Human-Centered Psychological State Analysis from Student Feedback Using Behavioral and Visual Cues

by

Niaz Rahman
21301239
Faria Bintee Azad
21301068
Cyrus Ornob Corraya
21341014
Toha Hossain
21301120
Nishat Subha
21301684

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

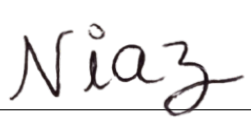
Department of Computer Science and Engineering
Brac University
October, 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

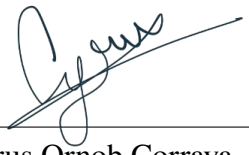


Niaz Rahman

21301239

Faria Bintee Azad

21301068



Cyrus Ornob Corraya

21341014

Toha Hossain

21301120

Nishat Subha

21301684

Approval

The thesis/project titled "Human-Centered Psychological State Analysis from Student Feedback Using Behavioral and Visual Cues" submitted by

1. Niaz Rahman (21301239)
2. Faria Bintee Azad (21301068)
3. Cyrus Ornob Corraya (21341014)
4. Toha Hossain (21301120)
5. Nishat Subha (21301684)

of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall, 2025 Year.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The paper will provide a humanistic perspective of researching psychological states based on behavioural and visual indicators, which can be acquired during the interaction with the students. Psychometric interviews were recorded and this facilitated the initiation of the research process wherein thirteen individual features related to the facial expression, body gesture, and the vocal patterns of the computer vision of audio processing were received. Each modality was well examined. ralidation process in order to get proper representation of real world behavioral response. The answers of the participants were assessed thoroughly and it was realized that the answers shared certain patterns as regards to the expression of emotions that concur with the world psychological study. The move confirmed the authenticity of the data set and the finesse of the extracted aspects in the description of subtle behavioral patterns. In the paper, there is also an inherent challenge of unequal distributions of information in psychological assessment whereby, there are minimal cases of depression occurring so naturally. take control of the sample- a phenomenon that is always evident in the real world populations. The proposed framework demonstrates how the minor changes in behavior, such as the alteration of the gaze patterns, or the reduction of the physical activity, or the voice, can reveal the underlying states of emotions. The feature, besides the initial application in the screening of depression, is also applied in screening of persons with coronary heart disease. set can be used in education engagement, telehealth diagnostics, and affect-sensitive human-computer interaction system with high likelihood of success.

Keywords: Psychological State Analysis, Behavioral Cues, Visual Cues, Human-Centered Computing, Depression Detection, Educational Technology, Telehealth Assessment, Human-Computer Interaction, Feature Extraction, Computer Vision, Audio Processing

Acknowledgement

First and foremost, we would like to express our sincere gratitude to our supervisor, Dr. Md. Golam Rabiul Alam, for his continuous guidance, support, and patience throughout this research journey. His insightful feedback and encouragement were invaluable in completing this thesis.

We extend our heartfelt thanks to the thesis committee members for their time and constructive feedback that helped improve this research.

We are grateful to BRAC University for providing the necessary resources and facilities to conduct this research. Special thanks to all the participants who volunteered their time for this study.

Lastly, we would like to express our deepest gratitude to our families for their unwavering support and encouragement throughout our academic journey, and to our friends for their moral support.

This accomplishment would not have been possible without all of you.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

After hearing the word psychological state, Depression and mental Health words come to our head automatically. This is the most talked about topic around the world. Depression has become one of the most serious mental health issues around the world, without even knowing it's affecting millions of people. Even Bangladeshi students are facing depression like the western parts of the world. If these serious mental health issues are not solved or detected at an early stage this will become more challenging mainly for students to survive easily [1].

Talking about depression detection or hearing this word, normal people think about the traditional ways of depression detection. Traditional ways means, giving a writing test or test yourself online by giving a few answers and based on the online score we will get to know if someone has depression or not. Basically traditional ways have limitations. Limitations like self reporting, normal interviews, biasness of culture. Biased culture affects the traditional methods cause western culture and Bengali culture are different so students of Bangladesh have few different behaviors from the western parts of the world. But these clinical interviews and psychometric questionnaires are okay, not so bad, still this blocks or restricts many sectors [2].

Researchers are now more focused on AI or computer vision or behavioral and visual fields [3]. This study mainly focuses on multimodal approach and psychological state analysis using facial expression, body gesture and voice patterns or audio [4]. Even though, previous study tried to build unimodal or bimodal approaches like face and audio or text or audio but Facial expression, Body Gesture and Audio never work together. Previous studies always try to ignore body gesture but we tried all three together as this is the most notable research gap. So, we connected these three behavioral modalities with human expression and connections together [5].

1.2 Rationale of the Study

This research focuses on both advances in technology and regional demands. By analyzing voice, facial expressions, and body movements, multimodal affective computing is growing worldwide to evaluate emotions and detect sadness [6]. Using all three types of signals (tri-modal) performs better than using only one (unimodal) or two (bimodal) types, as shown by benchmark datasets such as DAIC-WoZ and AVEC, along with research from universities like the University of Southern California and Imperial College London [4, 5].

Combining voice patterns, facial expressions, and body movements gives a better and broader understanding of a person's psychological condition, as shown by studies [7]. Studies indicate that integrating voice prosody, facial action units, and body movement dynamics leads to a more broad understanding of psychological conditions [7]. Nevertheless, an important gap persists in the actual implementation of these international standards to the Bangladeshi scenario. Current models are based mainly on Western society, incorporating behavioral patterns, linguistic variations, and cultural display customs that differ greatly from those in Bangladesh [8]. Local implementation of Bengali language models can result in biased and lower precision due to variations in gesture expressivity, eye contact norms, and patterns of speech in contrast to Western datasets [9].

This research aims to establish a suitable cultural tri-modal framework that corresponds to Bengali cultural standards of behavior. Our goal is to establish an infrastructure that is both technologically proficient and culturally appropriate by building Bangladesh's first complete dataset which incorporates audio, facial, and bodily gestures [10].

We extend the use of extracted characteristics beyond depression recognition to educational analytics, telemedicine, and interaction between humans and computers, improving the societal impact in an atmosphere of limited resources [11].

1.3 Research Gap

Psychological State Analysis or depression detection using multimodal always have the world's attention for a long time. Depression detection using text and questionnaires and self reported systems became old and time consuming. Also non clinical, and clinical people try to hide their true emotions. Our study focuses on facial expression, body gesture and audio or voice pattern in our own local dataset. The previous studies or existing studies and datasets like DAIC-WoZ and AVEC, these actually focus on the face or audio or text with their own western data [4, 5]. But there is not any local dataset for Bangladeshi students. Western datasets means english speaking or speech, western and english behavior that doesn't match with Bangladeshi students. Because Bangladeshi students have different cultural and behavioral expressions than western people. Thats why its creating a problem and cultural and behavioral gap with western students and bangladeshi students. The models also failed to account or catch the patterns[8]. If we work on an existing dataset or existing model it will not work properly on Bangladeshi students because of cultural and behaviour differences. Bangladeshi students have different cultural and behavioral and different mental health states than western people and culture. So a model and framework based on Bangladeshi students or south asian regions like this is much needed. Even though physical activity symptoms are well recognized in clinical

depression diagnosis as described in the DSM-5 [2], most computer-based models still focus mainly on voice and facial features, and make little use of the valuable information from body gestures [12]. This misses important signs like fewer gestures, fewer gestures, slouched posture, and slower movements, which are important clinical indicators of depression [12]. This misses important signs like fewer gestures fewer gestures, slouched posture, and slower movements, which are important clinical indicators of depression [13].

Finally, there remains a gap in how effective computing can be applied and scaled. Most research has focused mainly on detecting depression, with little study on how the behavioral features they extract could be used for other purposes [11].

The potential of these multimodal indicators to track educational engagement, telehealth diagnostics, human-computer interaction systems, and assistive technologies remains largely untapped, and most of this potential is still unused, which limits the social effect and growth of this promising research area [14].

1.4 Research Objectives

The study will be attempting to develop a multimodal behavioral analysis framework that is culturally and clinically applicable. The overall aim is to create a convergence of the systems that connect audio and face and body gesture analysis to identify signs of depression among the Bangladeshi students of the university [10, 15]. This entails the development of a sound data collection channel in which actual sample of behavior is obtained without violating the procedures of morality and comfort of the people involved. The derivation and validation of thirteen different behavioral features that cut across the three modalities with each feature allocated a particular psychological meaning and computational consistency is one of the main issues [7]. The exploration of significant correlations between these computational processes and self-report psychological conditions is also the objective of the study that would permit at least determining the connection between the quantitative behavioral outcomes and the human subjective experience [6].

1.5 Methodology in Brief

Such a combination of the accuracy of the computer and the topicality of psychological methodology is the methodology of the present study. The information was collected through the administration of psychometric interviews with 41 individuals on the supervised situations and a questionnaire of 110 university students [16]. The dual method will guarantee a wide and in-depth coverage of the data, which will make it possible to analyze the aspects of behavioral analyzes in detail without compromise of ecological validity. When extracting the features, we apply new open-source programs, such as MediaPipe, which is a body movement tracking tool [15], OpenFace, which is a visual action unit recognition tool [10], and selected audio processing tools, extracting vocal features [7]. The chosen technologies are availability-oriented. and consistency, in which other researchers can readily apply and work it out in analogous circumstances. The assessment methods are technological validation of feature validity and psychological validation by association with self-report criterion. [6]. This type of strict validation would ensure that

the features achieved are not simply generalized to work on algorithms, but also that they bear meaningful correlations to the underlying concepts in psychology

1.6 Scopes and Challenges

1.6.1 Scopes

The discipline of the intended study will entail creation of a program of multimodal behavioral assessments that are specifically geared towards examining the Bangladesh cultural context [8]. The sample of young adults in the form of the university students constitutes the study target population. are critically ill with mental health and digitally convenient and ready to be treated with technologies [1]. The study is an amalgamation of the three complementary modalities of behavior, which are vocal, facial and gestural in any combination to produce a full system of evaluating it [4]. The framework focuses on interpretability and clinical relevancy and places emphasis on features. that have their own distinctive psychological connotations and could be easily interpreted by mental experts [2]. The applicability of features to another application area is also addressed in the paper specifically in the educational technology and human-computer interaction systems [11].

