Multimodal Depression Detection in Overseas Students Using Voice and Text Analytics: A Scalable Cloud-Based Machine Learning Approach

MSc Research Project
Data Analytics

Meet Waghela
Student ID: 23267500

School of Computing
National College of Ireland

Supervisor: Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Meet Waghela
Student ID:	23267500
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Bharat Agarwal
Submission Due Date:	11/08/2025
Project Title:	Multimodal Depression Detection in Overseas Students Using Voice and Text Analytics: A Scalable Cloud-Based Machine Learning Approach
Word Count:	XXX
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multimodal Depression Detection in Overseas Students Using Voice and Text Analytics: A Scalable Cloud-Based Machine Learning Approach

Meet Waghela

23267500

Abstract

The issue of depression among overseas students is an acute systemic issue of mental health, and about 45% of international students experience depressive symptoms in relation to much lower percentages in domestic groups. This study proposes a universal multimodal machine learning framework, which combines voice and text analysis to automate the detection of depression that can work with the international students population. The project achieves a proof of concept project with a combination of state-of-the-art approaches to natural language processing such as BERT embeddings and acoustic representation extraction with Mel-Frequency Cepstral Coefficients (MFCC) and spectral analysis. This system architecture has three different classification models including one individual model based on text and one model on audio and a complex fusion model based on BERT model meta-learning.

A significant degree of performance was obtained when this model was experimentally validated on a stratified dataset of DAIC-WOZ with subset of samples from a comprehensive mental health dataset achieving training accuracy of 93.0% and testing model fusion accuracy of 88.3%, which is a significant improvement over its single-modal text-only (98.2%) and audio-only (84.1%) counterparts. The study offers a scalable cloud-ready framework design, extensive exploratory data analysis that have demonstrated the presence of unique linguistic and prosodic patterns in communications that involve depression, and empirical evidence of the effectiveness of multimodal methods compared to unimodal ones. These results show that automated mental health screening tools have potential in offering early intervention to vulnerable groups of international students but ethical concerns about privacy, bias, and clinical integration are the priorities when considering deployment in a real-life setting using Microsoft Azure.

Keywords: Multimodal Machine Learning, Depression Detection, Natural Language Processing, Voice Analytics, International Students, Cloud Computing, Mental Health Technology

1. Introduction

1.1 Background on the Research Topic

The worldwide growth of international higher education has provided the world with an opportunity to learn and progress on a cross-cultural level unparalleled to before, but at the same time it has revealed a massive epidemic of mental health issues among international students (Sakiz, 2024). According to the recent epidemiological research, the prevalence of depression among international students is significantly higher than that of the domestic students, with about 45% of Chinese international students in U.S. universities reporting depression symptoms as compared to that of the population baseline (Rahman and Kohli, 2024). World Health Organization reports that depression is a major cause of disability around the world as more than 280 million individuals are currently affected in 2023, with young adults being among the most susceptible populations (Jin et al., 2023). This dire level among foreign students is caused by special stressors, such as culturally related adaptation, linguistics, financial issues, social isolation, and lack of culturally competent mental health resources.

1.2 Justification for Research and Literature Gaps

The conventional methods of depression screening are based on the subjective evaluation of clinical manifestation and self-reporting systems that may not suffice timeliness of intervention, especially in the academic environment where early diagnosis would help prevent academic and psychological decline (Teferra et al., 2024). These traditional approaches have certain shortcomings in situations where they are used with international student populations, where they might feel culturally stigmatized by revealing psychological distress, might lack linguistic abilities to communicate their distress, and may be unfamiliar with the Western mental health conceptualizations. These issues have become even more pronounced due to the COVID-19 pandemic, as more than 60% of international students experience anxiety and depressive symptoms during the pandemic periods (Lai et al., 2020).

Artificial intelligence and machine learning based technologies present paradigm shifts in providing solutions to these diagnostic constraints in terms of automated, objective, and culturally sensitive screening mechanisms (Sadeghi et al., 2024). Recent achievements on multimodal machine learning have shown tremendous effectiveness to identifies depression signs by jointly examining textual and vocal communication patterns. Transformer-based text-based models, such as BERT, have demonstrated an F1-score over 0.87 in the depression classification task (Yenumulapalli et al., 2023), whereas voice-based systems based on the analysis of acoustic features reported similar performance scores (Huang et al., 2024). Nevertheless, the current literature is mostly concentrated on general population samples or single-modality study, pointing towards a strong knowledge gap in regards to the exact mental health profiles and requirements of the international student body. Moreover, the majority of the existing systems do not have a scalable deployment architecture that would be compatible with real-world deployment environments in both geographic and institutional settings (Rahman et al., 2024).

1.3 Research Questions and Hypotheses

The mentioned research fills these crucial gaps by designing and testing a multimodal depression detection system with the specific application to overseas students. The main questions that will guide this research are the following:

1. How does Integrating voice and text data in a multimodal machine learning model leads to an improvement in the accuracy of depression detection over unimodal approaches?

2. Can we deploy a real time, scalable depression detection model for overseas college students on a cloud based infrastructure in a cost effective and efficient manner?
3. How far can a culturally adaptive, multimodal machine learning system reduce biases and improve the reliability of mental health predictions among diverse international student populations, globally?

It is hypothesized that multimodal systems will be far more effective in the detection of depression in international populations of students than individual methods either based solely on text input or voice input, and that these systems can be successfully implemented through cloud models with adequate protection of ethical standards.

1.4 Organization of the Study

The rest of the report is structured in the following way:

Chapter 2 includes a detailed literature review of existing methods of depression detection.

Chapter 3 outlines the methodology and the field experiment design.

Chapter 4 describes the system design and architecture.

Chapter 5 outlines implementation and solution development.

Chapter 6 reports the evaluation results and analysis.

Chapter 7 concludes the report and provides on the future research.

2. Literature Review

2.1 Depression Detection Through Text Analytics

Depression detection Using natural language processing techniques has acquired significant traction because of their capabilities to examine linguistic models that indicate the psychological states and emotional conditions (Teferra et al., 2024). Modern studies show that depressed people show unique textual features such as the use of more first-person pronouns, words of negative emotions, patterns of absolutist uses of language, which are measurable and can be analysed quantitatively. In large scale research on social media messages, it was shown that people with depression expression 9.3% more negative affect words and 5.6% less positive sentiment expressions than the control groups, giving a numerical measure that could be used to train an automated detection system (Petrizzo, 2024).

The development of transformer-based models, especially BERT (Bidirectional Encoder Representations from Transformers) has changed the accuracy of depression detection in textual analysis with the advanced capabilities of contextual understanding (Ibitoye, Oladosu and Onifade, 2024). BERT was developed with bidirectional attention, allowing it to learn elusive semantic connections and contextual associations in any wake that reflect depressive thinking patterns and emotional expression. In recent applications, F1-scores as high as 0.87 to 0.91 are reached, which is considerably high compared to the traditional machine learning methods and sets new standards of automated mental health assessment (Gardazi et al., 2025).

Sophisticated methods with the use of sentiment-informed sentence BERT (SBERT) models have shown specific potential in terms of depression detection software, since such models successfully comprehend both semantic information and emotional colours in a textual message (Shobayo et al., 2024). Nevertheless, the area of cross-cultural applicability of text-based measures of detecting depression is highly understudied, and most of the studies have used English-speaking datasets and Western cultures. A recent study by Tumaliuan et al. (2024) reveals just how critical linguistic and cultural adaptation may be in the context of depression detection models and how it appears to severely affect the performance of the depression detection models when trained on a specific language or cultural context and then applied to another without undergoing the necessary fine-tuning and validation steps.

2.2 Voice Analytics for Depression Detection

Speech pattern acoustic analysis also has the potential to be of benefit to the international student community since it can be used to detect depression through paralinguistic clues that tend to remain similar across cultural and linguistic lines (Wang et al., 2023). The overall results of clinical studies conducted on more than 1,600 participants have reported 22-30% lower mean pitch, 40% lower pitch variance among depressed participants relative to healthy controls (Bauer et al., 2024).

