested models, VGG19 is best suited for this emotion recognition task, especially in applications aimed at supporting clinical assessments of depression.

8.3 VGG19 Model Results review

Despite VGG19 achieving the highest overall validation accuracy among the tested models, a closer examination of its classification report reveals several critical limitations, especially in its per-class performance. Emotions such as Disgust, Surprise, and Angry were particularly challenging for the model, with precision, recall, and F1-scores all falling near or below 0.10. This indicates that the model frequently misclassified these emotions, either by confusing them with other classes or failing to detect them altogether. Such low performance suggests that while the model may be reasonably good at identifying certain dominant emotional patterns, it lacks the nuanced understanding required to distinguish between less frequent or visually similar emotions.

On the other hand, emotions like Sad, Fear, and Happy showed relatively better results, although even these categories did not reach optimal performance levels. These emotions tend to have more distinguishable facial features and were likely more prevalent in the training dataset, which could explain the modest improvement in their classification metrics. However, the overall low scores across most categories point to the model's difficulty in handling class imbalance—a common issue in emotion recognition datasets where certain emotions appear far more frequently than others. Additionally, the semantic and visual similarities between some emotions (e.g., Fear vs. Surprise or Angry vs. Disgust) may have led to confusion during prediction, further degrading the model's precision.

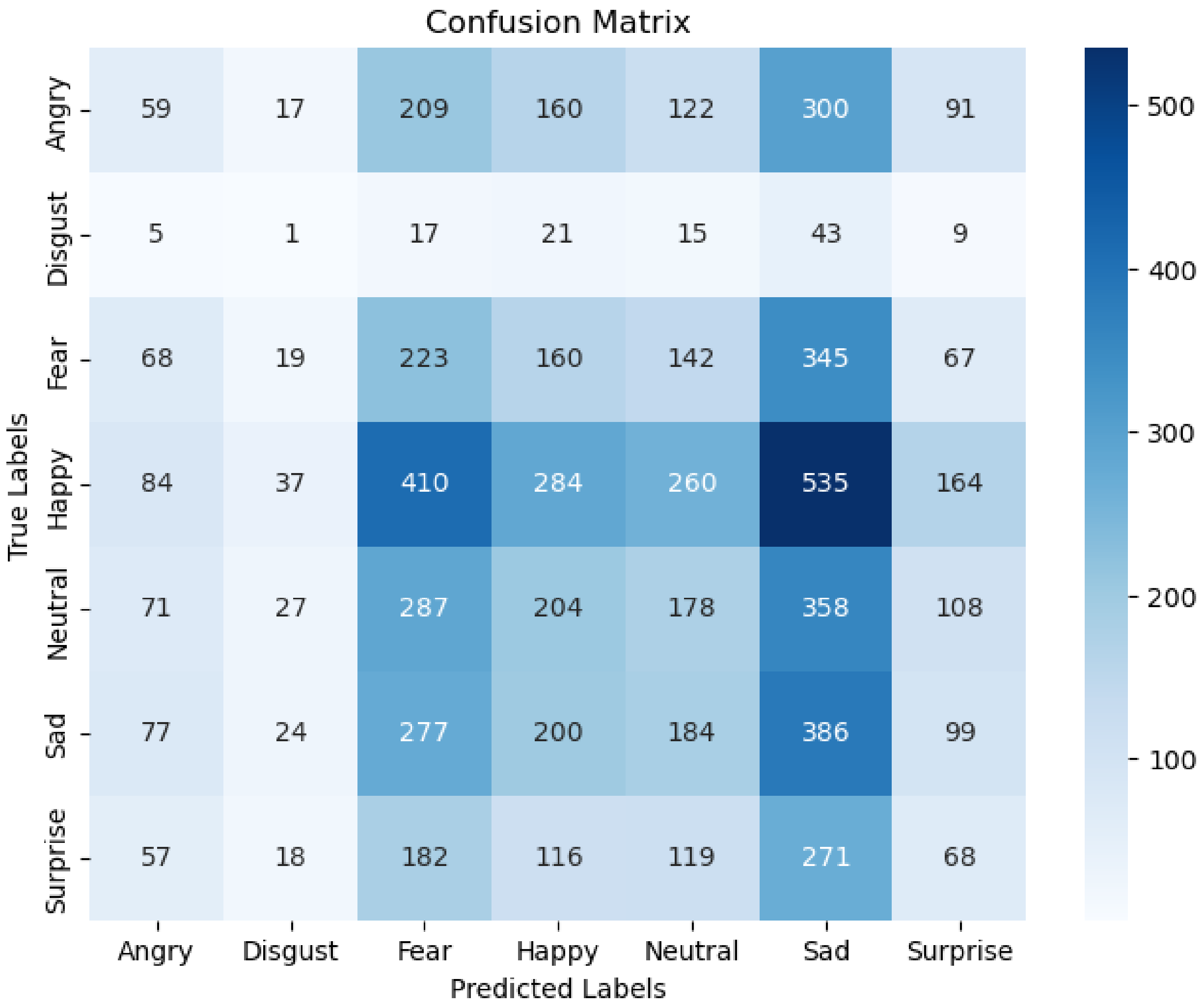


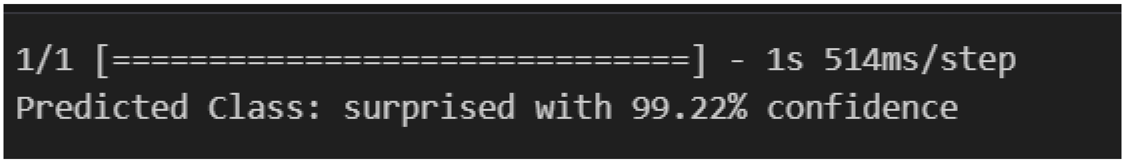
Fig 12. Confusion Matrix for Emotion Labels

8.4 Explainable AI with Grad-CAM