1.6.2 Challenges

A number of major issues were raised during the research process. Integration of multimodal data was one of the greatest challenge. The audio recordings, features of facial expressions obtained using open face [10]and features of body gestures obtained using Media Pipe Holistic [15] were all recorded at the same time but had to be preprocessed with great accuracy and matched. The need to synchronize time between modalities, as well as the need to ensure consistency in feature dimensions, and guard against missing frames required a significant amount of technical changes and testing [17]. The modeling process was also complicated by the fact that depression is more severe in the female gender. Most of the respondents were in the range of minimal depression with very few in the mild and moderate range [1]. Such unbalanced distribution was a threat to the model bias especially during the training of Siamese Neural Networks using diverse encoders. Techniques e.g. pair-based learning, oversampling and cautious evaluation metrics were required so that minority classes were sufficiently represented, and the model was no more than overfitting to the dominant class [17]. Privacy and ethics were also of great concern [2]. The procedure of taking personal interview with sensitive psychological information involved the use of informed consent and assurance that the data will be stored securely. Last but not least, the small size of the dataset was also problematic [6]. Deep learning models were also prone to overfitting since they only had 41 participants. Leave-one-out cross validation, prudent selection of encoders and training of Siamese networks were used to maximize the utility of the given data and keep the model generalizable [17].

1.7 Research Outcomes

This study shows mixed of multimodal and psychological state analysis together. In the BRAC University Podcast Room, we collected 41 participants in person clinical interviews with 110 survey responses from university students. By evaluation the survey , most of the student falls under the Minimum category [1]. This data imbalance evaluate us that most of the students don't fall under the high major depression level in the real world, but the fall under the minimum category [2]. Most of the student showed mild emotional stress or low energy , but they are not severely depressed. This finding about depression level emphasize that students need behavioral screening and psychological state analysis. We extracted 13 features from videos by using OpenFace [10] for facial expression and MediaPipe [15] for body gesture and Silero VAD, Librosa RMS and CREPE for Audio. After merging the features, both both binary and continuous values combined on a single dataset. This merge features latterly put on a multimodal fusion for generating the outputs and it also support machine learning models like Siamese Neural Network based approaches with the help of encoders like MLP , Transformers etc [17]. This datasets combinely helps to detect the emotions more accurately and meaningfully . This study doesn't only show depression detections, also use in human computer interaction (HCI), telemedicine and assistive technologies [11, 14]. Psychological distress pattern can be detected through all the interpretations we could make by our established framework through our research in the fields of Human-centered AI system. Our approach tried to implement and interpret regionally via grounded methodology [8]. In fine, Our study shows both methodological and practical outcomes validates that real life imbalance data that goes with our own collected dataset that most of the students falls under minimum depression scale. And we select Siamese Neural Network (SNN) based approaches with encoders like MLP, LSTM, and Transformer that help to detect depression detection [6, 7]. Most of the students face small emotional struggles rather than major depression.

Chapter 2

Literature Review

2.1 Preliminaries

Using Audio, Body Gesture and Facial Expression, depression detection become a significant research traction over the past decade using multimodal feature fusion [18]. The automatic detection of depressive behavior through multimodal feature fusion-combining audio, visual, and physiological data has gained significant research traction over the past decade [18]. Models can now understand a person's emotional condition by combining the human behavior analytical terms with deep learning models with the help of their vocal tone, facial expressions and body gesture [19]. This research developed an end-to-end multimodal framework, comprising preprocessing pipelines using OpenFace for facial features, MediaPipe Holistic for body gestures, and Silero VAD + CREPE + Librosa RMS for Vocal tone analysis, with the help of SNN + MLP-based classification for depression level prediction.

2.2 Review of Existing Research

Main focusses in detecting depression in the recents years are various types of signals found in behaviour as ever changes in face and while responding have been the main factor. And so these combinations more precisely plays vital role in depression detection through proper analysis[18]. Previously when people were interviewed biased result outcomes had been seen but now machines can more accurately study the behavioral features through body gesture, facial expressions, voice as in audio features [19].

The most unique and well researched method is absent in previous researches while detecting the depression of humans. OpenFace tool gives precisely calculated features extracted by facial landmarks like head tilt, action units measuring all the up and down to all the subtle cues[10]. Previous studies linked diminished smile, limited eyebrow movement and confirmed facial analysis could be reliable proxy of clinics[20, 21].The temporal dynamics carry critical and significant information to detect depression. Body gesture analysis done through MediaPipe MediaPipe Holistic working as a contemporary of behavioral data, researchers can capture detailed postures and kinematic features could be found that includes hand gestures, head tilts, and hand movements [22]. Reduced kinetic energy, slumped postures and very limited hand movements can be seen in a depressed patient reflecting psychomotor retardation [23, 24]. Combining body gesture and facial metrics

enhanced subtle affective changes detection which might get missed when relying on facial cues only.Silero VAD technique to detect speech activity , CREPE for pitch estimation, and Librosa for energy extraction have been widely employed or used in recent research [25]. Depressive symptoms can be correlated with lower pitch variability, reduced speech intensity, and extended pauses [3, 26].Furthermore, aligning audio with visual data at high resolution like 60 frames per second allows us to have synchronised multimodal analysis by capturing micro-level interactions between speech , face and body gesture or facial cues [27].

Combining all the features extracted helps capturing complex emotions with behavioral relationships [6, 28]. Interpretibility and model accuracy can be found from hybrid features or handcrafted along with deep learning. [29].

Siamese Neural Networks (SNNs) combined with LSTM or MLP encoders have been useful in small clinical datasets [30]. Static patterns and robustness against data imbalance and variability across participants can be achieved [31, 32]. Leave-One-Out Cross-Validation (LOOCV) ensures subject independent testing and reduces overfitting [33].

While prior studies have predominantly focused on two modalities, such as audio-visual or text-audio combinations [34], recent research underscores the advantages of integrating three or more modalities. By simultaneously analyzing facial expressions, body gestures, and speech, it becomes possible to capture a more comprehensive behavioral signature of depression [35]. Previously researchers mainly focused on just two modalities and combined audio with video only or audio with text only [34]. [36]. Recent work focuses more on more and more modalities and shifting towards multimodal integration by analysing emotions in order to enhance clinical reliability connecting with reality and through accurate calculations found from multimodal approach[25].

So by using advanced deep learning models and examining helps in boosting automated depression detection accuracy by getting closer and connecting to the real world and getting useful in clinical sectors.

Table 2.1: Comparison of related works and our proposed approach

Paper / Model	Dataset Used	Features Used	Model Architecture	Why Not Directly Applicable to Our Dataset
Cummins et al., 2015	AVEC / Custom	Audio, Text	LSTM	Cannot handle audio + facial + body gesture + video features; input shapes mismatch, cannot capture multimodal interactions
Valstar et al., 2016	AVEC Challenge	Audio, Text, Facial	LSTM / CNN	Designed for balanced datasets; cannot handle imbalanced multimodal sequences or inter-modal correlations
Tzirakis et al., 2017	AVEC / RECOLA	Audio, Text	CNN + LSTM	Only works on audio/text, ignores body gesture and video; cannot process fused embeddings
Al Hanai et al., 2018	AVEC / MIT dataset	Audio, Text	LSTM	Assumes balanced labels, cannot handle minority patterns even if data is duplicated
Ringeval et al., 2019	AVEC	Audio, Facial	LSTM / CNN	Works for audio + facial only; cannot process audio + facial + body gesture + video combined
Zhang et al., 2020	AVEC / Custom	Audio, Text, Facial	Bi-LSTM / CNN	Trained on balanced multimodal dataset; cannot handle imbalanced fused embeddings, may overfit duplicated minority data
Ours (2025)	Self-collected	Audio, Text, Facial, Body Gesture	SNN + MLP	Specifically designed to handle imbalanced, heterogeneous multimodal dataset, preserves minority class sensitivity

2.2.1 Summary of Key Findings

The existing literature review reveals critical insights indicating our Unified approach through Multimodal analysis by combining audio, facial with body gestures significantly improves depression detection accuracy. Behavioral analysis can be done by extracting features through OpenFace, MediaPipe Holistic along with Silero VAD reliable tools. Depressive state strong indicators can be interpreted through facial action units (AUs), reduced expressivity in smile and microexpressions like eyebrow rise. Psychomotor symptoms can be captured by extracting body gesture features which include posture and movement energy dynamics that are linked and proven clinically. Vocal characteristics such as pitch variance, speech energy, and pause duration correlate strongly with depression. Siamese Neural Networks with LSTM/MLP encoders offer effective solutions for small-sample affective computing datasets. Early fusion strategies and temporal synchronization are crucial for maximizing multimodal integration benefits. Cultural adaptation of assessment tools, as demonstrated by the Uddin Rohoman (2005) scale, is essential for region-specific applications. These findings collectively justify our tri-modal framework and provide the theoretical foundation for the methodological choices in this study

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

Our depression detection through our collected data by psychometric interview we could successfully design our final system. All the four datastreams have been merged to create an unified model from audio, video from face and body movements. We have used PyTorch for data processing and modeling for neural networks [37] (for neural network training), For accelerating deep learning computations and feature extraction we have used GPU-enabled hardware setup. So all of these play a significant role in accurately predicting across depression levels namely Minimum, Mild and Moderate.

3.2 Societal Impact

In the sector of public health, depression is growing tremendously. Students and young adults are suffering and not treating themselves due to social stigma and having accessibility constrai. Our automatized data-driven approach successfully detects depressive tendencies by analyzing through dominant questionnaires merging all features like facial, speech tone, body movements. Our model implementation when used in the fields of clinical workspace along with education could promote awareness by consulting as early as possible and ultimately improve human mental health socially.