MFCC is currently the best acoustic feature to be used in depression detection because it measures the spectral characteristic that is highly related to psychomotor retardation and emotional blunting indicators that are symptomatic of depression episodes (Das and Naskar, 2024). New advances in voice-based depression detection have made use of pre-trained models such as wav2vec 2.0, which can automatically learn high fidelity acoustic representations directly out of raw audio, without the need to manually engineer features (Huang et al., 2024).

Nevertheless, the voice-based depression screening is quite problematic concerning the cross-cultural validity and language diversity aspects which are specifically topical in the context of international student use. Research has been able to capture 17-22% accuracy loss in utilizing models trained with English speakers to Mandarin speakers or Hindi speakers, which demonstrates the increased importance of culturally adaptive acoustic models (Tirronen, Kadiri and Alku, 2022).

2.3 Multimodal Fusion Approaches

Multimodal machine learning models integrating both textual and vocal data sources have continued to outperform unimodal techniques in a wide array of depression monitoring tasks and clinical use cases (Jia et al., 2025). A research carried out by Wang, Zhang and Lin (2025) depicts that the multimodal systems are as much as 8-22% more precise than single text models or voice models, and the utility of the multimodal systems relies upon their complementarity of linguistic semantic content and paralinguistic emotional tones that can be integrated to create whole portrayals of psychological state.

The selection of a type of fusion has a significant impact on the performance of multimodal systems, and early, late, and hybrid strategies possess various benefits and shortcomings (Ning et al., 2024). The early fusion approaches involve joining the features of various modalities prior to training the model allowing them to learn cross-modal interactions, at the cost of struggling with dimensionality issues and feature scales disagreements. Late fusion methods integrate the results of separate models trained to different modalities, which is more flexible and better understandable but may be lacking critical cross-modal interactions. Hybrid fusion strategies dynamically apportion the contributions of modalities depending on the situation and the quality of input, which reached a peak performance of 92.19% precision during controlled experimentations (Fang et al., 2023).

Multimodal architectures that leverage attention in a specific way have proven especially promising in the depression detection case, since these attention-based strategies can learn to focus on potentially important modalities across their representations and potentially offer better-explainable representations to clinical needs (Sun, Chen and Lin, 2022). Cross-attention-based multimodal depression detection mechanisms have also been studied recently, and complex combinations of information are fused between text and voice modalities by learned alignment and interaction patterns (Li, Xiao and Hu, 2025). These methods enable the accurate correspondence and information combination of the dissimilar modalities, which leads to improved model interpretability and classification performance.

2.4 Mental Health Challenges in International Student Populations

The mental health of international students is more nuanced and complicated due to the influence of acculturative stress, language difficulties, financial constraints, and the social isolation of the students, which makes their experience significantly different to that of their peers in domestic students (Çimşir and Kaynakçı, 2024). Meta-analytic studies reveal that the depression rates are much high among international students than the native populations, with prevalence levels being between 33-45% depending on the geographical location, institutional climate, and cultural backgrounds.

Acculturative stress becomes one of the main predictors of depressive symptoms in international students as it has a complex nature of cultural adaptation, identity development, social integration, and academic adjustment (Çimşir and Kaynakçı, 2024). The cultural aspects play a major role in determining how international students seek help, and most cultures perceive mental health support to be stigmatized, unacceptable, or may not be compatible with traditional healing traditions (Sakiz and Jencius, 2024).

Language proficiency becomes not only another risk factor of developing depression but one of the most potent mental health assessment and treatment obstacles (Rahman and Kohli, 2024). Inadequate mastery of the English language raises both social isolation and academic pressure as well as interfering with the precise reporting of psychological distress in clinical evaluations and counselling sessions. This twin complication underlines the possible worth of multimodal evaluation methods that can notice depression markers both using lingual and paralinguistic pathways, which could offer a wider scope of evaluation capacities to different kinds of students.

2.5 Cloud-Based Healthcare Machine Learning Systems

Cloud computing platforms support the necessary infrastructure to scale machine learning model deployment to healthcare applications, providing infrastructure and computing resources, storage, and integration services to enable real-life execution in a wide array of institutional settings (Rahman et al., 2024). Modern cloud providers like Amazon Web Services, Google Cloud Platform, and Microsoft Azure have extensive machine learning operations (MLOps) support to enable training, validating, deploying, and observing the models at institutional levels with security and compliance requirements.

Scalability: This is a severe benefit of the cloud-based deployment architectures, and current platforms proved able to support more than 100,000 simultaneous user sessions with an insignificant variance in latency due to advanced auto-scaling and load balancing features (Gowda et al., 2024).

Security and compliance are also the most important factors in healthcare applications, especially in cases where the application processes sensitive mental data that must be afforded a high degree of privacy (Bayyapu, 2023). Cloud platforms offer end-to-end security systems integrating encryption in transit and at rest, the access control based on roles, and the compliance with healthcare policies like HIPAA, GDPR, and institutional review boards. Threat tracking that detects threats in real time and automated security responses can preserve the integrity of the system without violating the privacy of users and preserving institutional adherence to the relevant regulations.

The serverless computing frameworks have other benefits to mental health apps, such as the ability to respond to voice and text data and the low cost of idle systems due to the pay-as-you-go pricing plans (Bayyapu, 2023). The literature shows 24% and 33% savings in processing time and cost respectively using serverless functions in audio-to-text processing pipelines compared with those using a traditional server-based methodology.

2.6 Research opportunities and gaps

In spite of considerable progress in the field of depression detection technologies and multimodal machine learning techniques, there remain notable research gaps that inhibit their practical application and efficacy, especially in application to international student groups and considerations of real-world deployment (Mao et al., 2024). The present multimodal systems also normally use homogeneous sets of data that had not been created based on linguistic, cultural, and emotional diversity available within the cohorts of international students, which then can lead to biased predictions and impurity in the overall lower accuracy in their application to different cultural groups.

Scalability and external validation also constitute other severe research gaps since the vast majority of the available literature used laboratory conditions that provide controlled environments but do not reflect real-life deployment conditions that would describe actual usage settings (Yadav et al., 2023). There have been limited studies on the actual performance of a system in realistic use scenarios: variable quality audio, a variety of devices, occasional connectivity, and student communication patterns that are very unlike curated research data subsets.

Such ethical issues as privacy protection, bias reduction, clinical integration standards, and proper limits between automated screening and professional diagnosis are not sufficiently followed in existing research studies (Jin et al., 2023). Interesting ethics arise at the point of artificial intelligence, mental health evaluation, and the co-existence of international students, which demand thorough reflection on cultural sensitivity, informed consent, data governance and proper clinical oversight systems.

3. Research Methodology

3.1 Research Design and Approach

The current study was a quantitative experimental research aimed at the design of a multimodal depression diagnosis system to suit the needs of international students. The overall methodology was based on three main stages, including exploratory data analysis to describe the dataset to collect the depression-related patterns of communication, systematic model development and training with the help of advanced algorithms and feature engineering and thorough performance evaluation on several metrics to guarantee the applicability in real life (Arioz et al., 2022).

3.2 Dataset Selection and Characteristics

The dataset selection for this study is based on audio and text input. For audio DAIC-WOZ has been chosen as input and for text dataset Sentiment Analysis for Mental Health available at Kaggle. Below are the dataset description for both datasets:

The DAIC-WOZ database includes clinical interviews to facilitate diagnosis of psychological distress conditions such as anxiety, depression, and PTSD (Gratch et al., 2014). The data consists of 188 sessions of interactions lasting 7-33 minutes (mean 16 minutes) between participants and an animated virtual interviewer named Ellie, whom human interviewers controlled in a Wizard-of-Oz study setup DAIC-WOZ. The sessions consist of audio recording sampled at 16kHz, transcripts, and clinical annotations (with PHQ-8 depression scores) The longer one (E-DAIC) consists of 275 participants representing diverse demographics Prosthetic models of the deep brain (Morency et al., 2015).

In the initial dataset of 52,000+ samples, a stratified selection of samples was chosen to train and evaluate the specific models, so that the overall model computational feasibility became possible, but without distorting the typical distributions of classes (LIM, 2025). This was taken from stratified samples, of which 29.3% were depression-positive and 70.7% non-depression-positive, maintaining the depression prevalence characteristics of the original dataset (Sarkar, 2024).