To enhance the interpretability of the VGG19 model's decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was utilized to visualize the regions of facial images that the model focused on during emotion prediction. Grad-CAM generates class-discriminative heatmaps by computing the gradients of the target class flowing into the final convolutional layers, effectively highlighting the most influential areas of the image that contribute to the model's classification decision.

The generated heatmaps provided valuable insights into the internal workings of the model. In most cases, the model consistently attends to key expressive facial regions, including the eyes, mouth, eyebrows, and forehead. These areas are well-established in psychological and facial behavior research as the most expressive parts of the human face, often revealing subtle emotional cues such as tension, relaxation, and asymmetry. For instance, the mouth shape can differentiate between happiness and disgust, while the eye region can reflect emotions like fear or surprise. The fact that the model naturally focused on these regions indicates that it learned to extract meaningful and semantically relevant visual features, rather than relying on arbitrary or background patterns.

This visual confirmation through Grad-CAM not only increases trust in the model's decisions but also serves as a validation of the feature learning process, especially in a sensitive domain like emotion or depression detection. Explainability tools like Grad-CAM are essential when deploying deep learning models in healthcare-related applications, as they offer transparency and interpretability, helping researchers and clinicians understand why a model arrived at a certain conclusion. This is particularly important when dealing with complex emotions that may manifest subtly and vary across individuals.



```
1/1 [==============================] - 1s 514ms/step
Predicted Class: surprised with 99.22% confidence
```