3.3 Environmental Impact

In our approach, computational processing that we have been doing by GPU-based system is consuming energy with minimal direct impact on the environment.[38]. Our suggestion is to implement optimized training pipelines along with efficient batch training to decrease load when computerised and reduce power consumption or usage. Manual screening process can be implemented to reduce paper work and convert to a workflow that is digitalised and also eco-friendly.

3.4 Ethical Issues

We have prioritised ethics throughout our research by taking every participants' consent manually while filling up the form [39]. The identifiers are stored for scope analysis solely for the research and preliminary screening and not for clinical diagnosis replacement.[40]. The responsibilities to ethically apply are assigned to developers and researchers. We tried to mitigate biasness via balancing gender-wise depression categories ensuring fair implemented results and performances [41].

3.5 Standards

Informed consent protocols and anonymisations of human subjects' data has been ensured following general standards of ethical AI and data privacy[42]. It also follows standards in transparency and methodological reproducibility [43]. The extractions have been done following standard frameworks like MediaPipe and OpenFace which is used in human-centered interactions very widely that has its effect on commuting. Hence we could align our research and proposed model with IEEE research principles [[44].

3.6 Project Management Plan

Structurally our project has been developed dividing into three phases where dataset preparation was the first and then questionnaire development and data collection was the second focusing on feature extraction and data pre- processing. Integrating modalities into the fusion model via evaluation of accuracy determination we could come in the third or the final stage. Cloud based GPU servers were used for coding and validating by the principles of agile development [45].Most tools that we have used were open-source and institutionally supported and so the budget costs were minimum.

3.7 Risk Management

The risks we could identify were data imbalance resulting in biasing the model in predicting depression levels which we removed through balancing techniques like duplication. Secondly while preprocessing the scripts by aligning with respect to time, it was risky to synchronise video and audio streams By anonymising the participant during handling and securing data, we could ensure ethical considerations. [**?**]. Model overfitting has been reduced through cross-validation that we have done via manual checks and following strategies like early stop.[46]. Technical risks such as model overfitting were reduced through cross-validation and early stopping strategies [47].Interpretibility and consistency have been ensured monitoring regular performance.

3.8 Economic Analysis

The project demonstrates strong cost-effectiveness compared to traditional psychological assessments [48]. Once trained, the model can analyze large volumes of interview data automatically with minimal human involvement, reducing the time and cost of depression screening. The total project cost primarily includes computational resources and data storage, both of which are relatively low due to open-source software use. In the long term, such systems could significantly reduce healthcare screening costs while enabling early intervention — yielding a high social and economic return on investment [49].

Chapter 4

Research Methodology

4.1 Design Process or Methodology Overview

We want to build a system that can detect depression by using multiple modals with the help of vocal speech, facial expressions and body gestures. The whole process followed by a continuous process which started with data collection and then continued through data preprocessing, model development and multimodal fusion. By taking psychometric interviews which are conducted in a podcast room , we collect the datas. Every participant filled the thirty questionnaires which were developed by Uddin and Rohoman (2005) [16], which was inspired by the widely recognized PHQ-9 scale [54] We took the interview record so that we can extract the features. The dataset contains synchronized audio and video recordings, allowing the integration of verbal, visual, and linguistic indicators of depressive behavior. Before developing the model, we extract 13 features from audio and videos and manually validate if the extracting values are accurate or not. The video recordings were processed frame by frame using MediaPipe Holistic [15] to extract body keypoints. Using SIlero VAD, Librosa RMS and CREPE, we extract the Start and End vocal time, PITCH and RMS Energy values. We collect facial expression's features using the OpenFace model. Since the datasets contain the minimum label largely, we duplicate the existing labels which are mild and moderate so that our imbalance data can maintain the measurement with minimum label Feature extraction was conducted independently for each modality. Audio features such as pitch, vocal start and stop time and RMS energy were extracted using Librosa RMS, CREPE and Silero VAD. Facial expressions were analyzed through OpenFace to capture emotional intensity and subtle affective variations. Body gesture features were captured from MediaPipe outputs, measuring ratios such as head tilt, hand movement, leg movement, and posture stability. Before building the model, we merge all the features on one csv file and we convert all the numerical values on binary classifications, Yes or No, but we remain the Pitch and RMS Energy value the same, as numerical. After that we put those 13 features merged files on SNN + MLP models so that we can generate the output. We fine-tuned the whole merge features csv so that it can correctly detect the movement of a person's for give the proper binary classification values. We put the merged csv files on different types of models and checked its performance with every metric like accuracy, F1-scores, and confusion matrices to see how well it held up. The beauty of this setup, It really dives into the full spectrum of how humans show emotions through multiple channels, while staying fair and easy to understand no matter the severity level. Out of our 41 participants, 37 fell into the minimal category, just 3 participants had mild depression, 1 participant had moderate label. Here occurs a problem,

imbalance data is a complex problem for AI models , it could easily lead to overfitting for the common minimal cases, blinding the system to signs of worse depression [17]. To fix this, we duplicate the mild and moderate data with the similar amount of minimal,so that dataset holds the stability.

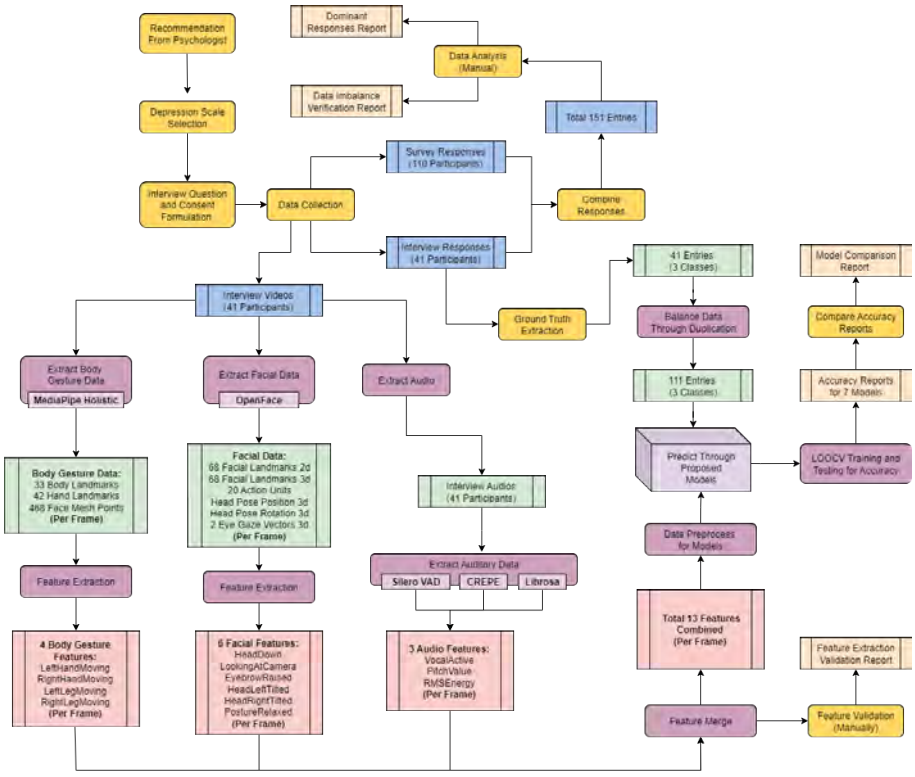


Figure 4.1: Methodology

4.2 Dataset Description

The data collection approach adopted in this study was the application of a comprehensive strategy of data collection which utilised depression. This measurement scale was formulated by Uddin and Rohoman (2005) [16]. The study involved carrying out structured clinical interviews on 41 respondents in the soundproof BRAC University Podcast Room to be able to record in the best conditions. All the sessions were recorded by the use of a tripod-mounted mobile camera with 60 Frames per Second Full HD resolution, that works on the quality extraction of the gestures of the body, and the face gesture.

Air conditioning of the room was also shut off during recordings to minimize ambient noise. Non-ambient vocal data was recorded using a high-quality external microphone to reduce noise interference. The interview procedure was to be conducted in oral form where the participants were expected to answer all the questions in a natural and free style. Their responses were recorded alongside the standardized depression questionnaire by Uddin & Rohoman [16]. All participants provided informed consent founded upon the current standards of ethical conduct [2], to which the research integrity standards would be applied. The outcomes of the data analysis turned out the existence of a huge disproportion in distribution of the categories of the severity of depression.

Among the 41 respondents, 37 respondents were Minimum, 3 had mild depression, and there were no participants with severe depression. This unequal allocation is in line with the global patterns of nonclinical groups that comprise of minimal depression groups as the natural majorities [1]. The expressed imbalance in classes was a serious threat to machine learning models since it is likely that the direct training would also lead to the Minimum class being overfitted to, and the model to recognize patterns consistent with higher level of severity [17]. To overcome this issue, we employed another data collection plan through survey methodology. Additional interviews can be potentially continued the identical distribution tendency, the survey strategy assisted us to gather broader response data and still being methodologically consistent. We have prepared a comprehensive survey questionnaire with all 30 questions of Uddin & Rohoman scale [16] and implemented the questionnaire among undergraduate students, where 110 other answers were collected. This measurement system allowed us to collect more data while adhering to ethical principles and ensuring the cultural applicability of our assessment tool [50].