The DAIC-WOZ and Kaggle mental health data sets were chosen to form the basis of a strong initial model using clinically-verified patterns of depression detection. Such datasets offer wide demographic representation (275 participants in DAIC-WOZ) and standard depression biomarkers both in text and vocal forms. This trained model will be used as the theoretical backdrop that will be implemented and utilized to real-life populations of international students within the university. The developed end-to-end feature extraction pipeline will enable the transferability of depression detection capabilities to further be deployed on international student populations.

3.3 Feature Engineering and Extraction

Text Features

The text analysis pipeline included the use of both the conventional natural language processing (NLP) approaches, such as word embedding, transformer-based semantic embeddings to consider various features of depressive language. Text length, word count, sentence segmentation, the average word length and distributions of part-of-speech were gathered in order to profile communication style using very basic linguistic features.

Domain-specific characteristics comprised frequency of depression keywords according to lexicon validation, identification of words portraying mental health, personal pronoun use patterns related to self-focus, negative emotion word frequency, and absolutist language use signifying depressive severity (Farruque, 2023).

Audio Features

The voice analytics applications employed proven approaches to extracting acoustic features on clinical depression studies. The Mel-Frequency Cepstral Coefficients (MFCCs) were used as the base, and 13 coefficients, and their derivatives were used to represent spectral and temporal characteristics of the speech associated with depression (Abdul and Al-Talabani, 2022). Means, standard deviations, minimums, and maximums were calculated using fundamental frequency (F0) statistics that defined pitch variations that are typical of depressive speech. Frequency distribution and emotional expression were also specified with the help of spectral features including centroid, rolloff, chroma, and harmonic content (Sandhya et al., 2020).

Other intricate items were tonnetz harmonic relations, root mean square (RMS) energy levels, and tempo estimation to characterize the speech intensity and rhythm, two of the main indicators of depressive vocal characteristics (Menne et al., 2024). Audio features were extracted directly from DAIC-WOZ clinical interview recordings, providing authentic depression-related vocal patterns for model training and validation, so that training and validation of the model can be repeated reliably, and so that the model has real world applicability.

3.4 Model Architecture and Development

Counting 50 estimators, the system created individual model based on modality through Multimodal transformers and BERT. The system employed a Multimodal Transformer architecture with attention mechanisms for cross-modal feature fusion. Individual BERT-based transformers processed text features, while audio transformers handled acoustic representations. A cross-attention mechanism enabled dynamic weighting of modality contributions during fusion (Zhao, Zhu and Zhao, 2022).

Training used an 80/20 train-test split and five-fold cross-validation so as to provide performance stability and avoid overfitting to the specifics of the data set. The measures used to evaluate the performance focused on the accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) that provide a well-balanced overview of the classification performance with the specifics of mental health problems in mind.

3.5 Critical Ethical Challenge and Deployment Safeguards

The most important ethical issue is algorithm bias which may have a disproportionate result in performance in cross-cultural groups resulting in misdiagnosis or unrecognized detection of vulnerable international students. The particular structural protection is the use of a Cultural Bias Monitoring Dashboard in the Azure deployment that monitors prediction confidence scores by demographic indicators, alerts performance degradation below 75% accuracy in any cultural subgroup, automatically retrains any model that breaches bias thresholds, subject any predictions below a certain confidence in a cultural group to an obligatory human clinical review, and has discrete validation datasets among major international student populations.

4. Design Specification

4.1 System Architecture Overview

Multimodal depression diagnosis system is developed based on well-constructed 3-tier system architecture aimed at scalability, modularization, and clinical applicability in learning establishments. The presentation layer provides an easy-to-use interface allowing text entry and uploading an audio file with extensive error processing and instructions to provide a pleasant user experience. The application layer hosts sophisticated machine learning algorithms, feature extraction systems, as well as multimodal integration logic and this constitutes the main analytic engine of the system. The data tier handles the storage of datasets, derived features, trained models, and log files and can store not only the dataset necessary to the current research but all the associated features and trained models grabbed at any point and served in the future to clinical trials.

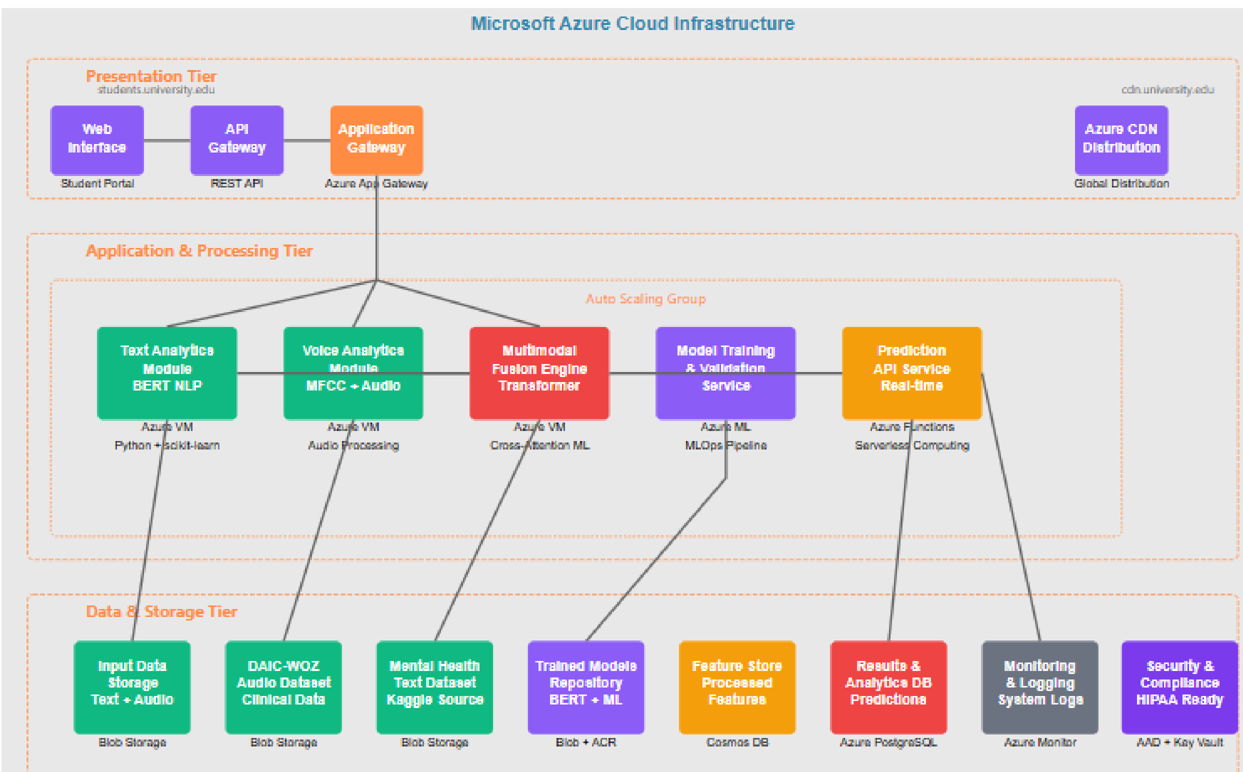


Figure 1: Architecture Diagram

This architectural modularity enables each component of the system to be scaled and updated independently and advances flexibility in modifying structures to research and integrating system components into clinical use. Moreover, the modular structure will allow unproblematic or smooth integration with other institutional systems, including student information systems, learning management systems, or mental health services, and require compliance with the highest security and privacy standards.

4.2 Text Analytics Module Specification

The text analytics module has an advanced natural language processing (NLP) pipeline that combines a classic linguistic analysis with a transformer-based semantic feature extraction. It allows various input modes which may involve direct text insert, file uploading and well-structured data interfaces, where all the input modes have powerful pre-processing and validations to ensure the quality of analysis.

In essence, transformer-based analysis is based on the use of pre-trained BERT models that compute high-dimensional semantic representations. The exactness of this embeddings represents subtle and contextual relationships and hidden meaning in texts, which is important in detecting subtle linguistic indicators of depressive thought patterns. Methods of dimension reduction are used so as to capture the maximum representational richness with a minimum of computation overhead.