Fig 13. Sample VGG19 Model Output with XAI

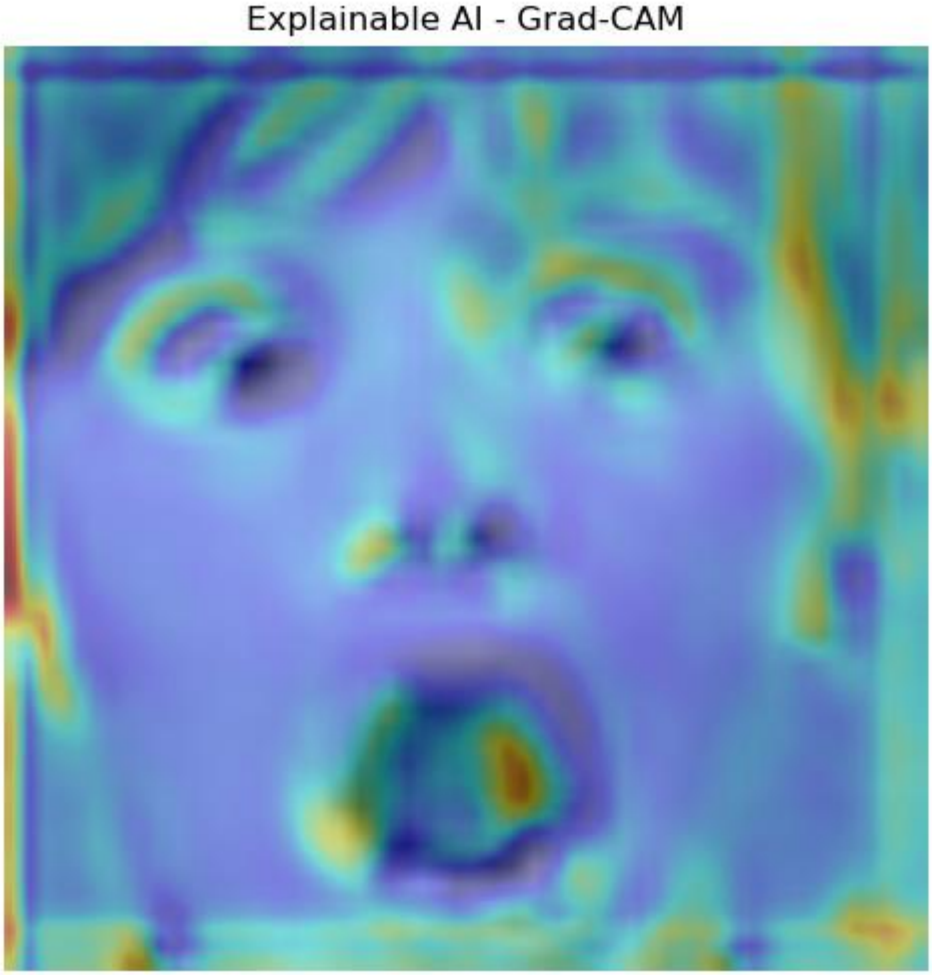


Fig 14. Sample VGG19 Model Output with XAI

Grad-CAM helps identify potential biases or inconsistencies in the model's focus. For example, if the model were found to rely on irrelevant background elements or non-expressive areas, it would suggest a need for dataset cleaning or architectural changes. However, in this case, the attention to facially relevant regions confirms that VGG19 was at least partially successful in learning discriminative patterns aligned with human perception, supporting its potential use as a foundational model in emotion and depression detection pipelines.

Moreover, explainability supports accountability by making it possible to audit and evaluate the model's behavior in real-world scenarios. If a model consistently misclassifies certain emotions or fixates on non-informative areas, developers and healthcare professionals can use XAI tools to identify and correct these issues. This not only improves model performance but also contributes to ethical AI deployment by reducing bias and ensuring fair treatment across diverse user groups. Additionally, explainable outputs can serve as educational tools—helping clinicians communicate AI insights to patients in an accessible manner, fostering greater acceptance and collaboration between human expertise and machine intelligence. Ultimately, XAI transforms deep learning models from opaque black boxes into transparent and trustworthy systems, making them more suitable and responsible for use in the delicate field of mental health diagnosis and monitoring.

8.5 Multi-modal fusion methods

In the multimodal fusion framework adopted for depression detection, two integration strategies were explored: the Weighted Average and the Multi-modal Neural Network approaches. Both aim to combine the insights drawn from facial image-based emotion classification and text-based sentiment analysis (using a Bi-LSTM model), producing a single, unified prediction regarding a subject's mental health state.

In the Weighted Average method, predictions from both the image and text models are combined by assigning weights that reflect the relative importance of each modality in the context of detection depression. This weighting process is not arbitrary; it requires careful evaluation and expert guidance, particularly from clinical psychologists or psychiatrists who understand the psychological indicators of depression. For instance, in some cases, facial expressions might be more expressive or telling (e.g., prolonged sadness, emotional flatness), while in others, the textual content of a person's writing may reveal deeper cognitive distortions, hopelessness, or suicidal ideation. These expert insights help determine whether the image model (emotion cues) or the text model (language and tone) should be given more influence in the final prediction. This personalized weight assignment is especially crucial when models show conflicting outputs, as it ensures the decision is driven by clinical relevance rather than just algorithmic confidence.

The advantages of the Weighted Average approach lie in its simplicity, interpretability, and the ability to incorporate domain knowledge into the model fusion process. It allows flexibility in balancing the two models based on situational or contextual needs. However, it also has limitations. The method assumes linear combinability of modalities and does not account for potential dependencies or interactions between image and text features. Additionally, determining the correct weights can be subjective and requires continuous refinement through expert feedback and model validation.