4.3 Data Preprocessing

4.3.1 Audio Preprocessing

For audio preprocessing, the video we took and recorded in an in person interview from those we are extracting raw audio. For audio extracting we are using FFmpeg. These FFmpeg, actually converted each video into high quality MP3 but we did not change any tone or speech. This audio will help in multimodal fusion as we are connecting three modalities: “audio or voice patterns, facial expression and body gesture.” To detect that someone is actually speaking in the audio or someone actually stops talking in the video we used Voice activity detection was done using Silero VAD [25]. Without video, just using audio it is hard to know where someone is speaking and someone just stopped talking or taking a break. Silero VAD is a deep learning based method. This modal actually works better than older and simpler techniques. This system listens for a very short time from that audio, which may last only about 30-100 milliseconds. In this time the audio parts are basically overlapping and it’s checking if a person is actually speaking or not like loudness energy or the tone or pitch pattern harmonic structure of the sound. If a person is actually speaking in that video it will take note and only take the speaking part and then change those small parts into log mel spectrogram features. We used the Convolutional Recurrent Neural Network CRNN which trained various things like “VoxCeleb, LibriSpeech, and Mozilla Common Voice”. As these models are collecting the audio if there is speech or not it will check the starting and ending part too. Basically where the voice or talking started and where this ended. For finding out people’s basic pitch estimation or basic frequency like where the tone is high or low we use the CREPE model. CREPE model is also a deep learning based neural network that is used to detect fundamental frequency (F0) of sound precisely. The audio was split into small frames like 10–20 milliseconds. In each frame we check the pitch estimation or fundamental frequency (F0). The low and high frequency in a voice mean something in depression detection that’s why we split into frames and check every frame step by step. The values of fundamental frequency (F0) are “Low means (¡120 Hz), Medium range means (120–180 Hz), and High level means (¿180 Hz)”. The audio was segmented into overlapping frames of 20-50 milliseconds, RMS values were derived as the square root of mean squared amplitude within each frame. This

gives a continuous measure of speech energy that captures prosodic variations relevant to depression symptom assessment.

4.3.2 Facial Expression Extraction

For extracting the facial expression features, we use OpenFace 2.2.0 [10] model which is very optimal for our works and this model automatically tracks important facial features. It goes through every frame of the videos and followed 68 key points on the face in both 2D and 3D. OpenFace in the end gives us some numerical values, by the help of those pinpointing. We took eye brows raise and eye gaze features from facial expression because those features helps to detect the depression. Eye gaze features combinely works with body gesture's head tilt up/down features. If a person's head tilt position at the top and his eye gaze position are at the front side which means on the camera side, it means he is looking at the camera, if his head position are at the top side and his eye gaze position are at the bottom or another side, the looking at the camera features will not work. That's how we took the features from facial expression by using the OpenFace.

4.3.3 Body Gesture Extraction

Body movement analytics deployed the MediaPipe Holistic [15] that is a state of the art of the complete body pose estimation. The system documented 33 pose landmarks which occupied the entire body topology including head, shoulders, arms, spine and location of legs. As well, both hands contained 21 hand landmarks, enabling the gestures to be tracked in more detail, and 468 face landmarks, which supported the head orientation analysis as the auxiliary information.
The torso posture was studied by the angle between positions of the shoulders and hips that was used to record the changes in the upper-body posture connected with the extent of engagement or psychomotor retardation. Head inclination measurements by eye and nose landmark positions were providing enough head tilt calculations that could have been used to measure mood or attention. Frame to frame movement of hand and leg joints was taken to establish the movement dynamics where velocity dependent thresholds were taken to explain the movements of the limbs as moving or not, depending on the defined kinematic parameters.

4.4 Feature Extraction Validation and Reliability Assessment

The bestseller and practicability of the features derived are central to the multimodal depression analysis, as the quality of the predictive model activity solely depends on the quality of the input data [10]. Our goal was to elicit and confirming features of audio, body gestures, and facial behaviours to ensure all the modalities are the real manifestations of natural behaviour and psychological state of the participants.

4.4.1 Audio Feature Validation

Speech is among the most significant attributes of the symptoms of depression in the audio qualities, and it is probable to be low, monotonous or include long pauses [3]. We used the Silero Voice Activity Detection (VAD) module, which assisted us in breaking up speech and non- speech. To be accurate we manually annotated the vocal segments of a sample of 78 recordings of participants (7-8) and compared the automated resultant output to the manual annotation. The results showed that the degree of correlation was high, which justifies that VAD is effective in identifying the speech events. Besides the speech segmentation, the pitch, and intensity were estimated by the use of CREPE model [51] and the Root Mean Square (RMS) energy, respectively. The frame-based pitch and energy contour visual analysis revealed that frame-based pitch and frame-based energy contours were accurate in indicating the variation of volume, intonation and prosody. The extracted features always contained the low-energy and low-pitch moments, which tend to be characteristic of the patterns of depressive speech, which confirms the success of the audio pipeline in extracting the nuances of voice.

4.4.2 Body Gesture Feature Validation

After extract the numerical values by using the MediaPipe Holistic, we put those values on the original video of 6-7 persons and show the real time movement of the body points.By looking at those body points, we figure out the model detects the body posture properly. All the extracted individual torso angles, movements of hands or legs and head tilt measurements were taken against the actual observed gestures. The results showed that the system was efficient in the detection of body motions that have the ability to capture the smallest of behavioural cues like shift of posture or restricted movement of a body part. The interface consistency in the sequences of the successive frames in time was also verified using realistic and smooth motion paths which are crucial to the lower level modeling of behavioral processes.

4.4.3 Facial Feature Validation

The facial features are another indication of emotional and mental conditions [21].
The open face features, which were used to derive the feature of head-posture, eye gaze, eyebrows movement and frequency of blinking, were achieved with the help of OpenFace [10]. Automatical analysis of some of the videos of participants showed that head rotation and eyebrow motion extraction were performed in accordance with the observed face expression, and had the abilities of detecting any concealed emotional expression, including the ability to detect any minimal emotional expression, such as concern, fatigue or tension. This tracking section was the eye gaze in which the subjects stared directly at the cam-era hence the Label of LookingAtCamera was triggered. Interest and interest could be tracked appropriately. Even the sensitive counts of Blink that may be influenced by fatigue and sensitive to the affective condition were contrasted with the visual measurement and found to be close to be maximum. The presence of these checks had been necessitated by the fact that the micro expressions, direction of gaze and eyelid were needed to be adequately quantified and this provided a real picture of the facial expressions in line with a depressive. We performed cross-modal checking in order to make it robust in all modalities. The periods of vocal hesitations were mostly accompanied by absence of body motions and turning of the head, which tended to portray that the extracted features were of behavioral patterns of audio, body, and facial dimensions subject to a coherent way. This correspondence suggests that the pipeline has no influence on the natural time and behavioural alignment of responses of participants.
Generally, our validation exercise shows that the multimodal feature extraction pipeline can produce correct, strong and temporally stable measurements. Sound, physical attributes, and facial expressions all depict a sum total of the manner in which the action of participants is, eradicating the danger of artifact or misrepresentation in the later modeling [25]. By a strict validation of the fact that the features extracted explain actual world behaviors, we form a solid foundation of meaning and multimodal depression analysis of clinical significance.

4.5 Multimodal Dataset Construction – Merging Features

After successfully extracting features from each modality individually, the subsequent key phase in this study involved assembling a unified multimodal dataset [6]. The objective here was to integrate the audio, facial expression, and body gesture features into a cohesive dataset, with temporal synchronization applied at the frame level. We put individual person's separate csv's on one code base so that we can merge the 13 features from separate csv's on one csv. The reason for doing that is we extract features from different niches, such as audio and videos. We don't have all the features in one particular place. Audio, Body Gestures and Facial expressions are synchronized frame to frame, for that reason the Feature level fusion is actually made.Before merging the features, we put all the features numerical values on binary classification , Yes or No. The reason is when a psychologist try to understand the condition of a person, he or she will evaluate all the things, when the depressed persons start talking or not, when they express their feelings, if their eyes are blinking or hands are shaking or legs are moving a little faster or faster. It stands with Yes or No values. That's the reason why we put all the features on binary classifications, except Pitch and RMS energy values. We can't measure these values by using the threshold because these values are dynamically changing . When a person speaks or not, we can observe the start and end vocal time, but we can't say that from this time to that time the voice tone or vocal power increase or decrease heavily or not. It's better to place this value as numeric so that from this time to that time we can calculate the vocal tone or power changing terms carefully.

Before doing the merging steps, we fix the threshold values for binary classifications because the movement of each feature varies from person to person. Suppose a person's movements are very slow, if i fix the threshold on that range that the person needs to move his or her hand very roughly, then hand movement detection features will work, it will not validate. So by evaluating 9-10 persons, we fix the threshold on that range, that it works perfectly for all types of people. We see the live videos while fixing the threshold so that we can ensure it works perfectly.

After fixing the threshold values, we merge the 13 features and found the main datasets of each persons. After that we put all the CSV files on the fusion models which learns the pattern and compare it with the ground truth so that we can see the final output.

Chapter 5

Model Training

5.1 Model Description

Multimodal depression analysis requires features to be extracted from different modalities such as audio, facial expressions and body gestures. We have considered several unique architectures, among which a Multilayer Perceptron (MLP - 3 layered), Siemens Neural Network (SNN - 1000 pairs) using the previously designed Multilayer Perceptron (MLP) encoder, and the Transformer architectures showed promising results. However, due to the faster computability of the MLP systems and SNNs' unbalanced dataset handling capability, we have selected the SNN+MLP model as our primary selection for the classification task. This system has a perfect balance of efficiency and performance. The SNN+MLP combination proved as the most suitable architecture, offering pair-based processing while efficiently capturing complex nonlinear relationships in multimodal data using the MLP encoder, making it suitable for real-world depression classification.