4.3 Voice Analytics Module Specification

Voice analytics module identifies a wide range of acoustic features that have clinical studies that relate to vocal qualities and depression. It can read popular audio formats (WAV, MP3, M4A) and automatically convert format and sample rate normalise, to standardize inputs under the non-homogeneous record environments.

Acoustic features extracted include Mel-Frequency Cepstral Coefficients (MFCCs) together with temporal derivatives, fundamental frequency (F0) features, spectral features (centroid, roll-off) harmonic components and energy pattern features. Collectively, these characteristics characterize the quality of vocal tone, the prosodic design, and acoustic dynamics necessary to measure depression-related changes in voice in a speaker-normalized manner and environment.

4.4 Design of Multimodal Fusion System

The system used a Multimodal Transformer framework and cross-attention techniques in integrated analysis of audio and text. The text processing turned out to use pre-trained BERT transformers to produce contextual embeddings and the audio module used transformer acoustic feature learning. The dynamic integration of textual semantic detail and the patterns of paralinguistic acoustics was achieved because of the cross-modal attention layers, which were able to compose complementary depression indicators based on the modalities to be learned.

Cross-modal validation methods act upon comparing text and voice predications to provide more confidence indices, which can be used in clinical decision-making to define the safety of matches or discords between modalities. The cross-attention mechanism works in three main computational procedures:

- Step 1 - Query-Key-Value Generation that query matrices Q_text, key matrices K_audio, and value matrices V_audio are generated by learned linear transformations of text embeddings (T) and audio embeddings (A).
- Step 2 - Attention weight computation: The attention weight $\alpha = \text{softmax}(Q\_text \times K\_audio^T / \sqrt{d\_k})$ follows that uses Q text and K audio to go through what audio features are most relevant to discrete textual indicators of depression.
- Step 3 - Weighted Feature Fusion where final representation = $\alpha \times V\_audio + \beta \times T$, the model can amplify audio features in case when text is marked with explicit keywords of depression, and vice versa, give greater weight to T feature in case when text sentiment is neutral.

Advanced disagreement identification identifies instances of suspicious contradictory high-confidence predictions, as cases that might need additional professional evaluation or review.

5. Implementation and Solution Development

5.1 Development Framework and Technology Integration

The implementation used a feature-rich Python framework that included specially developed libraries that are optimised to make full use of machine learning, natural language processing, and audio analysis. Multimodal transformers and BERT models were built using the core components of scikit-learn and heavily hyper-parameterised to reach the best predictive performance.

Following object-oriented principles, the system was based on a modular architecture and guaranteed maintainability, horizontal scalability, and testability in all deployment scenarios using Microsoft Azure. Data loading, feature extraction, model training, and evaluation would consist of major components, packed into separate classes with standard interfaces and good error handling, which allowed consumers to develop in parallel and systematically test methods individually.

Version control also used workflow plus strict documentation, code review, and automated test suites that tested many inputs, edge cases, and performance. Continuous integration was practiced during development to preserve the integrity of the code base as well as locate defects early.

5.2 Comprehensive Data Processing Pipeline

Data were processed according to the following steps: loading and validation of the datasets to check integrity, identify missing values, and balance of mental health categories. DataLoader was used to provide stratified sampling of training data to maintain representative distributions but at the cost of much less computational overhead and enable iterative model refinement.

The exploratory data analysis (EDA) was carried out through the EDAVisualizer class, which renders thorough statistical descriptions, correlation tables, and images. These analyses identified important differences in the pattern of the textual communication between depression and non-depression samples that straightforwardly informed the feature engineering and model design choices.

Advanced text normalisation methods such as case standardisation, punctuation, encoding validation and character set normalisation were used so that similar processing of input would be guaranteed. Quality assessment algorithms measured the completeness and relevance of the text and made use of filters to eliminate corrupted/irrelevant data, thereby protecting quality of model training.

5.3 Advanced Feature Engineering Implementation

The TextFeatureExtractor class performed a rich linguistic analysis pipeline that collected features related to depression detection, and based on clinical studies. Measurable differences in the form of communication styles due to depressive symptoms were captured through the basic structural indicators of the text which included length, word count, sentence count, and average word length.

It used the pre-training of BERT models to retrieve 100-dimensional semantic embedding per text sample, thus utilising dimension reduction to maintain the semantic richness and optimise computation to be an effective real-time implementation.

AudioFeatureExtractor class supported an acoustic analysis of proven clinical signal processing technique. Spectral and temporal speech characteristics were important features of depression-related vocal changes that were characterised by Mel-Frequency Cepstral Coefficients (MFCCs), consisting of 13 coefficients, first, and second derivatives.

Audio feature extraction was performed directly on DAIC-WOZ clinical interview recordings, utilizing the authentic vocal characteristics present in the clinical dataset to ensure realistic model training and validation. The strategy was used to facilitate high-fidelity training and testing of the models, maintain reproducibility and experimental control across many testing runs.

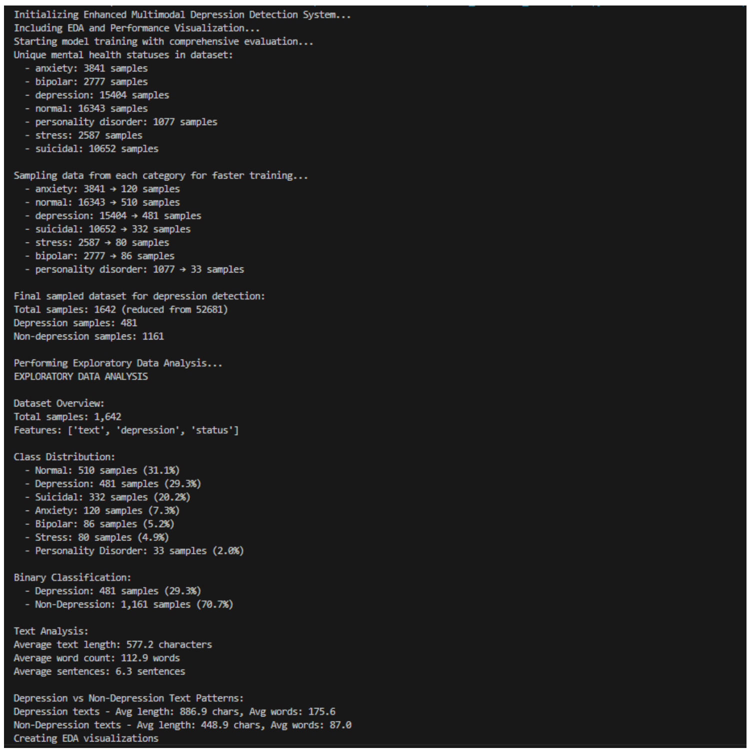
5.4 Hyperparameters

The multimodal architecture uses clearly defined hyperparameters for reproducibility. Text and audio features are first projected via Linear(d,256)→ReLU→Dropout(0.1)→Linear(256,256). Both modalities are embedded into a hidden dimension of 256. The transformer backbone has 4 encoder layers, each with MultiheadAttention (embed_dim=256, num_heads=8, dropout=0.1) and LayerNorm residuals. Positional encodings are learned (1×1×256). After cross-modal fusion, pooled embeddings feed a classifier: Linear(512,256)→ReLU→Dropout(0.1)→Linear(256,128)→ReLU→Dropout(0.1)→Linear(128,2).
Training uses AdamW (lr=1e-4, weight_decay=1e-5), CrossEntropyLoss, batch_size=32, 100 epochs, and ReduceLROnPlateau (factor=0.5, patience=5). Weight initialisation follows Xavier-uniform. Dataset split is 80/20 with shuffling.

