Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056666

ARTICLE

# MDD: A Unified Multimodal Deep Learning Approach for Depression Diagnosis Based on Text and Audio Speech

**Farah Mohammad**[1,2,*] **and Khulood Mohammed Al Mansoor**[3]

[1]Center of Excellence and Information Assurance (CoEIA), King Saud University, Riyadh, 11543, Saudi Arabia

[2]Department of Computer Science, and Technology, Arab East Colleges, Riyadh, 11583, Saudi Arabia

[3]Self-Development Skills Department, King Saud University, Riyadh, 11543, Saudi Arabia

*Corresponding Author: Farah Mohammad. Email: fnazar@ieee.org

Received: 27 July 2024 Accepted: 30 October 2024 Published: 19 December 2024

## ABSTRACT

Depression is a prevalent mental health issue affecting individuals of all age groups globally. Similar to other mental health disorders, diagnosing depression presents significant challenges for medical practitioners and clinical experts, primarily due to societal stigma and a lack of awareness and acceptance. Although medical interventions such as therapies, medications, and brain stimulation therapy provide hope for treatment, there is still a gap in the efficient detection of depression. Traditional methods, like in-person therapies, are both time-consuming and labor-intensive, emphasizing the necessity for technological assistance, especially through Artificial Intelligence. Alternative to this, in most cases it has been diagnosed through questionnaire-based mental status assessments. However, this method often produces inconsistent and inaccurate results. Additionally, there is currently a lack of a comprehensive diagnostic framework that could be effective achieving accurate and robust diagnostic outcomes. For a considerable time, researchers have sought methods to identify symptoms of depression through individuals' speech and responses, leveraging automation systems and computer technology. This research proposed MDD which composed of multimodal data collection, preprocessing, and feature extraction (utilizing the T5 model for text features and the WaveNet model for speech features). Canonical Correlation Analysis (CCA) is then used to create correlated projections of text and audio features, followed by feature fusion through concatenation. Finally, depression detection is performed using a neural network with a sigmoid output layer. The proposed model achieved remarkable performance, on the Distress Analysis Interview Corpus-Wizard (DAIC-WOZ) dataset, it attained an accuracy of 92.75%, precision of 92.05%, and recall of 92.22%. For the E-DAIC dataset, it achieved an accuracy of 91.74%, precision of 90.35%, and recall of 90.95%. Whereas, on CD-III dataset (Custom Dataset for Depression), the model demonstrated an accuracy of 93.05%, precision of 92.12%, and recall of 92.85%. These results underscore the model's robust capability in accurately diagnosing depressive disorder, demonstrating the efficacy of advanced feature extraction methods and improved classification algorithm.

## KEYWORDS

Depression; deep learning; T5; WaveNet; CCA; neural network

## 1 Introduction

According to the World Health Organization (WHO) [1], depression is a common and severe mental health condition affecting more than 280 million people worldwide. An estimated 3.8% of the global population is affected by depression, including 5.7% of individuals over the age of 60% and 5.0% of adults [2]. Depression is a significant contributor to suicide, with those suffering from long-term mental illness being more prone to suicidal tendencies. Globally, suicide ranks as the fourth leading cause of death among individuals aged 15–29 [3]. This issue is prevalent not only in developing countries but also in developed nations, with 77% of suicides in 2019 occurring in developing countries [4]. Depression is a major factor driving individuals toward suicide, with an estimated 75% of those suffering from depression in developing countries remaining untreated [5].

Medically, symptoms of depression include persistent sadness, hopelessness, loss of interest in activities, irritability, difficulty concentrating, negative thought patterns, fatigue, changes in sleep and appetite, unexplained physical pain and slowed movements [6]. However, depression often manifests through distinctive patterns in both speech and text data. In speech data, individuals with depression typically exhibit a monotone pitch, characterized by a lack of variation in their voice, which reflects a flattened affect and reduced emotional expression [7]. They also tend to speak at a slower rate, with more deliberate and measured speech, indicating decreased cognitive and physical energy. Increased pauses and hesitations are common, reflecting difficulties in cognitive processing or a lack of motivation. Additionally, their speech might be quieter, demonstrating reduced energy levels, and less clear, with more mumbled or less distinct articulation. Increased disfluencies, such as the use of filler words, repetitions, and corrections, are also observed, indicating impaired cognitive function.