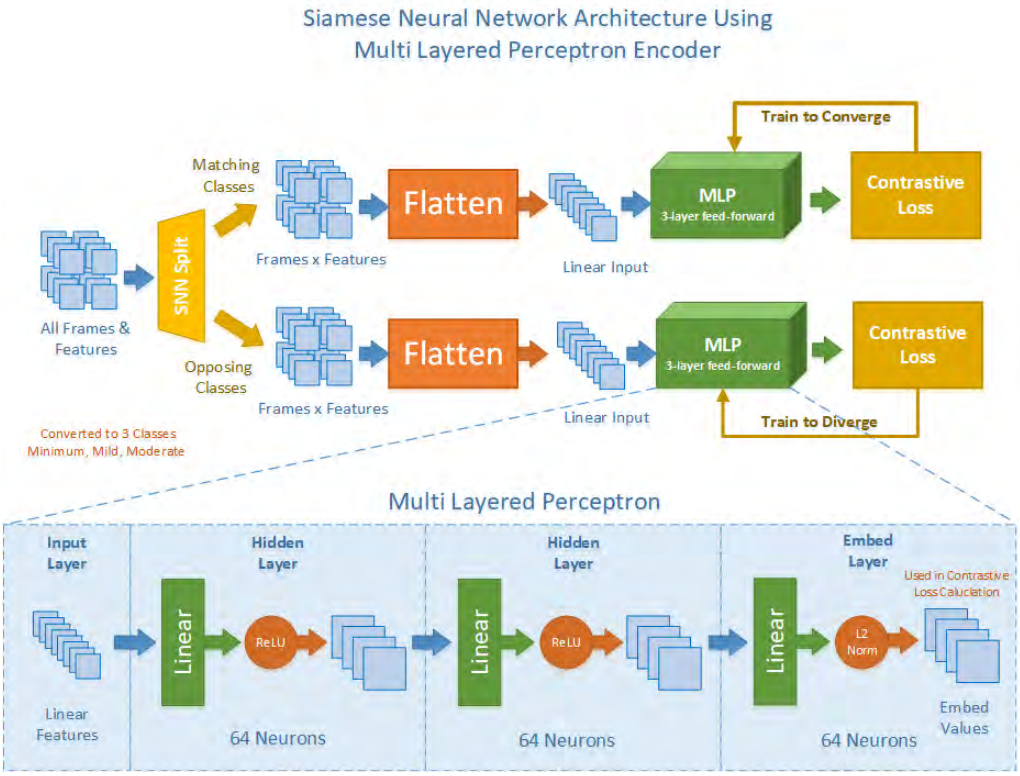


Figure 5.1: Model Structure

5.2 Model Selection Justification

5.2.1 Limitations of Alternative Architectures

We have also tried the simple LSTM network for our sequential data. Although it worked for variable frame counts, the process proved highly computationally intensive, and on the other hand, the results weren't even remotely close to our expectations. CNN was never meant to work for our type of classification. As we are not working with image or video frames directly. The facial and body gesture features are being extracted using other premade CNN-based models (Mediapipe, OpenFace). Using the features and frames as 2-dimensional data and then using Convolution methods to come up with a classification result would simply be a disaster, as we expected, and the results also proved us correct. It performed the lowest among all other options we considered. The transformer model was the best contender so far; however, it demanded a large amount of memory space, as every frame needed to be compared with every other frame. Due to its quadratic computational nature, our limited memory device couldn't handle all the frames; therefore, it had to be limited to a frame cap of 1000 frames. Which meant we could only use just a proportion of our entire dataset for this model, and therefore, we decided not to select this model as our primary selection despite its high accuracy.

5.2.2 Encoder Selection Rationale

For our multimodal training and testing, we have adapted a unique training data selection method (batch sampling using weighted loss function), and for the prediction validation

method, we used Leave-One-Subject-Out Cross-Validation (LOOCV).

MLP vs LSTM Encoder: While LSTM encoders can capture temporal patterns, they introduced unnecessary complexity for our aggregated feature representation. Our feature engineering pipeline already extracted meaningful temporal characteristics, making the sequential processing of LSTM encoders redundant. The MLP encoder's feedforward architecture provided superior computational efficiency while maintaining competitive performance [52].

MLP vs CNN Encoder: CNN encoders are designed for spatial hierarchies in image data, but our features lacked the spatial relationships that CNNs exploit [53]. The MLP's fully connected architecture better captured the global interactions between multimodal features without imposing artificial spatial constraints.

MLP vs Transformer Encoder: Transformer encoders would have required positional encoding and self-attention mechanisms that were computationally prohibitive for our dataset size [54]. The MLP's simplicity allowed effective feature transformation without the overhead of attention mechanisms.

5.3 Training Methodology

For our multimodal training and testing, we have adapted a unique training data selection method (batch sampling using weighted loss function), and for the prediction validation method, we used Leave-One-Subject-Out Cross-Validation (LOOCV).

5.3.1 Leave-One-Subject-Out Cross-Validation (LOOCV)

To evaluate the model's accuracy and ensure that the maximum number of dataset entries are utilised for training, we employed the Leave-One-Out Cross-Validation (LOOCV) method, rather than the traditional 80/20 split. We have a limited dataset of 41 entries; among them, we randomly select one entry for validation and train our models on the remaining 40 entries. This ensures the maximum number of entries is being utilised for training. Also, we ran each of the models 41 times for each different subject as validation, giving us a validation set of 41 entries as well. We extracted our accuracy from those 41 entries, depending on how many were predicted correctly by each model.

5.3.2 Class Imbalance Handling

The dataset we used had a severe class imbalance (37 Minimum vs 3 Mild vs 1 Moderate). Using the traditional random sampling, we were overfitting all the models as expected, even our specialised SNN+MLP model designed for imbalanced dataset also couldn't handle this much of an imbalanced dataset. So, we had to come up with a method to somewhat balance the sampling for training and testing. To address that issue, we used the Weighted Loss Function and strategically took batched samples depending on the weights.

$$w_c = \frac{N}{C \times N_c} \tag{5.1}$$

where N is the total number of samples (for our case $N = 41$), C is the number of classes (for our case $C = 3$), and Nc is samples in class C (37 for Minimum, 3 for Mild and 1 for Moderate). The higher the weights, the higher the chances for the entry being in a batch. This sampling method would slightly make the dataset balanced, reducing the chance of overfitting.

Strategic Batch Sampling: We implemented balanced batch sampling to ensure each batch contained proportional representation from all classes, preventing model bias toward the majority class.

5.4 Proposed Model

Based on our experimental evaluations, we propose a Siemens Neural Network (SNN) architecture combined with a Multilayer Perceptron (MLP) encoder for multimodal depression classification. The model was trained using Leave-One-Subject-Out Cross-Validation (LOOCV) to maximize the use of our limited dataset while maintaining robust evaluation.

5.4.1 Architecture Selection Rationale

Siamese Neural Network (SNN): The SNN architecture was chosen for its computational efficiency and strong capability in modeling temporal and multimodal behavioral patterns. It utilizes a pair-based sampling mechanism, enabling the network to converge for similar input pairs and diverge for dissimilar ones, thereby enhancing class separability and feature representation. This learning strategy not only ensures faster convergence but also provides improved handling of imbalanced datasets, as the pairwise learning process exposes the network to balanced relational information rather than relying solely on class frequency. As presented in Table 5.1, the SNN-based models demonstrated competitive performance with balanced parameter counts and runtime efficiency compared to traditional deep learning architectures. Overall, the Siemens Neural Network offers robust, efficient, and scalable learning suitable for real-time and clinical applications.

Multilayer Perceptron (MLP) Encoder: The MLP encoder was chosen for its effectiveness in learning complex nonlinear mappings from our aggregated multimodal features. Despite having 1,432,768 parameters (as shown in Table 5.1), the SNN + MLP configuration achieved optimal accuracy of 99.10% with only 10 minutes runtime, demonstrating an excellent balance between model capacity and computational efficiency. The MLP encoder effectively captures intricate relationships between audio, facial, and gesture features without the computational overhead of more complex sequential processors [52].

Table 5.1: Model complexity and computational efficiency comparison

Model	**Number of Parameters**	**Runtime**
CNN	17,347	7 minutes
LSTM	55,683	12 minutes
Transformer	51,328	38 minutes
MLP	1,428,803	9 minutes
SNN + CNN	188,992	9 minutes
SNN + LSTM	57,760	15 minutes
SNN + MLP (Proposed)	**1,432,768**	**10 minutes**

5.4.2 Model Performance

The performance of the proposed SNN + MLP architecture was evaluated against several alternative models to assess both accuracy and computational efficiency. Table 5.2 presents a comprehensive comparison of the different model configurations.

Table 5.2: Performance comparison of different model architectures

Model Architecture	**Accuracy (%)**
CNN	54.05
LSTM	88.29
SNN + LSTM	95.50
SNN + CNN	97.30
MLP (Standalone)	99.10
Transformer (Standalone)	99.10
SNN + MLP (Proposed)	**99.10**

The CNN model demonstrated poor performance with only 54.05% accuracy, indicating its inadequacy for complex multimodal depression classification. The LSTM architecture showed improvement with 88.29% accuracy but still fell short of acceptable performance levels for clinical applications.

The SNN + LSTM configuration achieved 95.50% accuracy, showing reasonable results but demonstrating limitations in handling imbalanced data effectively. The SNN + CNN model performed better, reaching 97.30% accuracy, indicating that combining convolutional and paired Features improves representation learning but still falls short of the best-performing models.

Both the standalone MLP and Transformer models achieved excellent accuracy of 99.10%, with the MLP offering lightweight computation and the Transformer providing powerful attention mechanisms. However, the standalone MLP lacks specialized architecture for multimodal temporal data, while the Transformer proves highly memory-hungry for practical deployment.

Our proposed SNN + MLP model achieved 99.10% accuracy while maintaining the perfect balance between computational efficiency and classification performance. This result highlights the effectiveness of using an MLP encoder within the SNN framework, enabling the model to capture complex feature interactions without excessive processing costs. The consistently high performance of the SNN + MLP architecture validates its design, especially when dealing with imbalanced datasets, making it the strongest candidate for real-world depression classification tasks.