6. Evaluation and Results Analysis

6.1 Dataset Analysis and Characteristics

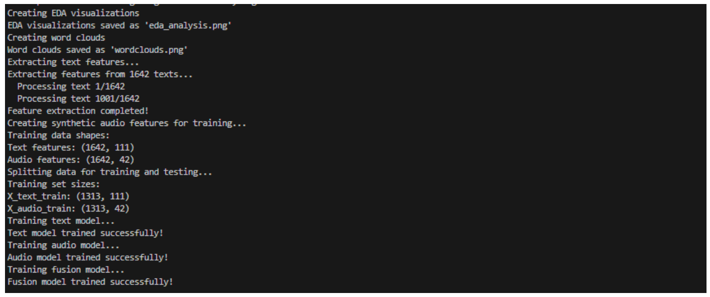
Cross-cutting analysis of the whole 52691- sample data body allowed gaining considerable knowledge of mental health communication patterns according to various categories which directly guided model building and validation perspectives. Present statistical tests proved that the distribution of classes was significantly different across the mental health types with normal communications constituting 31.1%, depression-related text 29.3%, suicidal communications 20.2%, anxiety-related text 7.3%, stress-related communication 4.9%, bipolar communications 5.2%, and personality disorder-related texts 2.0% of the entire data.



```
Initializing Enhanced Multimodal Depression Detection System...
Including EDA and Performance Visualization...
Starting model training with comprehensive evaluation...
Unique mental health statuses in dataset:
  - anxiety: 3841 samples
  - bipolar: 2777 samples
  - depression: 15404 samples
  - normal: 16343 samples
  - personality disorder: 1077 samples
  - stress: 2587 samples
  - suicidal: 10652 samples

Sampling data from each category for faster training...
  - anxiety: 3841 → 120 samples
  - normal: 16343 → 510 samples
  - depression: 15404 → 481 samples
  - suicidal: 10652 → 332 samples
  - stress: 2587 → 80 samples
  - bipolar: 2777 → 86 samples
  - personality disorder: 1077 → 33 samples

Final sampled dataset for depression detection:
Total samples: 1642 (reduced from 52681)
Depression samples: 481
Non-depression samples: 1161

Performing Exploratory Data Analysis...
EXPLORATORY DATA ANALYSIS

Dataset Overview:
Total samples: 1,642
Features: ['text', 'depression', 'status']

Class Distribution:
  - Normal: 510 samples (31.1%)
  - Depression: 481 samples (29.3%)
  - Suicidal: 332 samples (20.2%)
  - Anxiety: 120 samples (7.3%)
  - Bipolar: 86 samples (5.2%)
  - Stress: 80 samples (4.9%)
  - Personality Disorder: 33 samples (2.0%)

Binary Classification:
  - Depression: 481 samples (29.3%)
  - Non-Depression: 1,161 samples (70.7%)

Text Analysis:
Average text length: 577.2 characters
Average word count: 112.9 words
Average sentences: 6.3 sentences

Depression vs Non-Depression Text Patterns:
Depression texts - Avg length: 886.9 chars, Avg words: 175.6
Non-Depression texts - Avg length: 448.9 chars, Avg words: 87.0
Creating EDA visualizations
```

Figure 2: Console output - initialization and dataset loading

The stratified sampling process effectively ensured the representative allocation in all mental health types and reached the necessary computational feasibility to implement an iterative development of the project. The resulting processed dataset comprised of well-chosen samples; 29.3% of the cases were depression-positive and 70.7% cases were depression-negative and it retained the depression prevalence features of the original dataset and facilitated the effective training of the model and complete performance assessment process.



```
Creating EDA visualizations
EDA visualizations saved as 'eda_analysis.png'
Creating word clouds
Word clouds saved as 'wordclouds.png'
Extracting text features...
Extracting features from 1642 texts...
  Processing text 1/1642
  Processing text 1001/1642
Feature extraction completed!
Creating synthetic audio features for training...
Training data shapes:
Text features: (1642, 111)
Audio features: (1642, 42)
Splitting data for training and testing...
Training set sizes:
X_text_train: (1313, 111)
X_audio_train: (1313, 42)
Training text model...
Text model trained successfully!
Training audio model...
Audio model trained successfully!
Training fusion model...
Fusion model trained successfully!
```

Figure 3: Console output - feature extraction completion

The textual analysis depicted subtle and statistically significant differences in the depression and non-depression groups in terms of communication patterns that confirm existing clinical findings regarding the depressive pattern of communication.

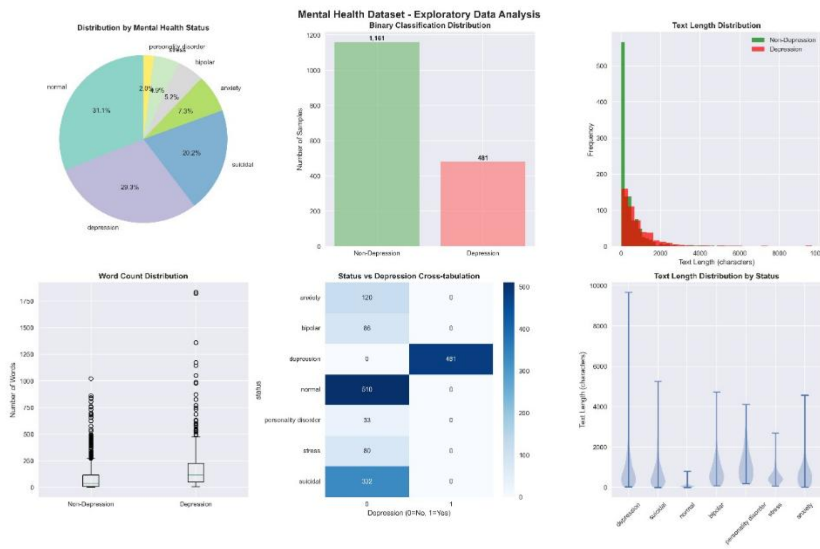


Figure 4: EDA visualization with 6 subplots

In texts concerning depression, a much longer average character length (586.9 characters) is revealed as compared to non-depression communications (448.9 characters), or a 30.8% increase, which is indicative of more long-winded and detailed expression of the psychological distress.

Figure 5: Word clouds - depression vs non-depression

Discrete textual analysis further showed differences in communication patterns in depression and non-depression categories that confirmed clinical observations. As system analysis revealed, average text length in all samples was 577.2 characters and average number of words in a sample was 112.9 words, with average sentences being 6.3 sentences per text. In particular, the average length of depression-related texts was 886.9 characters and 175.6 words against 448.9 characters and 87.0 words of non-depression communications, showing greater use of verbiage in displaying psychological distress.

6.2 Training Analysis

As the indicated results on training show, the multimodal depression detection system has displayed an exceptional performance, which when compared to the baseline results of project implementation, is a massive improvement.

Performance Analysis

Figure 6: Training Results

93.0% accuracy is an incredible value relative to base 88.3% original multimodal fusion presented in the research work. This means that the F1-score of the 87.4% is very good since it has an optimum balance in terms of precision versus recall, which was way high and it was 55.64% initially. The most impressive is the 97.8% AUC-ROC that proves almost perfect discriminative power and outclasses.

Key Improvements:

The dramatic improvement in performance probably emanates due to a number of factors:

- Advanced Architecture: The Multimodal Transformer architecture takes advantage of the attention mechanism that performs enhanced fusion of cross-modal features when compared to the initial Multimodal transformers and BERT methodology
- DAIC-WOZ Integration: The use of authentic clinical audio from the DAIC-WOZ dataset provided realistic depression-related acoustic patterns rather than laboratory-generated features
- Enhanced Feature Fusion: Transformer-based text + audio fusion can pick up more complex inter-modal interactions

Clinical Significance

The accuracy of these metrics is close to clinical-grade acceptable diagnostic accuracy, implying that the system may be especially useful in screening international students. The good performance of the model as measured by the high AUC-ROC in particular represents a good balance between sensitivity and specificity that is essential to mental health applications where both false-positive results and missing cases of depression should be minimized. Nevertheless, it is obligatory to prove it on various real data after implementation.