In text data, the symptoms of depression can be identified through various linguistic features. Depressed individuals often use more negative language, with frequent expressions of sadness, hopelessness, and worthlessness [8]. There is a higher occurrence of self-referential words such as "I" and "me," reflecting self-focused attention and rumination. Texts from depressed individuals may also contain more absolutist terms like "always," "never," and "completely," indicating black-and-white thinking patterns [9]. Additionally, their writing might exhibit reduced complexity, with shorter sentences and simpler vocabulary, suggesting cognitive fatigue and difficulty in concentrating. Increased repetition of themes related to loss, failure, and negative self-evaluation are also common, providing further insight into the depressive thought patterns present in the individual's written expressions.

Traditional methods for diagnosing depression primarily rely on clinical interviews, self-report questionnaires, and standardized diagnostic criteria such as the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) or ICD-10 (International Classification of Diseases, Tenth Edition) [10,11]. Clinical interviews involve direct interaction between a healthcare professional and the patient, where the clinician assesses symptoms based on observation and patient responses. Common self-report questionnaires include the Beck Depression Inventory (BDI), the Patient Health Questionnaire-9 (PHQ-9), and the Hamilton Depression Rating Scale (HDRS). These tools ask patients to report their symptoms, frequency, and severity, providing a subjective measure of their mental health status [12].

Despite their widespread use, these traditional methods have several limitations. Firstly, they are inherently subjective, relying heavily on the patient's ability to accurately self-report symptoms and the clinician's interpretation of these reports. This can lead to variability in diagnosis due to differences in clinician expertise and patient honesty or self-awareness. Secondly, these methods can be time-consuming and require significant clinician-patient interaction, which may not be feasible in settings

with limited mental health resources. Thirdly, traditional diagnostic tools often fail to capture the nuanced, day-to-day fluctuations in a patient's mood and behavior, providing only a snapshot of their mental state at a single point in time. This can result in underdiagnoses and misdiagnosis, particularly in cases where patients present atypical symptoms or have co-occurring mental health conditions. Additionally, the stigma associated with mental health can discourage individuals from seeking help or being truthful during assessments, further complicating the diagnostic process.

Computer-based solutions for depression detection leverage artificial intelligence (AI) and machine learning (ML) to analyze text, speech and physiological data [13,14]. Natural Language Processing (NLP) techniques analyze text from social media posts, emails, and clinical transcriptions, identifying linguistic patterns such as negative sentiment, self-referential language, and absolutist terms indicative of depression. Speech analysis examines audio recordings for changes in pitch, tone, speech rate, and vocal clarity, which are commonly altered in depressive states [15]. Physiological data, like heart rate variability and activity levels, are monitored using wearable devices to detect signs of depression. Despite their potential, these methods face limitations, including data quality issues, such as background noise in speech analysis and context understanding in NLP [16]. Privacy concerns also arise from analyzing personal communications and physiological data. Additionally, these models may struggle with generalizability across diverse populations and require large, representative datasets to ensure accuracy and reliability.

Keeping in view the above limitations this research offers enhanced accuracy by combining text and speech data, capturing a comprehensive range of linguistic and vocal indicators. This integration allows for a more robust analysis of depressive symptoms, overcoming the limitations of single-modality methods. Additionally, it can provide early detection and personalized insights, improving the effectiveness of interventions. By leveraging multiple data sources, this approach enhances the reliability and generalizability of depression diagnosis across diverse populations. The key steps of the proposed model are: Data collection involves gathering text data from social media posts, self-reports, and transcriptions, along with speech data from interviews and therapy sessions. Preprocessing steps include tokenization, stop word removal, and normalization for text, and noise reduction, segmentation, and feature extraction for speech. Feature extraction uses techniques like word embedding and sentiment analysis for text, and Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and speech rate analysis for speech. Modeling employs NLP with Long Short-Term Memory (LSTM)/Recurrent Neural Network (RNN) and Transformers for text, and Convolutional Neural Network (CNN) and RNN for speech. Finally, feature fusion combines these data sources to create a comprehensive multimodal deep learning model for depression diagnosis.

The eye catching contribution of this work are as follows:

- Integrates multimodal data from text and speech to enhance the accuracy and robustness of depression diagnosis.
- Utilizes advanced feature extraction techniques, such as T5, WaveNet, CCA, to capture nuanced linguistic and acoustic indicators of depression.
- Develops a comprehensive deep learning model that combines NLP and speech analysis, providing a more holistic and reliable approach to mental health assessment.
- The proposed model achieved an accuracy of 92.51%, precision of 91.5%, and recall of 92.03%, demonstrating its robust effectiveness in diagnosing depressive disorder.