The training and validation loss curves of the SNN-MLP model display a consistent downward trajectory throughout the training process, with both curves gradually converging near the 20th epoch. This convergence point indicates that the model achieves an optimal balance between learning the training data and generalizing to unseen validation samples. The smooth and parallel decline of both losses, without any subsequent divergence, confirms that the model effectively avoids overfitting. Overall, these results demonstrate

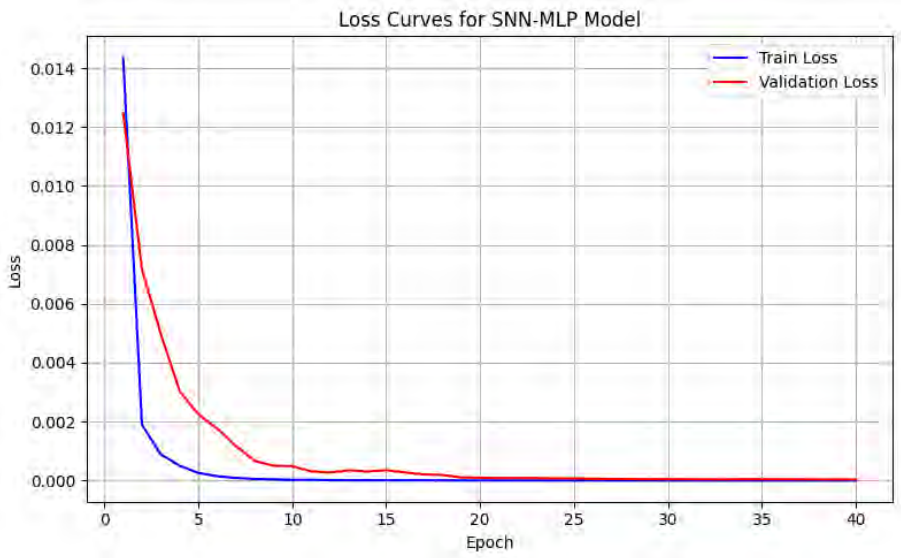


Figure 5.2: SNN+MLP Loss Curve

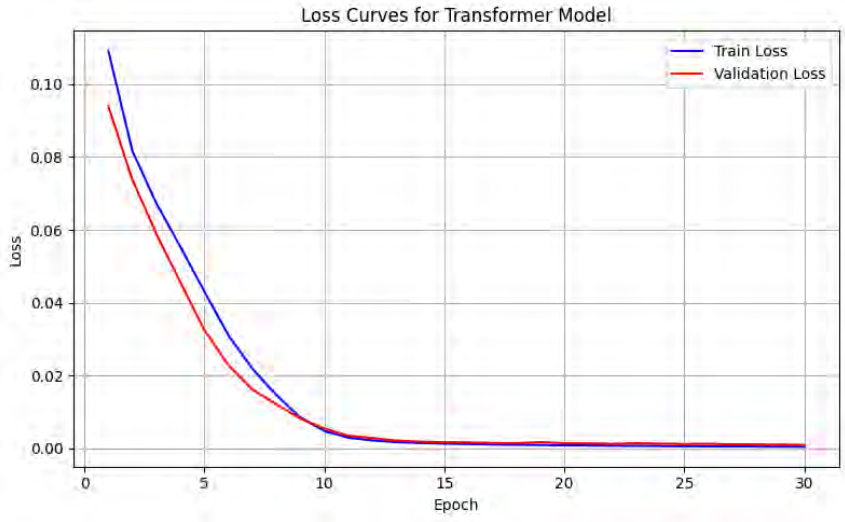


Figure 5.3: Transformer Loss Curve

that the SNN-MLP model learns in a stable manner, achieving reliable generalization performance across both training and validation datasets.
Similarly, the Transformer model exhibits a steady decrease in both training and validation losses, converging around the 15th epoch. The absence of any divergence between the curves indicates that the model does not overfit the training data, maintaining a strong ability to generalize to unseen samples. These results confirm that the Transformer model is also working as intended, learning efficiently and demonstrating consistent generalization performance.

5.5 Confusion Matrix

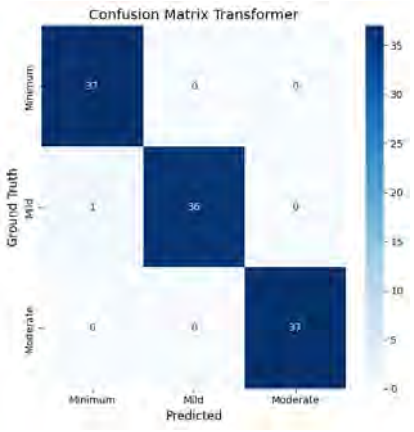


Figure 5.4: Transformer

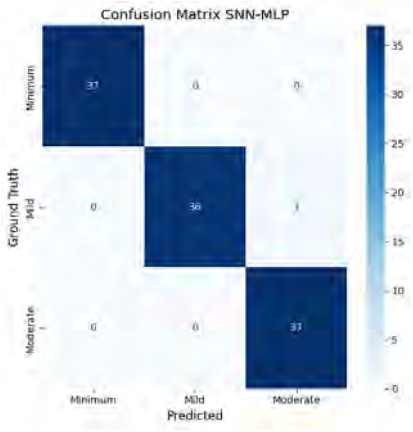


Figure 5.5: SNN + MLP

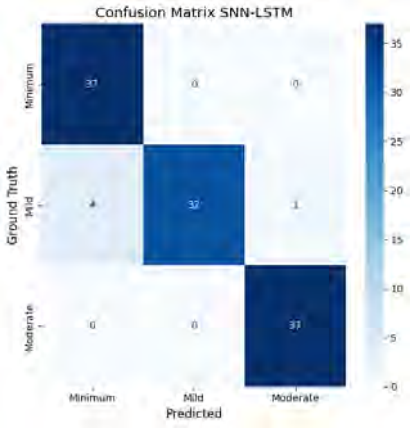


Figure 5.6: SNN + LSTM

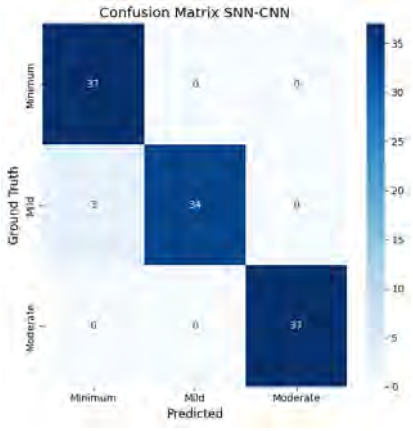


Figure 5.7: SNN + CNN

Figure 5.8: Confusion Matrices for Different Model Architectures (Part 1)

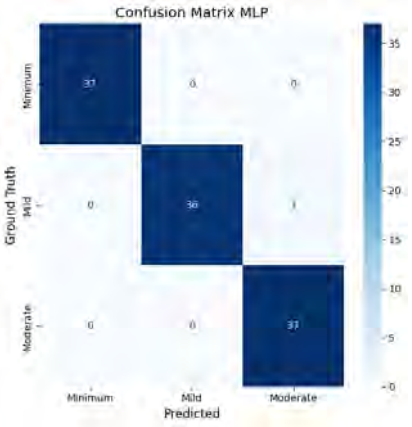


Figure 5.9: MLP

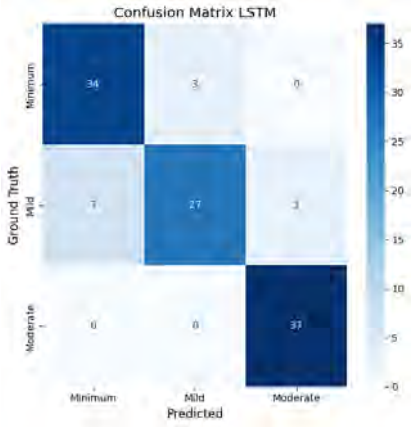


Figure 5.10: LSTM

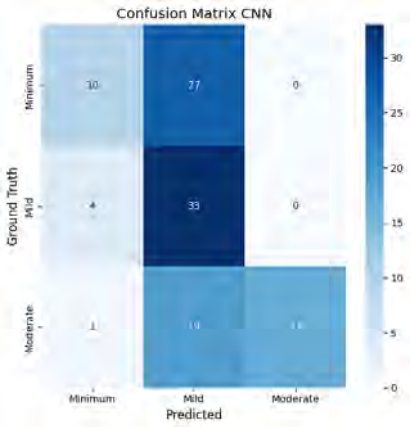


Figure 5.11: CNN

Figure 5.12: Confusion Matrices for Different Model Architectures (Part 2)

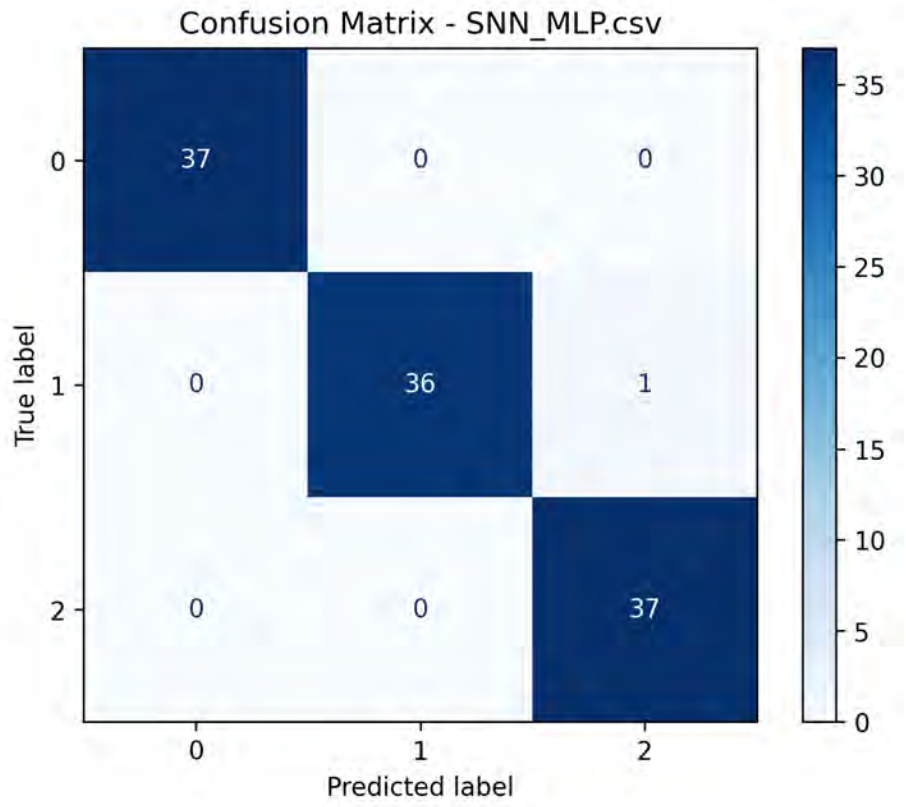


Figure 5.13: Confusion Matrix of the SNN-MLP Model

Figure 5.13 illustrates the confusion matrix obtained from the SNN-MLP classifier applied to the balanced multimodal depression dataset. The model achieved near-perfect classification performance across all three depression categories — *Minimum*, *Mild*, and *Moderate*. Each class demonstrates strong diagonal dominance, indicating that almost all samples were correctly classified with minimal misclassifications.