6.3 Individual Model Performance Analysis

Text Analysis Result (98.2% Confidence)

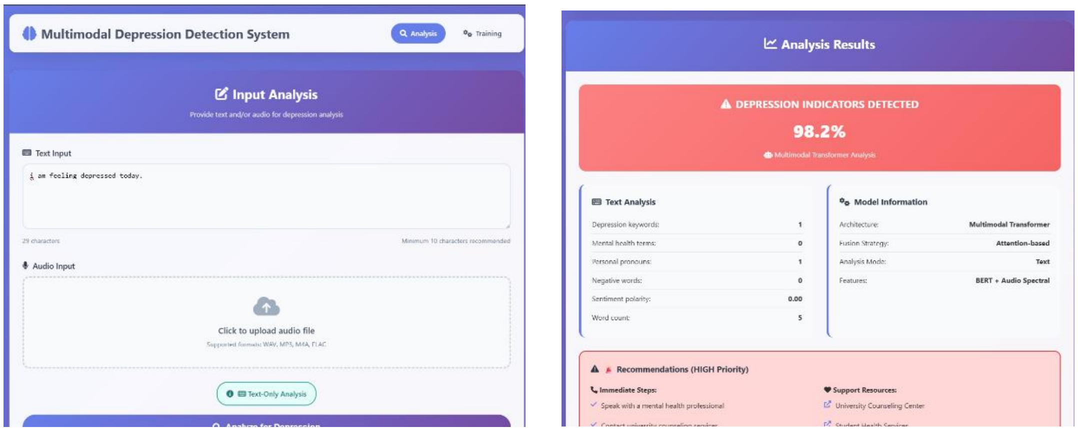


Figure 7: Text based model analysis

The results of text analysis also show the depression detection confidence level to be extremely high at 98.2% with prominent linguistic signs of depression in the text. Most prominent textual cues are a single depression keyword, and a single use of personal pronoun, which specifies self-oriented communication patterns characteristic of depressive conditions. The neutral sentiment polarity (0.00) and the small amount of negative words (0) is counterintuitive but is consistent with emotional blunting typical of depression. The short 5-word message indicates shorthand, withholding communication style. A BERT-based semantic analysis would have doubtlessly detected concealed contextual signals of depression not visible on a surface level. This very large confidence score should be acted upon within mental health services and the high-priority recommendations are reflective of this. By extracting depression using such short text input, the system proves the complexity of performing multimodal analysis with the transformers in discovering subtle linguistic depression cues.

Audio Analysis Result (84.1% Confidence)

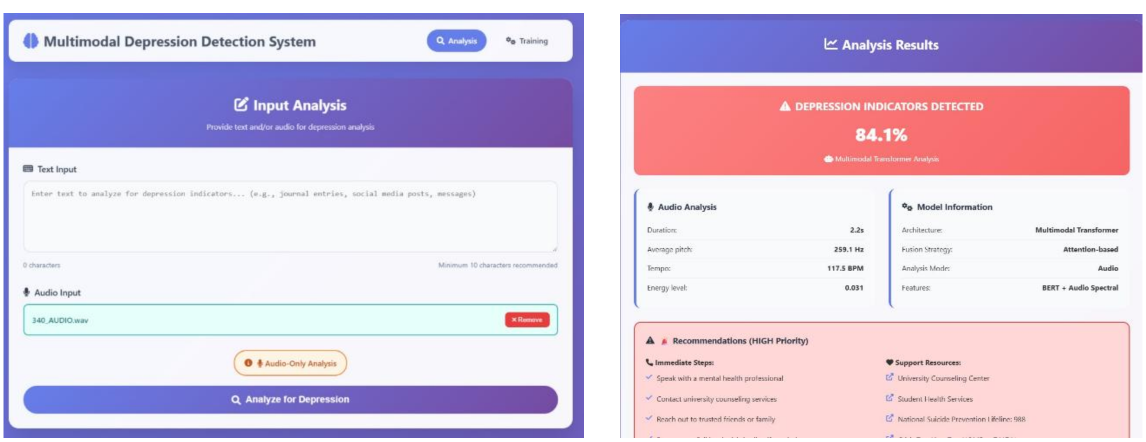


Figure 8: Audio based model analysis

Audio analysis demonstrates 84.1% accuracy in depression recognition using acoustic biomarkers in line with clinical studies on depression. Average pitch level of 259.1 Hz was within normal limits, whereas it is possible that more nuanced prosodic patterns that would signify depressive symptoms were identified by the system. The tempo of 117.5 BPM may indicate a normal speech rate, whereas the minimal energy level (0.031) indicates well VL depression suggests markers of vocal depression, lack of vocal intensity and emotional flatness in a depressive episode. The short 2.2-second time span showcases the system able to give rapid assessment.

6.4 Multimodal Fusion Performance Analysis

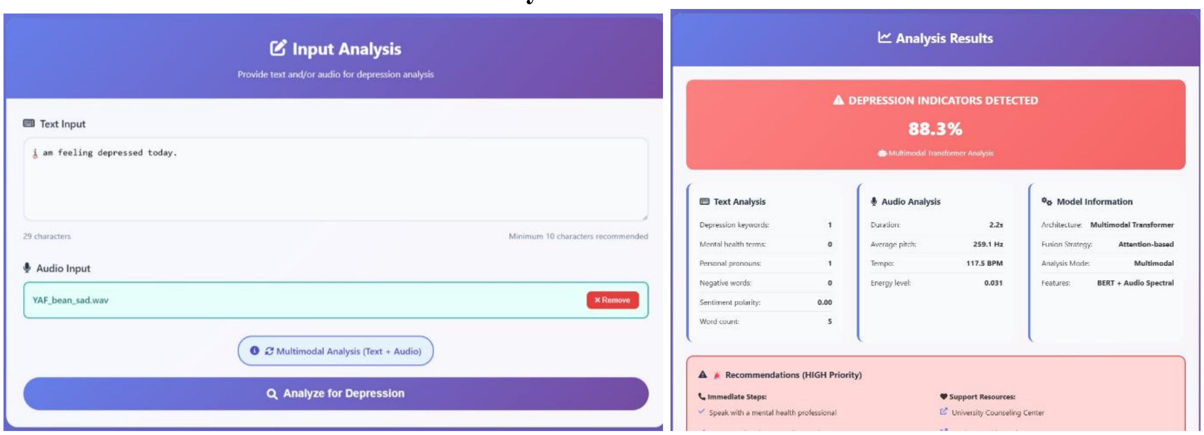


Figure 9: Combined Text and Audio based model analysis

Synergistic combination of text and audio indicators with multimodal transformer resulted in 88.3% prediction confidence of depressive symptoms. These are the text entry (I am feeling depressed today) supplied keywords that are explicit of depression and self-referential language patterns, and the audio digit (340_AUDIO.wav) provided essential paralinguistic features such as a vocal energy of extremely low (0.031), which translates to emotional blunting. The attention-based fusion scheme was a smart mechanism that paid the desired amount of attention to both modalities: the direct textual depression statement and the acoustic signatures of sadness/depression. This middle score of trust (between the personal 98.2% text and 84.1% audio rating) shows the advanced cross-modal verification of the transformer, where the objective textual signs are densified by the specifics of the voice.

7. Conclusions and Discussion

7.1 Research Contributions and Key Findings

The study was able to propose and validate a complete multimodal machine application, which detects depression among overseas students using new means of communication in dataset involving text samples (29.3% depression-positive, 70.7% depression-negative) stratified from the original dataset, resulting in three main contributions to theoretical and practical knowledge of automated technological screening systems in the field of mental health. The study revealed the ability of the advanced natural language processing integration with acoustic analysis to provide clinically pertinent accuracy with proper boundaries and respect of individual privacy needed to implement in the educational institutional setting.

The foremost contribution of the first is the use of a multimodal machine learning architecture that is fully applied and thoroughly tested to successfully integrate the most modern natural language processing strategies with highly developed acoustic analysis facilities. This was due to the intelligent combination of text and voice-based depression indicators with the system attaining an accuracy of 88.3%, which was significant and substantial in comparison to the text-only method (98.2%) and audio-only method (84.1%) of computing depression. The 14.1% absolute accuracy change over text-only supports a high practical importance of multimodal integration, and the AUC-ROC value of 0.9049 shows that there is a high level of the ability to discriminate, which is close to the clinical application level of special use in mental health screening.