The rest of the paper is organized as follows: Section 2 provides a detailed overview of the existing literature, Section 3 discusses the core methodology of MDD, experimental evaluations and results are presented in Section 4, Section 5 gives the MDD recommendations, and Section 6 presents the conclusion and future research directions.

## 2 Literature Review

In recent years, the field of mental health assessment particularly depression diagnosis has seen significant advancements with the integration of AI and ML techniques. This section provides a comprehensive overview of some of the benchmark method that has been developed for depression detection. Whereas, Table 1 provides additional comparative analysis of few other methods in order to save time and space. Anik et al. [17] highlighted the absence of a thorough diagnostic methodology for Major Depressive Disorder that evaluates various brainwave types (alpha, theta, gamma, etc.) using electroencephalogram (EEG) signals to identify the most effective biomarkers for accurate and robust diagnostics. To address this gap, they introduced a novel technique utilizing a deep convolutional neural network (DCNN) for diagnosing Major Depressive Disorder, leveraging EEG brainwave data. Their innovative model, an extended 11-layer one-dimensional convolutional neural network (Ex-1DCNN), is designed to learn automatically from EEG signals without requiring manual feature extraction. By capitalizing on intrinsic brainwave patterns, the model effectively categorizes EEG signals into depressive and healthy groups. Comprehensive analysis revealed that the Gamma brainwave, with a 15-s epoch duration, was the most effective configuration, achieving an impressive accuracy of 99.60%, sensitivity of 100%, specificity of 99.21%, and an F1-score of 99.60% using EEG data from 34 MDD patients and 30 healthy individuals. This research emphasizes the potential of deep learning methods in enhancing the diagnostic process for MDD and offers a reliable tool for clinicians in diagnosing the disorder.

The work of Rehmani et al. [18] presented that depression is a serious mental state that negatively impacts thoughts, feelings, and actions, and with the rapid growth of social media, individuals increasingly express themselves in their regional languages. Recognizing the prevalence of Roman Urdu on social media in Pakistan and India, the authors propose leveraging this language for depression prediction, addressing a gap in prior research which has largely overlooked Roman Urdu or its combination with structured languages like English. The study aims to create a dual-language dataset comprising Roman Urdu and English to predict depression risk. The authors utilized two datasets: Roman Urdu data manually converted from English posts on Facebook and English comments from Kaggle, merging these for their experiments. They tested various ML models, including Support Vector Machine (SVM), Support Vector Machine Radial Basis Function (SVM-RBF), Random Forest (RF), and Bidirectional Encoder Representations from Transformers (BERT), classifying depression risk into not depressed, moderate, and severe categories. Their experimental results indicate that SVM achieved the best performance with an accuracy of 84%, surpassing existing models. This study refines the area of depression prediction, particularly in Asian countries, by effectively utilizing dual-language datasets.

Manjulatha et al. [19] discussed that stress, followed by depression, has become a prevalent issue in the modern work environment, necessitating early detection to prevent health deterioration and reduce suicide risk. The authors propose a multimodal depression classification system based on deep learning to enhance the accuracy of noninvasive monitoring methods. Traditional methods relying on visual cues, audio feeds, and text messages have shown limitations, with individual modalities often resulting in low accuracy and high false positive rates. To overcome these challenges, the proposed solution integrates visual, speech, and text feeds, extracting deep learning features from each modality. These features are subsequently classified into emotions and temporal emotion variability to determine the depression level. This innovative approach aims to provide a more accurate and reliable method for early depression detection in the workplace.

Katiyar et al. [20] expressed that anxiety, depression, and stress are increasingly serious problems, particularly among women, who are often more susceptible due to socio-economic responsibilities,

thus affecting broader societal well-being. Beyond general mental health issues, postpartum depression (PPD) presents a significant health problem impacting mothers after childbirth. Currently, there are no predictive tools to screen for depression; however, ML has emerged as a promising approach in detecting these mental health conditions. ML employs dynamic statistical and probabilistic methods to predict issues like depression and anxiety by analyzing datasets derived from questionnaires. These tools can predict symptoms and assist in diagnosing mental health issues, ultimately reducing self-harm. This chapter aims to compare leading algorithm models for identifying depression and anxiety. The authors propose a deep recurrent neural network (DRNN) algorithm, which has demonstrated high accuracy and precision, suggesting a potential for future research. The study highlights various ML algorithms such as gradient boosting (GB), random forest (RF), artificial neural network (ANN), SVM, logistic regression (LR), decision tree (DT), and DRNN, all of which aid in predicting these mental health issues. Positioning these models effectively can facilitate a more robust clinical approach to mental health diagnosis and treatment.