On the other hand, the Multi-modal Neural Network approach involves learning a joint representation of both image and text features through a neural architecture that combines outputs from the Bi-LSTM (for text) and VGG19 (for images). This approach allows the model to learn complex inter-modal relationships, capturing dependencies that may not be obvious when considering modalities in isolation. The fused features are passed through dense layers to produce the final classification. The advantages here include a potentially more robust and flexible model that adapts to nuances in both modalities without requiring manual weight assignments. However, this method demands significantly more training data and computational resources. Additionally, it lacks interpretability, which is a drawback in clinical settings where understanding why a prediction was made is often as important as the prediction itself.

In conclusion, both fusion strategies offer valuable pathways for integrating visual and textual cues in depression detection. The Weighted Average approach provides clinical control and interpretability, while the Multi-modal Neural Network offers deeper, potentially more accurate integration of heterogeneous features. These approaches are still evolving, especially in the mental health domain, where ethical considerations, explainability, and clinical validity play a central role. Ongoing collaboration with mental health professionals is essential to refine these fusion techniques and ensure they provide meaningful, trustworthy support in real-world diagnostic settings.

The Bi-LSTM with Attention model for text classification demonstrated strong performance, achieving an AUC of 0.97 and high accuracy as indicated by the confusion matrix, with minimal false positives and false negatives. This highlights the model’s excellent capability in identifying depressive and non-depressive text patterns. However, rather than relying solely on a single model to determine depressive states, the study incorporated a multimodal approach by combining this text-based Bi-LSTM model with a VGG19-based emotion detection model for facial expression analysis. This decision is backed by multiple research findings which prove that using both text and image modalities enhances depression detection accuracy and reliability.

The VGG19 model effectively captured subtle emotional cues from facial expressions, while the Bi-LSTM model interpreted contextual and linguistic indicators of mental health from text. Together, these models provided complementary insights—text data often reflects internal thoughts, while facial expressions can reveal non-verbal emotional states. Proven by research in multimodal learning, such combinations reduce ambiguity, increase robustness, and better mimic real-world clinical evaluations, where both verbal and non-verbal cues are assessed. Therefore, integrating both models forms a more comprehensive and dependable system for identifying depressive symptoms than relying on either model alone.

9. Contribute to the Domain / Commercialization

Mental health disorders affect a significant portion of the global population, making this a large and critical market to address. Traditional methods of diagnosing and managing mental disorders involve in-person consultations, which can be time-consuming and costly for patients. Our solution aims to bridge this gap by providing a convenient and cost-effective tool for diagnosing and managing depression through a mobile application. This approach not only reduces the challenges associated with traditional healthcare methods but also offers continuous monitoring and personalized insights.

We plan to introduce multiple subscription plans for different user segments:

- Monthly Subscription: Rs. 500
- Annual Subscription: Rs. 4,500

Our Target Market is General Practitioners of the MOH centers, Institutes who learn about mental disorders, General population who wish to take self-mental care. We will offer the initial depression diagnosis feature free of charge, allowing users to assess their mental health at no cost. However, to access more detailed reports, continuous monitoring, and personalized treatment plans, users will need to subscribe to one of the available plans. This model ensures accessibility while also generating revenue to sustain and enhance the platform. By offering flexible pricing and targeting a wide range of users, including partnerships with community clinics, we aim to create a scalable and impactful solution in the mental health space, addressing a critical global need.

10. Conclusion

In conclusion, this research presents a robust multimodal approach for depression detection by integrating a Bi-LSTM with Attention model for textual data and a VGG19-based CNN for emotion recognition from images. The text model achieved high accuracy with an AUC of 0.97, supported by a strong confusion matrix indicating reliable classification of both depressive and non-depressive instances. The VGG19 model effectively detected a wide range of facial emotions, enhancing the interpretability and context of user states. Combining both models allowed for a more holistic and accurate assessment compared to using a single modality, aligning with proven findings in literature that multimodal systems outperform unimodal ones in complex psychological tasks. Overall, the proposed framework demonstrates significant potential for real-world mental health applications, offering a reliable and explainable tool for early depression detection and emotional monitoring.

References

[1] "Depressive disoders (Depression)," World Heath Organization , 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] Khushi Shah, Urmi Patel, Yogesh Kumar, "Machine Learning-Based Approaches for Early Prediction of Depression," *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE),* pp. 1-7, 2024.