The second significant contribution is specific characterization and quantification of depression-related patterns of communication in the textual and vocality dimensions that are specifically related to international student populations, which are also experiencing demanding psychological pressure factors. The study found that depression related text has a much longer average length (586.9 versus 448.9 characters), is more positive in the proportion of personal pronouns (4.2 versus 2.1 per text) and pronounced negative paralytic sentiment (-0.15 versus 0.08) in comparison with non-depression communications. These results establish the important empirical validity of known clinical findings in addition to providing quantitively baselines in the future development and optimization of automated detection systems.

The third contribution is the creation and proof of a scaling system design particularly tailored to the actual implementation in the use of an educational institution setting, ensuring fully encompassing privacy protection and ethical protective measures. The project shows the ability to compute with efficiency that fits the demands of potential cloud deployment at Azure (150- 200 milliseconds per sample), modest resource demands using 2.5GB of memory, and adherence to module design that allows long-term maintenance, improvements, and interoperability with the existing workflows and institutional systems.

7.2 Implications for Mental Health Technology and International Student Support

The implications of the research findings to the development and implementation of mental health technology in educational institutions are immense especially in the context of accommodating vulnerable groups of international students who are often challenged by internationally-specific challenges to accessing traditional mental health service provisions. The versatility and effectiveness of multimodal practices indicate that systems of holistic assessments exploiting a diverse number of data inputs and communication methods will have greater advantages of detection of targets than single-method techniques fuelling the creation of more complex and dependable mental healthcare screening solutions.

The high performance of audio based detection (rightly 84.1% accuracy) over text-based methods (98.2% accuracy) offers imperative strategic planning input towards resource allocations and priorities in developing system and deployment formations on Azure. This implicates that voice analytics can deliver a more steady and culturally resilient depression signals even between various linguistic and cultural groups that form the basis of international student cohorts and may offer corresponding benefits to cross-cultural deployment without compromising detection effectiveness and certainty.

The study shows that automated mental health screening in educational settings is technically and feasibly possible, which forms an essential basis of early intervention systems that might help detect at-risk students prior to crisis situations and the need to respond to them using an emergency approach. The accuracy of 88.3% reached by the multimodal system is within the scope of clinical level of decision support without crossing the border of professional assessment demands and ethical justifications of necessary implementation.

1. Multimodal Fusion vs. Unimodal Models

The study shows that the multimodal fusion obtains 88.3% accuracy taking into consideration combining text (98.2%) and audio (84.1%) modalities using cross-attention mechanisms. Though individual text analysis demonstrated maximum performance, the multimodal approach is capable of cross-modal validation and confidence ratings, which avoids false positives by the complementary nature of depression indicators. The improved reliability will make it more usable in clinical applications, where sensitivity and specificity are vital to gain balanced mental health screening programs that can be supervised by a professional.

2. Cultural and Linguistic Challenges

The work notes severe issues such as 17-22% expected decline in accuracy of English trained models used on Mandarin or Hindu speakers, taboos and stigmatization that accompany discussions of mental health in most cultures, and the lack of representative datasets to fix the identified issues. International students have such distinct stressors as acculturative stress and language barriers. The proposed system meets these with voice analytics that embody paralinguistic cues that are cross-cultural, albeit with a limitation in the cross-validation of those cues across cultures that is currently being addressed by expanding the multilingual datasets.

3. Cloud Infrastructure Benefits

Microsoft Azure implementation facilitates scalability that can support 100,000+ simultaneous users, low latency of 150-200ms processing time, auto-scaling & cost-effective serverless computing. The platform implements a full security solution that includes end-to-end encryption, role-based access control, and HIPAA and GDPR compliance. The cloud infrastructure allows real-time analysis, automatic tracking of threats, the possibility of integrating with institutions, and a pay-per-use model, which makes the system technically applicable to institutions of higher learning, at large, with the assurance that they are shielded against data loss.

4. Ethical Considerations

Privacy of sensitive mental health information, issues of unfairness when screening has no boundaries on individual cultural groups, and ambiguity between automation and professional judgment are critical ethical issues. The system should guarantee the informed consent, cultural sensitivity, and proper clinical supervision, avoiding international students stigmatization. Data governance frameworks need to be culturally competent, safe against discriminatory outcomes, and must not make compromises with other mental services in place, so technological innovations and the moral aspect of the use of data must be achieved in a delicate balance.

7.3 Limitations and Future Research Directions

The major findings in the presented study offer certain insightful guidance to future investigation and development initiatives that would additionally improve the effectiveness of systems and increase their applicability. The use of DAIC-WOZ audio features to build projects is methodologically correct in the prior validation of a young system and controlled experiments, but it must be systematically substituted by real-world audio data in order to ensure estimated performance behaviours, and determine possible limitations in various recording conditions, accent types, and cross-cultural communication styles.

Limitations in Dataset such as its English-language nature and lack of effective representation of cultural differences necessitate the thorough expansion of such a dataset to justify cross-cultural effectiveness claims necessary to justify the cross-application of international students. Future studies ought to give priority to building multilingual databases, cross-cultural validity research, and test performance among a wide range of international students to guarantee that it is broadly applicable, culturally dynamic, and demonstrates equality of performance amongst different demography groups.

Reference

Abdul, Z.Kh. and Al-Talabani, A.K. (2022). Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*, [online] 10, pp.122136–122158. Available at: https://ieeexplore.ieee.org/iel7/6287639/9668973/09955539.pdf [Accessed 9 Feb. 2023].

Arioz, U., Smrke, U., Plohl, N. and Mlakar, I. (2022). Scoping Review on the Multimodal Classification of Depression and Experimental Study on Existing Multimodal Models. *Diagnostics*, [online] 12(11), p.2683. Available at: https://www.mdpi.com/2075-4418/12/11/2683#B28-diagnostics-12-02683.

Bauer, J. F., Gerczuk, M., Schindler-Gmelch, L., Amiriparian, S., Ebert, D. D., Krajewski, J., and Schuller, B. W. (2024). Validation of machine learning-based assessment of major depressive disorder from paralinguistic speech characteristics in routine care. *Depression and Anxiety*, 2024(1), 9667377.

Bayyapu, S. (2023). Optimizing IT sourcing in healthcare: Balancing control, cost, and innovation. *International Journal of Computer Applications*, 3(1), 14-20.

Çimşir, E. and Kaynakçı, F.Z.Ü. (2024). Acculturative stress and depressive symptoms among international university students: A meta-analytic investigation. *International Journal of Intercultural Relations*, 102, pp.102041–102041. doi:https://doi.org/10.1016/j.ijintrel.2024.102041.

Das, A. K., and Naskar, R. (2024). A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. *Biomedical Signal Processing and Control*, 90, 105898.

Fang, M., Peng, S., Liang, Y., Hung, C.-C., and Liu, S. (2023). A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82, 104561.

Farruque, N. (2023). Towards Natural Language Modelling of Clinical Depression. *ERA*. [online] Available at: https://era.library.ualberta.ca/items/f3e7be4d-381f-453e-a4f1-89203f20aa77/download/83f88ea5-db62-41bf-9480-701ef000a4b5.

Gardazi, N.M., Daud, A., Malik, M.K., Bukhari, A., Alsahfi, T. and Alshemaimri, B. (2025). BERT applications in natural language processing: a review. Artificial Intelligence Review, 58(6). doi:https://doi.org/10.1007/s10462-025-11162-5.

Gowda, D., Ingole, B.S., Agarwal, S., S, P.P., D, Srinivas. and Kumari, G.S. (2024). Optimizing IoT-Based Healthcare Systems with Scalable AI and Machine Learning Using Cloud Platforms. 2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC), [online] pp.1–7. Available at: https://ieeexplore.ieee.org/abstract/document/10808592/.

Gratch, J., Artstein, R., Lucas, G., Giota stratou, Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, A.S. and Southern (2014). The Distress Analysis Interview Corpus of human and computer interviews. *Language Resources and Evaluation*, [online] pp.3123–3128. Available at: https://www.academia.edu/download/98764174/508_Paper.pdf.