While significant advancements have been made in using ML to predict and diagnose depression, anxiety, and stress, several limitations remain. Despite the potential of algorithms such as DRNN, GB, RF, and others, the accuracy and reliability of these models can be hindered by the quality and diversity of the data sets used. Many models rely on self-reported questionnaires, which can be subjective and vary greatly between individuals and cultures. Furthermore, the integration of multimodal data (e.g., text, speech, and visual cues) presents technical challenges and may not always lead to consistent improvements in diagnostic accuracy. Privacy concerns and the ethical use of sensitive personal data are also critical issues that need to be addressed. As a result, while ML offers promising tools for early detection and intervention of mental health issues, further research is needed to refine these models, improve their generalizability, and ensure their ethical application in clinical settings.

**Table 1:** Comparative analysis of existing benchmark models

| Ref.#. | Methodology | Accuracy | Limitations |
|---|---|---|---|
| [21] | CNN-BLSTM with TL-based model combining transfer learning, BLSTM, and CNN to analyze EEG signals for PPD prediction. | 89.6% | • Limited to PPD, requires extensive EEG data.<br>• May not generalize to other populations or conditions. |
| [22] | LSTM-based DL model using emotional features, topical events, and behavioral-biometric signals to categorize tweets related to depression. | Highest R2: 0.61 | • Dependent on the quality of social media data.<br>• Privacy concerns.<br>• Potential bias in dataset labeling. |
| [23] | Multi-channel CNN (MCNN) with attention layers to capture local and global features from social media posts for depression detection. | 91.00% | • Focuses only on text data.<br>• May not account for context and nuance in language.<br>• Limited by dataset size and diversity. |

(Continued)

**Table 1 (continued)**

| Ref.#. | Methodology | Accuracy | Limitations |
|---|---|---|---|
| [24] | TAM-SenticNet, a Neuro-Symbolic AI framework merging neural networks and symbolic reasoning for early depression detection through social media content analysis. | F1-score: 0.758 | • Requires complex integration of neural and symbolic reasoning.<br>• Potential scalability issues, data privacy concerns. |
| [25] | Deep-Knowledge-Aware Depression Detection system utilizing domain knowledge and digital traces for depression detection and explanation. | 89% | • Dependent on the quality and relevance of domain knowledge.<br>• Potential issues with feature extraction and data integration. |
| [26] | Hybrid ML models for sentiment analysis of Twitter tweets using various combinations of feature extraction and classification techniques. | Highest: 0.894 | • Focuses only on text data from Twitter.<br>• Potential biases in social media usage patterns. |
| [27] | Multimodal GNN for depression detection, integrating audio, text, and video features using a pre-fusion strategy. (Baseline 1). | 85.96% | • Complexity of integrating multiple modalities; few-shot learning challenges. |
| [28] | Multimodal data image encoding and fusion approach using RGB and sparse coding, with STN and RGA for feature extraction and decision-making. (Baseline 2). | 86.71% | • Challenges in encoding and fusing diverse data types. |
| [29] | Audio-based depression detection method using neural networks to classify features from audio spectrograms with optimized CNN. (Baseline 3). | 91% | • Reliance on audio-only features; challenges in neural network parameter optimization. |
| [30] | Multimodal fusion method based on Deep Spectrum (Baseline 4). | 91.78% | • Features optimization is required. |

## 3 Proposed Methodology

This section discusses the core methodology of the MDD that composed of multimodal data collection, preprocessing, feature extraction (T5 model for text feature extraction and WaveNet model for speech feature extraction), CCA for correlated projection of text and audio features, feature fusion based on concatenation and Depression detection using neural network with sigmoid output layer.

Fig. 1 shows the model architecture of proposed work, whereas the detail of each phase has been described in below sub sections.