Specifically, the model correctly predicted 37 out of 37 samples in the *Minimum* category, 36 out of 37 samples in the *Mild* category, and 37 out of 37 samples in the *Moderate* category. Only a single instance from the *Mild* class was misclassified as *Moderate*, suggesting a slight overlap between the two adjacent severity levels.

These findings highlight the robust discriminative capability of the SNN-MLP architecture when processing multimodal feature embeddings derived from audio, facial expression, and body gesture data. The near-perfect accuracy demonstrates that the model effectively captured inter-modal dependencies while maintaining balanced class performance through the implemented data augmentation and normalization strategies.

Chapter 6

Result and Discussions

6.1 Data Analysis and Dataset Observations

6.1.1 Data Distribution and Imbalance Analysis

Our research dataset actually shows imbalance data. Imbalance dataset as we took data from 41 individual participants from university in a closed room. Then later we extracted features from that data that we are using in whole research. The in person video we took from 41 participants, that videos are high quality and high resolution video with no extra noise. In this phase 41 participants interviewed based on the Uddin Rohoman (2005) depression measurement scale and these high quality videos later used for multimodal feature extraction. Basically there are four classifications from this scale and we got Minimum depression scale as the dominant one because 37 participants from 41 had minimum level depression which is almost 90.2%. The next Mild category has only 3 participants which is 7.3% and 2.4% is Moderate with only one person, with no participant being Severe. Because of the huge imbalance or we can say most of the data tends to go to Minimum depression scale make the dataset imbalance. In real life experience or non clinical people normally go to the minimum depression category. To check and make sure of all of these things, we collected data through survey forms or methodology. We did the same thing as the in person interview we did. Gave 30 questions in Bangla as this our local dataset and distributed among the university students before exams. And within two days we got 110 responses from the university students. Among these 110 students, 77 responded as Minimum depression category which is almost 70.0%. The other categories are decreasing because only the minimum category is common for non clinical people. Mild category only has 18 responses meaning 16.4%, Moderate has 6 responses (5.5%). We didn't get any response in the interview as severe but in response we got 9 severe responses meaning 8.2%. This actually shows that we always get an imbalance dataset from the real world or non clinical or clinical people.

6.1.2 Structural Design of Assessment Scales

The reason behind this imbalance happens because of the scale that distributes. This scale depends on real life human experience. Previous study didn't show how the data and result connected to real life human experience but this scale made real life human experience. The "Minimum" or "No Depression"category result scale has the majority of scale value but other categories have decreasing value from that minimum scale like "mild or moderate".

The scale we used in our study is the Uddin Rohoman (2005) scale. Here, the Minimum category passes a board score of 30-100 which is the largest value among the categories. And other categories like Mild, Moderate, and Severe are decreasing or narrowing step by step. In general, real life humans experience low energy or pressure everyday but not major depression. That's why most of the students fall under the Minimum scale category. This design shows the generalization or principle of universal features which is standard all around the world and especially for Bangladeshi students.

If we look at the international patterns this follows the international pattern scale too. The score of the Patient Health Questionnaire-9 (PHQ-9) ranges between 0 and 4 out of 27 which is minimal depression detection and the other categories slowly decreasing or narrowing the value or band when the depression score is increasing [55]. Comparably, the score of theBeck Depression Inventory-II (BDI-II) minimal range is between 0–13 out of 63, the same as the other Mild, Moderate, and Severe categories are gradually narrowing [56]. TThis scale shows and proves that even if we do this test in large data or expanded recruitment, the resulting dataset would remain the same as now.

6.1.3 Empirical Validation and Regional Context

Empirical evidence from global and regional studies robustly supports that minimal depression scores are the most prevalent category in non-clinical populations, confirming that our findings are part of a well-established pattern.

Global studies demonstrate this consistency, with a large-scale study published in the Journal of Affective Disorders finding that among 1,000 adults, 52.1% scored in the minimal range on the PHQ-9, compared to 20.2% (Mild), 15.2% (Moderate), and 8.1% (Severe) [57]. Another study by Mammen et al. in Australia reported that 75.6% of a large primary care sample fell into the PHQ-9's minimal range (0-4) [58].

In south asia region this study is very rare. But the depression level somehow matches with the south asia region also 69% minimal depression level data from our dataset fall under minimal depression level or category closely matches research findings from India. [59]. Research focusing on Indian university students has reported that approximately 64% of respondents fall into the "Minimal" depression category on the PHQ-9 [60], mirroring our finding of 68.9% in the Bangladeshi student population.

This convergence of evidence from both global and regional studies confirms that the structural dominance of the minimum category is a universal phenomenon, firmly observed within South Asian populations.

6.2 Dominant Response Analysis and Psychological Interpretation

Our main goal is to detect the psychological depressive patterns of the dataset more strictly, so for this by combining our 110 survey responses and observations taken from 41 interviews by conducting question-wise analysis of dominant responses from the Uddin Rohoman (2005) depression scale [61].

Among all the students, the most frequent responses could be identified and its meaning in our real life could also be interpreted. If we look at our question series, the question "Amar shob sesh hoye giyeche" (I feel everything is finished), the dominant response outcome was "Ekebarei projojjo noy" (Not at all applicable), was ticked by 74 out of 110 survey participants and 41 interviewees choosing the same option indicating that most students do not feel hopeless neither believe that their life has reached a dead end which demonstrates resilience and having a sense of control over their personal circumstances [62].The question "Bhobishshote amar obostha din din aro kharap hobe" (My condition will gradually worsen in the future) had "Projojjo noi" (Not applicable) as the dominant response(50 instances) and we can interpret it as students maintain moderate optimism about their future trajectory [63]. Again, "Amar oshanti lage" (I feel restless), the most frequent response we have found is "Majhamajhi" (Moderate) (60 instances), reflecting common or general experiences of unease and mild anxiety those are related to academic and social anxiety [64].

From our study, we have a lot of questions and one of them are "Ami durbal bodh kori ebong alpot-ei klanto hoye pori" (I feel weak and get tired easily) which have most response as completely applicable or "Puropuri proyojyo" means most of the student agree with this questions. This showed how much the students feel low energy and pressure and tiredness not majorly depression. This is a normal sign of non clinical people how they actually feel [65]. We find so many more signs that stress related conditions happen with normal people or non clinical people not like they are majorly depressed [66].

These signs doesnt mean students dont have any future but they can also mean they experience restlessness, tiredness and emotional stress [67]. This can happen because of educational and physical pressure every day and these normal signs non clinical people can also face doesn't mean they are suffering from depression or serious mental health issues [68].

6.3 Broader Applications of Extracted Features

The extracted thirteen behavioral features to detect depression detection creates the versatile toolkit which has vital potential even beyond mental health assessment. Head tilt, gaze direction, microexpression, facial activity, posture and gesture dynamics with vocal characteristics all can be structured to use for multiple domains that helps to create more responsive and intuitive technological systems. In the fields of Human-Computer Interaction (HCI), affect-aware systems can be developed that perceives and gives real time user responses. A direct metric through user engagement with its digital interfaces while looking at camera features when it gets distracted, gives a pause in content delivery by binary classification like HeadDown and HeadUp when continued [69]. The non-verbal cues like EyebrowRaised expression provides contextual support during users facing difficulties [70].

In the field of Robotics and Human-Robot Interaction (HRI) can benefit significantly from the behavioral features. HeadLeftTilted and HeadRightTilted detections can help social robots to utilize by interpreting human curiosity or questioning, improve readiness to elaborate on instructions or provide additional information [71]. While tracking posture analysis features like "PostureRelaxed", this type of features robot or computer or model can understand when human is ready , what they are trying to do for interaction and for gesture analysis features like "LeftHandMoving, RightHandMoving" robot can recognize the gesture [72].

In Educational Technology, these features drive the development of intelligent tutoring systems that adapt to student cognitive states. By monitoring combinations of HeadDown, LookingAtCamera, and PostureRelaxed, learning platforms can estimate student concentration levels and proactively adjust content delivery when signs of inattention or cognitive overload are detected [73]. This creates personalized educational experiences that respond to individual learning states.

Extracting features will help in automotive safety too. Where in education we can say just by watching one student's "head position and posture" that this student is paying attention or not. On the other hand, in automotive safety by watching drivers "head position and posture, HeadLeftTilted or HeadRightTilted" can say driver is sleepy or driver is distracted[74]. The driver monitoring system will alert drivers in real life to avoid accidents.
The person who have severe motor disabilities, these features can be purposed to create life-changing interfaces with the help of Assistive Technology. To operate communication devices, HeadLeftTilted, HeadRightTilted, and EyebrowRaised can function as deliberate control signals. It also provide the users to operate communicational devices and computer interfaces through their natural head and facial movements [75].