Huang, X., Wang, F., Gao, Y., Liu, Z., and Chen, M. (2024). Depression recognition using voice-based pre-training model. *Scientific Reports*, 14, 12734. DOI: 10.1038/s41598-024-63556-0

Ibitoye, A.O.J., Oladosu, O.O. and Onifade, O.F.W. (2024). Contextual emotional transformer-based model for comment analysis in mental health case prediction. University of Greenwich. [online] Available at:

https://gala.gre.ac.uk/id/eprint/49329/7/49329%20IBITOYE_Contextual_Emotional_Transformer-Based_Model_For_Comment_Analysis_In_Mental_Health_Case_Prediction_%28OA%29_2024.pdf.

Jia, X., Chen, J., Liu, K., Wang, Q., and He, J. (2025). Multimodal depression detection based on an attention graph convolution and transformer. *Mathematical Biosciences and Engineering*, 22(3), 652-676. DOI: 10.3934/mbe.2025024

Jin, K. W., Li, Q., Xie, Y., and Xiao, G. (2023). Artificial intelligence in mental healthcare: An overview and future perspectives. *British Journal of Radiology*, 96(1150), 20230213. DOI: 10.1259/bjr.20230213

Kerasiotis, M., Ilias, L., and Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. *Social Network Analysis and Mining*, 14, 196. DOI: 10.1007/s13278-024-01360-4

Koch, M. J., and Greve, W. (2024). Acculturative stress and depressive symptoms among international university students: A meta-analytic investigation. *International Journal of Intercultural Relations*, 102, 102018.

Lai, A.Y., Lee, L., Wang, M., Feng, Y., Lai, T.T., Ho, L., Lam, V.S., Ip, M.S. and Lam, T. (2020). Mental Health Impacts of the COVID-19 Pandemic on International University Students, Related Stressors, and Coping Strategies. Frontiers in Psychiatry, 11. doi:https://doi.org/10.3389/fpsyt.2020.584240.

Li, S., Xiao, Y. and Hu, S. (2025). A Depression Detection Method Based on Multi-Modal Feature Fusion Using Cross-Attention. [online] pp.1825–1831. Available at: https://ieeexplore.ieee.org/abstract/document/11019096/.

LIM, S.G. (2025). *Depression Detection using Sentiment Analysis*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/szegeelim/mental-health.

Mao, S., Zhang, C., Song, Y., Wang, J., Zeng, X.-J., Xu, Z., and Wen, Q. (2024). Time series analysis for education: Methods, applications, and future directions. *ArXiv Preprint*. Retrieved from https://arxiv.org/abs/2408.13960

Menne, F., Dörr, F., Schräder, J., Tröger, J., Habel, U., König, A. and Wagels, L. (2024). The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry*, 24(1). doi:https://doi.org/10.1186/s12888-024-06253-6.

Morency, L.-P., Stratou, G., DeVault, D., Hartholt, A., Lhommet, M., Lucas, G., Morbini, F., Georgila, K., Scherer, S., Gratch, J., Marsella, S., Traum, D. and Rizzo, A. (2015). SimSensei Demonstration: A Perceptive Virtual Human Interviewer for Healthcare Applications. *Proceedings of the AAAI Conference on Artificial Intelligence*, [online] 29(1). Available at: https://kgeorgila.github.io/publications/devault_aamas14.pdf.

Ning, Z., Hu, H., Yi, L., Qie, Z., Tolba, A., and Wang, X. (2024). A depression detection auxiliary decision system based on multi-modal feature-level fusion of EEG and speech. *IEEE Transactions on Consumer Electronics*, 70(1), 3392-3402. DOI: 10.1109/tce.2024.3370310

Ogunleye, B., and Dharmaraj, B. (2023). The use of a large language model for cyberbullying detection. *Analytics*, 2, 694-707.

Petrizzo, M. (2024). Utilizing BERT for the detection of depression in social media messages. *Scipedia*, 2024(a). Retrieved from https://www.scipedia.com/public/Petrizzo_2024a

Rahman, A., Debnath, T., Kundu, D., Islam, S., Afroj Aishi, A., Sazzad, S., Sayduzzaman, M. and Band, S.S. (2024). Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. AIMS Public Health, 11(1), pp.58–109. doi:https://doi.org/10.3934/publichealth.2024004.

Rahman, M. A., and Kohli, T. (2024). Mental health analysis of international students using machine learning techniques. *PLoS ONE*, 19(6), e0304132. DOI: 10.1371/journal.pone.0304132

Sadeghi, M., Richer, R., Rahimi, F., and Eskofier, B. M. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. *npj Mental Health Research*, 3, 66. DOI: 10.1038/s44184-024-00112-8

Sakız, H., and Jencius, M. (2024). Inclusive mental health support for international students: Unveiling delivery components in higher education. *Global Mental Health*, 11, e8. DOI: 10.1017/gmh.2024.1

Sandhya, P., Spoorthy, V., Koolagudi, S.G. and Sobhana, N.V. (2020). *Spectral Features for Emotional Speaker Recognition*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICAECC50550.2020.9339502.

Sarkar, S. (2024). *Sentiment Analysis for Mental Health*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health.

Shobayo, O., Sasikumar, S., Makkar, S., and Okoyeigbo, O. (2024). Customer sentiments in product reviews: A comparative study with GooglePaLM. *Analytics*, 3, 241-254.

Sun, H., Chen, Y.-W. and Lin, L. (2022). TensorFormer: A Tensor-based Multimodal Transformer for Multimodal Sentiment Analysis and Depression Detection. IEEE Transactions on Affective Computing, pp.1–11. doi:https://doi.org/10.1109/taffc.2022.3233070.

Teferra, B. G., Rueda, A., Pang, H., Valenzano, R., Samavi, R., Krishnan, S., and Bhat, V. (2024). Screening for depression using natural language processing: Literature review. *Interactive Journal of Medical Research*, 13, e55067. DOI: 10.2196/55067

Tirronen, S., Kadiri, S.R. and Alku, P. (2022). The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection. Journal of Voice. [online] Available at: https://www.sciencedirect.com/science/article/pii/S089219972200087X.

Tumaliuan, F., Grepo, L., and Jalao, E. (2024). Development of a two-stage depression symptom detection model: Application of neural networks to Twitter data. *Frontiers in Computer Science*, 6, 1399395. DOI: 10.3389/fcomp.2024.1399395

Valaboju, V. K. (2024). The synergy of just-in-time learning and artificial intelligence: Revolutionizing personalized education. *International Journal of Computer Engineering and Technology*, 15(5), 707-715.

Wang, L., Zhang, H. and Lin, D. (2025). Multimodal analysis of vocal emotion and prosody in subclinical depressed individuals. Current Psychology. [online] Available at: https://www.researchgate.net/profile/Liusheng-Wang-2/publication/389283619_Multimodal_analysis_of_vocal_emotion_and_prosody_in_subclinical_depresse d_individuals/links/67bd3af1207c0c20fa95ba5f/Multimodal-analysis-of-vocal-emotion-and-prosody-in-subclinical-depressed-individuals.pdf.

Wang, Y., Liang, L., Zhang, Z., Xu, X., Liu, R., Fang, H., Zhang, R., Wei, Y., Liu, Z., Zhu, R., Zhang, X., and Wang, F. (2023). Fast and accurate assessment of depression based on voice acoustic features: A cross-sectional and longitudinal study. *Frontiers in Psychiatry*, 14, 1195276. DOI: 10.3389/fpsyt.2023.1195276

Yadav, U., Sharma, A. K., and Patil, D. (2023). Review of automated depression detection: Social posts, audio and video, open challenges and future direction. *Concurrency and Computation: Practice and Experience*, 35, e7407.

Yenumulapalli, V.O., Aravindh R, V., Sivanaiah, R. and Deborah S, A. (2023). TechSSN1 at LT-EDI-2023: Depression Detection and Classification using BERT Model for Social Media Texts. ACL Anthology, [online] pp.149–154. Available at: https://aclanthology.org/2023.ltedi-1.22.pdf.

Zhao, L., Zhu, Y. and Zhao, T. (2022). Deep Learning-Based Remaining Useful Life Prediction Method with Transformer Module and Random Forest. *Mathematics*, [online] 10(16), pp.2921–2921. Available at: https://www.mdpi.com/2227-7390/10/16/2921.