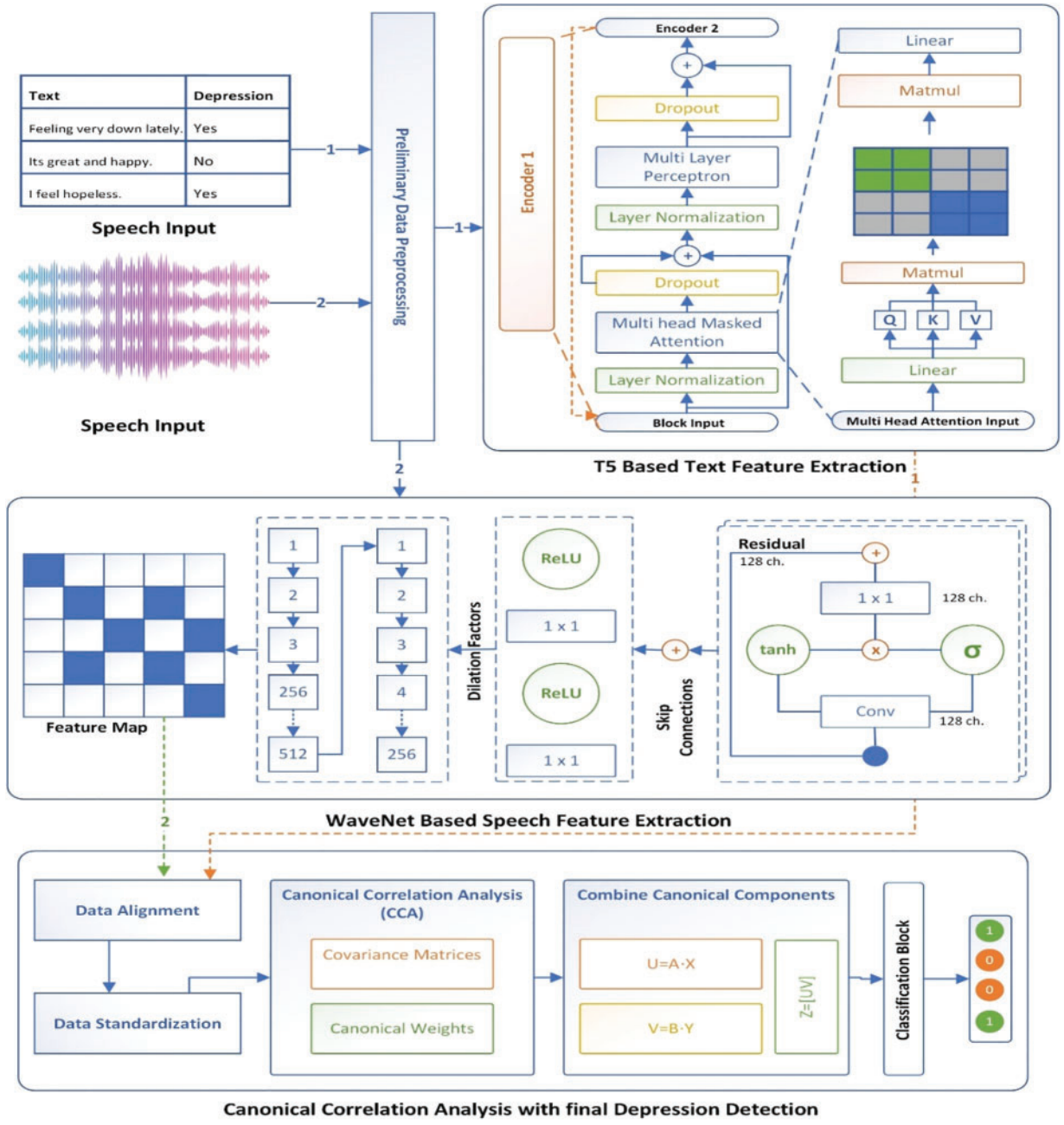



**Figure 1:** Proposed conceptual model of MDD

This figure represents a sophisticated system for detecting depression by analyzing both text and speech inputs through a combination of neural network models and CCA. The process begins with the collection of text and speech inputs, where the text might include phrases with associated depression labels, and the speech is captured as audio waveforms. Both types of data undergo

preliminary preprocessing, which prepares them for feature extraction by normalizing, tokenizing, or otherwise refining the data. For text feature extraction, a Transformer-based model, specifically the T5 architecture, is employed. This model uses multiple layers, including multi-layer perceptrons, layer normalization, and multi-head masked attention mechanisms, to extract relevant features from the text. These features are then transformed into a final text feature vector through linear layers and matrix multiplication operations.