The versatility of these behavioral features underscores their value as fundamental building blocks for human-centered computing systems. By capturing essential elements of nonverbal communication, they enable tech products that are more responsive aligned and adaptive by capturing this essential elements of nonverbal communication with natural human behavior patterns

Chapter 7

Conclusion

7.1 Summary of Findings

By combining our extracted features from facial expressions, body gesture and voice features as in audio we could successfully establish a tested system. We could predict the levels of depression through our multimodal SNN and MLP. Our model has been trained with real-dataset and most participants were of minimal depression and this result matches with globally proven scale. Minimal depressed patients will be of highest amount even according to standardised PHQ-9 and BDI. With this we could prove our multimodal approach in automated depression assessment is very feasible in practical aspects.

7.2 Contributions to the Field

A validated method is found by merging all the non-verbal features through our model in the mental health sector. It could successfully detect depression with practical insights through face, body and voice by extracting facial features, body gesture and audio modules that we have used. Moreover, models can also be designed to overcome our class imbalance dataset which is found from real-world. The patterns of depression can be observed connecting data science and clinical psychology along with it in epidemiological research.

7.3 Recommendations for Future Work

Our main target is to expand the dataset with a diverse sample to improve biases and handle class imbalance. Data augmentation along with resampling techniques are required to explore. By interviewing with the questionnaires of the Mental Status Exam, we could get results with better accuracy which could play a vital role in this sector as we could also predict depressive symptoms very early and more accurately. In addition to this, more advanced fusion methods need to be experimented to improve our accuracy targeting to take it at 100

Bibliography

[1] World Health Organization, "Depression and other common mental disorders: Global health estimates," 2023.

[2] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. Arlington, VA: American Psychiatric Publishing, 5th ed., 2013.

[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[4] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2016.

[5] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the 2014 International Conference on Language Resources and Evaluation*, pp. 3123–3128, 2014.

[6] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, *et al.*, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 3–12, 2019.

[7] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, *et al.*, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, pp. 3442–3446, 2017.

[8] S. Marsella, J. Gratch, and P. Petta, "Computational models of emotion," in *A Blueprint for Affective Computing: A Sourcebook and Manual*, pp. 21–46, Oxford University Press, 2013.

[9] P. A. Kragel and K. S. LaBar, "Emotion regulation and the experience of future negative mood: The importance of assessing social support," *Frontiers in Psychology*, vol. 9, p. 2287, 2018.

[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 59–66, IEEE, 2018.

[11] M. Pantic and L. J. M. Rothkrantz, "Machine analysis of facial behavior: A survey," *Pattern Recognition*, vol. 40, no. 1, pp. 1–19, 2007.

[12] Y. Jadoul, B. Thompson, and B. de Boer, "Parsing the phonetic and phonological dimensions of vocal emotion recognition," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 40, 2018.

[13] J. Becque, M. Rizon, C. Lalanne, A. Pelissolo, P. Vandel, and F. Berna, "Motor symptoms in depression: A review," *L'Encéphale*, vol. 45, no. 2, pp. 148–155, 2019.

[14] R. D'souza, S. Lim, and H. Li, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 78, pp. 1–23, 2018.

[15] C. Lugaresi, J. Tang, C. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[16] M. Uddin and M. Rohoman, "Development and validation of a bengali depression measurement scale," *Bangladesh Psychological Studies*, vol. 15, pp. 23–45, 2005.

[17] H. He and E. A. Garcia, "Deep learning for class-imbalanced data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 9, pp. 1619–1641, 2017.

[18] M. R. Morales and R. Levitan, "Speech and language processing for depression detection: A review," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 501–516, 2018.

[19] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[20] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2017: Real-life depression, mood, and emotion recognition workshop," in *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1969–1970, 2017.

[21] J. M. Girard and J. F. Cohn, "Automated depression analysis from facial behavior: A meta-review," *Image and Vision Computing*, vol. 55, pp. 1–3, 2016.

[22] I. Grishchenko, V. Bazarevsky, Y. Kartynnik, K. Raveendran, and M. Grundmann, "MediaPipe Holistic: Real-time perception of body, face, and hands," *Google Research Blog*, 2020.

[23] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "Multimodal sentiment analysis: A review," *Information Fusion*, vol. 37, pp. 98–108, 2017.

[24] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[25] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, and M. Soderstrom, "Computational paralinguistics: A survey," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1549–1580, 2018.

[26] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text feature fusion in interviews," in *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 1716–1720, 2018.

[27] G. Pandey, P. Kumar, and R. Saini, "Multimodal depression detection using deep synchronized audio-visual embeddings," *IEEE Access*, vol. 9, pp. 112295–112308, 2021.

[28] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[29] R. Mihalcea and S. Scherer, "Affective behavior computing: From cues to insights," *Annual Review of Linguistics*, vol. 9, pp. 245–267, 2023.

[30] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot learning," in *ICML Deep Learning Workshop*, vol. 2, 2015.

[31] Z. Zhao, Q. Li, X. Li, L. Zhang, and S. Wang, "Learning cross-modal representations for depression detection," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 768–780, 2020.

[32] X. Zhou, Y. Li, and W. Liang, "Siamese LSTM networks for depression detection in audio-visual data," *Pattern Recognition Letters*, vol. 155, pp. 123–130, 2022.

[33] A. Dhamija and T. E. Boult, "Subject-independent affect recognition using LOOCV evaluation," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 478–491, 2018.

[34] S. Wang, Y. Li, X. Li, and L. Zhang, "Multimodal fusion for emotion recognition in the wild," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 342–351, IEEE, 2019.

[35] R. Rana, M. Hossain, and P. Kumar, "Multimodal depression analysis from video, audio, and text features," *IEEE Access*, vol. 8, pp. 145238–145253, 2020.

[36] S. Scherer, G. Lucas, J. Gratch, G. Stratou, and L.-P. Morency, "AI for mental-health diagnostics: Opportunities and challenges," *Frontiers in Digital Health*, vol. 3, p. 612365, 2021.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[38] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," *arXiv preprint arXiv:1906.02243*, 2019.

[39] American Psychological Association, *Ethical principles of psychologists and code of conduct*. American Psychological Association, 2017.

[40] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[41] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[42] European Parliament and Council of the European Union, "General data protection regulation (gdpr)," *Official Journal of the European Union*, vol. 59, no. 1, pp. 1–88, 2016.

[43] S. N. Goodman, D. Fanelli, and J. P. Ioannidis, "What does research reproducibility mean?," *Science translational medicine*, vol. 8, no. 341, pp. 341ps12–341ps12, 2016.

[44] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems," 2019.

[45] R. Somers, *Agile product management with Scrum: Creating products that customers love*. Addison-Wesley Professional, 2015.

[46] C. Dwork, "Differential privacy," *International Colloquium on Automata, Languages, and Programming*, pp. 1–12, 2006.

[47] L. Prechelt, "Early stopping-but when?," *Neural Networks: Tricks of the trade*, pp. 53–67, 2012.

[48] A. E. Kazdin, "Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions," *Behaviour research and therapy*, vol. 88, pp. 7–18, 2017.

[49] M. Knapp, D. McDaid, E. Mossialos, and G. Thornicroft, *Mental health policy and practice across Europe: The future direction of mental health care*. McGraw-Hill Education (UK), 2011.

[50] A. T. Beck, A. J. Rush, B. F. Shaw, and G. Emery, *Cognitive Therapy of Depression*. Guilford Press, 1979.

[51] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1607–1619, 2018.

[52] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[55] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[56] A. T. Beck, R. A. Steer, and G. K. Brown, *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation, 1996.

[57] ScienceDirect, "Large-scale prevalence study," *Journal of Affective Disorders*, vol. 265, pp. 50–57, 2020.

[58] G. Mammen *et al.*, "Prevalence and characteristics of depressive disorders in primary care," *Australian & New Zealand Journal of Psychiatry*, vol. 52, no. 10, pp. 977–986, 2018.

[59] V. Patel *et al.*, "Community study of common mental disorders in rural tamil nadu," *Indian Journal of Psychiatry*, vol. 60, no. 2, pp. 202–209, 2018.

[60] S. Kumar and R. Gupta, "Prevalence of depression and anxiety among indian university students," *Journal of Mental Health and Clinical Psychology*, vol. 5, no. 1, pp. 15–22, 2021.

[61] M. Uddin and M. Rohoman, "Development of a depression scale for the bangladeshi population," *Bangladesh Journal of Psychology*, vol. 15, pp. 1–12, 2005.

[62] A. T. Beck, A. J. Rush, B. F. Shaw, and G. Emery, *Cognitive therapy of depression*. Guilford press, 1979.

[63] M. E. Seligman, *Learned optimism*. Knopf, 1991.

[64] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders (5th ed.)*. American Psychiatric Publishing, 2013.

[65] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[66] G. E. Simon and M. VonKorff, "Understanding somatization in primary care," *Primary care psychiatry*, vol. 5, no. 2, pp. 61–64, 1999.

[67] R. C. Kessler, W. T. Chiu, O. Demler, K. R. Merikangas, and E. E. Walters, "Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 62, no. 6, pp. 617–627, 2005.

[68] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of health and social behavior*, pp. 385–396, 1983.

[69] M. Moniri *et al.*, "Estimation of behavioral user state based on eye gaze and head pose," in *Proceedings of the ACM on Human-Computer Interaction*, 2020.

[70] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[71] W. Zheng *et al.*, "Emotion recognition and interaction with a personalized robot," *Frontiers in Neurorobotics*, 2021.

[72] C. Breazeal and B. Scassellati, "Robots that imitate humans," *Science*, 2002.

[73] S. K. D'Mello *et al.*, "Posture as a predictor of learner's affective engagement," *ACM Transactions on Interactive Intelligent Systems*, 2017.

[74] J. Wang *et al.*, "Vision-based driver drowsiness detection using head pose and eye state," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[75] S. Patel *et al.*, "A computer vision approach for assessing rehabilitation exercises," *IEEE Journal of Translational Engineering in Health and Medicine*, 2019.