[3] A. Agrawal, S. Dey, and G. C. Jana, "Depressive and Non-depressive Tweets Classification using a Sequential Deep Learning Model," *International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC),* pp. 1-6, 2023.

[4] M. k. Jha, R. Ranjan, G. K. Dixit, and K. Kumar, "An Efficient Machine Learning Classification with Feature Selection Techniques for Depression Detection from Social Media," *023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI),* pp. 481-486, 2023.

[5] A. Mulay, A. Dhekne, R. Wani, S. Kadam, P. Deshpande, and P. Deshpande, "Automatic Depression Level Detection Through Visual Input," *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4),* pp. 19-22, 2020.

[6] S. A. Nasser, I. A. Hashim and W. H. Ali, "A review on depression detection and diagnoses based on visual facial cues," *2020 3rd International Conference on Engineering Technology and its Applications (IICETA),* pp. 35-40, 2020.

[7] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal Spatiotemporal Representation for Automatic Depression Level Detection," *IEEE Transactions on Affective Computing,* vol. 14, pp. 294-307, 1 January - March 2023.

[8] Jing Wang and Ci Zhang, "Cross-modality fusion with EEG and text for enhanced emotion detection in English Writing," vol. 18, pp. 162-175, 17 January 2025.

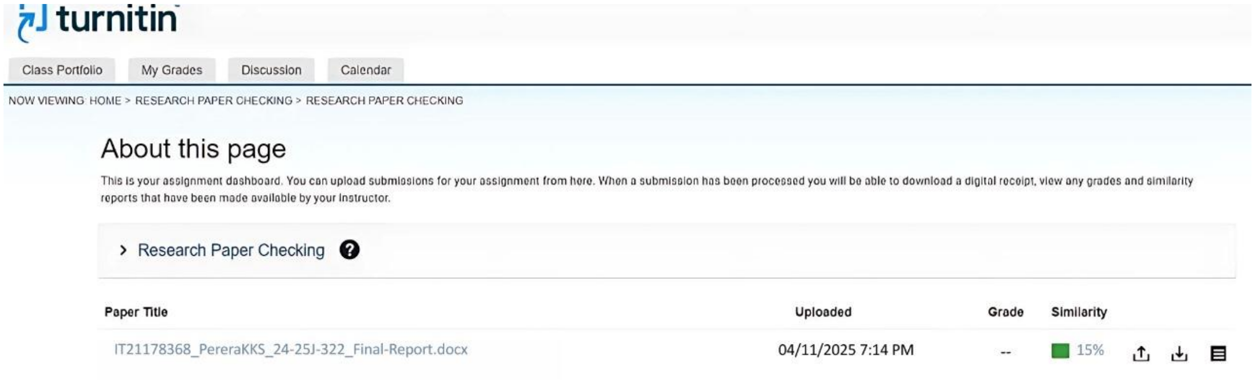
[9] Farhan Kabir, Md. Ali Hossain, A. F. M. Minhazur Rahman, and Sadia Zaman Mishu, "Depression Detection From Social Media Textual," *2023 26th International Conference on Computer and Information Technology (ICCIT),* pp. 13-15, 2023.

[10] Thrinith Fernando, Samadhi Rathnayake, and Kapila Dissanayaka, "Galaxy Morphology Classification Based on," *2024 6th International Conference on Advancements in Computing (ICAC),* pp. 1-6, 2024.

[11] Loukas Ilias, Spiros Mouzakitis, and Dimitris Askounis, "Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media," *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS,* vol. 11, pp. 1-12, 2 April 2024.

[12] Li Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, and Bin Hu, "TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection," *IEEE TRANSACTIONS ON NEURALSYSTEMSANDREHABILITATIONENGINEERING,* vol. 31, pp. 1-11, 2023.

Appendices



NOW VIEWING: HOME > RESEARCH PAPER CHECKING > RESEARCH PAPER CHECKING

About this page

This is your assignment dashboard. You can upload submissions for your assignment from here. When a submission has been processed you will be able to download a digital receipt, view any grades and similarity reports that have been made available by your instructor.

› Research Paper Checking

Paper Title	Uploaded	Grade	Similarity
IT21178368_PereraKKS_24-25J-322_Final-Report.docx	04/11/2025 7:14 PM	--	🟩 15%

Fig 15. Plagiarism Value