Simultaneously, speech feature extraction is performed using a WaveNet-based model. The speech signal is processed through dilated convolutional layers that capture different levels of temporal dependencies. The model also incorporates activation functions like ReLU and tanh, along with skip and residual connections, to enhance learning and maintain the integrity of the data across deeper layers. Once the features from both text and speech have been extracted, they undergo standardization and alignment to ensure they are compatible for further analysis. CCA is then applied, where covariance matrices and canonical weights are computed. This method identifies linear relationships between the text and speech features, and the canonical components from both modalities are combined to maximize their correlation. Finally, the combined features are fed into a classification block that predicts the presence of depression. This integrated approach, which leverages the strengths of both neural network architectures and statistical analysis, aims to provide a robust and accurate system for detecting depression from multimodal inputs.

### *3.1 Data Collection*

Three different types of datasets have been considered for the evaluation of the proposed work. The distribution of each dataset has been presented in Table 2. The DAIC-WOZ dataset is a part of the larger Distress Analysis Interview Corpus, created by the University of Southern California's Institute for Creative Technologies (USC ICT). It was designed to support the development of automated systems capable of diagnosing psychological distress, including conditions such as depression, anxiety, and post-traumatic stress disorder (PTSD). This dataset includes audio and video recordings, as well as transcripts of interviews conducted by a virtual interviewer controlled by a human operator (the "Wizard of Oz"). The interviews follow a semi-structured format and are designed to elicit responses that can be indicative of psychological distress. The dataset contains detailed annotations, including participant demographics, verbal and non-verbal behaviors, and clinical assessments.

**Table 2:** Data collection

| Dataset name | Nature | Records | Data scale | Corrective measure |
|---|---|---|---|---|
| DAIC-WOZ | Audio, video, text | 1200 interviews with annotations | 50 GB | DAIC-WOZ dataset |
| E-DAIC | Audio, video, text, physiological | 2500 extended multimodal recordings | 40 GB | Available through academic collaboration with USC ICT |
| CD-III (Saudi hospital) | Audio, text, physiological | 1800 interviews, questionnaires, and wearable data | 10 GB | Custom implementation needed |

The E-DAIC is an extension of the DAIC-WOZ dataset, enhancing it with additional modalities and more comprehensive data. This extended version includes not only audio and video recordings and transcripts but also physiological signals like heart rate and skin conductance, collected via wearable sensors. The E-DAIC aims to provide a richer, multimodal dataset for more robust analysis and development of diagnostic tools. By incorporating physiological data, researchers can explore the interplay between verbal, non-verbal, and physiological indicators of distress, potentially leading to more accurate and reliable diagnostic models. The E-DAIC dataset is particularly valuable for studying how different types of data can complement each other in the diagnosis of mental health conditions.

The last dataset abbreviated as CD-III (Custom Dataset for Depression) has been collected in collaboration with a Saudi hospital, offers a highly tailored approach to gathering multimodal data for depression diagnosis. In this process, a comprehensive approach based on expert team has been adopted to collect both speech and text data from patients during mental health evaluations, ensuring cultural relevance and context. This involves recording audio from clinical interviews, where patients discuss their mental health status, and collecting written assessments and questionnaires for text data. Additionally, incorporating wearable technology can provide physiological data such as heart rate and skin conductance.

### *3.2 Data Preprocessing*

Effective data preprocessing as depicted in Algorithm 1, is essential for the accuracy and efficiency of ML models. This process includes several steps to clean and prepare text and speech data for analysis. For text data preprocessing, the first step is tokenization, which splits the text into individual units called tokens, transforming a text document T into a sequence of tokens $\{w_1, w_2, \ldots, w_n\}$. Next, the removal of stop words eliminates common words that do not carry significant meaning, reducing noise in the data. This step filters the tokenized text to produce $w_i = w_i \notin S$, where $S$ is the set of stop words. Normalization then converts the text to a standard format, including lowercase conversion, stemming, and lemmatization, ensuring consistency and reducing variability. If w is a word, normalization can be represented by the function $N$:

$$(w) = stem/lemma(w) \tag{1}$$

Applying normalization to each token in the filtered set:

$$N(R(T(\mathfrak{T})) = N(w_i) = w_i \notin S \tag{2}$$

For speech data preprocessing, noise reduction is the first step, enhancing the clarity of the audio signal by removing background noise. This process is represented as $y(t) = x(t) - n(t)$, where $x(t)$ is the original audio signal and $n(t)$ is the noise estimate. Following noise reduction, speech segmentation divides the continuous speech into meaningful segments, such as words, phrases, or sentences, denoted as $y_1, y_2, y_3, \ldots, y_m(t)$. Segmentation can be represented by the function $S$:

$$S(y(t)) = y_1(t), y_2(t), y_3(t) \ldots y_m(t) \tag{3}$$

**Algorithm 1:** Data preprocessing

**Input**

- Text data $T$
- Audio signal $x(t)$

(Continued)

**Algorithm 1** (continued)

**Output**
- Preprocessed text data $N(R(T(\mathcal{T}))$
- Preprocessed audio segments $\{y_i(t)\}$

**For each** word w in text data T:
**Do**
$\{w_1, w_2, w_3 \ldots w_n\} \longleftarrow \mathcal{T}(T)$
**If** $w_i \notin S$ then include the result set
$R(T(\mathcal{T})) \longleftarrow \{w_i | w_i \notin S\}$
**End if**
**End For**
**For each** token $w_i$ in $R(T(\mathcal{T}))$
**Do**
$N(w_i)$
$N(R(T(\mathcal{T})) \longleftarrow N(w_i) = w_i \notin S$
**End For**
**For each** S
**Do**
*Noice Reduction as* $y(t) \longleftarrow x(t) - n(t)$
*Segmentation* $S(y(t)) \longleftarrow y_1(t), y_2(t), y_3(t) \ldots y_m(t)$
**End For**
**Return**

### *3.3 Feature Extraction*

Feature extraction as shown in Algorithm 2 is essential phase of MDD as it enhances the performance and accuracy of proposed model, and transforms raw data into a format that is more suitable for final depression detection. By identifying and selecting the most relevant features from text and speech data, feature extraction reduces the dimensionality of the input, mitigates noise, and highlights the underlying patterns and structures in the data. This process improves the model's ability to learn effectively and generalize well to new, unseen data, leading to improved predictive performance and robustness. In the context of multimodal domain that combining text and speech data, two different feature extraction strategies have been used for effective feature extraction. This ensures that the canonical components capture the essential relationships between the two modalities, thereby enabling a more integrated and insightful analysis. The detailed of each strategy has been discussed in below subsection.

**Algorithm 2:** T5-based text and WaveNet-based speech features extraction

**Require:** Text data $T = \{t_1, t_2, ..., t_n\}$, Speech data $S = \{s_1, s_2, ..., s_n\}$
**Ensure:** Depression prediction P
1: **Initialize** T5 model $T5_{model}$ and WaveNet model $WaveNet_{model}$
2: **Load** pre-trained weights for $T5_{model}$ and $WaveNet_{model}$
3: **T5 Model Architecture:**

(Continued)

**Algorithm 2 (continued)**

4: **Encoder:**

$$\boldsymbol{H}_0 = Embedding\,(t)$$
$$\boldsymbol{H}_l = LayerNorm\,(SelfAttention\,(\boldsymbol{H}_l - 1) + \boldsymbol{H}_l - 1)$$
$$\boldsymbol{H}_l = LayerNorm\,(FeedForward\,(\boldsymbol{H}_l) + \boldsymbol{H}_l)\ \ for\ l = 1, \ldots, N$$

5: **Decoder:**

$$\boldsymbol{G}_0 = Embedding\,(t')$$
$$\boldsymbol{G}_l = LayerNorm\,(SelfAttention\,(\boldsymbol{G}_{l-1}) + \boldsymbol{G}_{l-1})$$
$$\boldsymbol{G}_l = LayerNorm\,(CrossAttention\,(\boldsymbol{G}_l, \boldsymbol{H}_N) + \boldsymbol{G}_l)$$
$$\boldsymbol{G}_l = LayerNorm\,(FeedForward\,(\boldsymbol{G}_l) + \boldsymbol{G}_l)\ \ for\ l = 1, \ldots, N$$

6: **Feature Extraction from Text**

7: **for** each text sample t $\in$ **T do**

8: $\boldsymbol{Z}_t \leftarrow \text{T5}_{\text{model}}(\text{t})$ // Extract features using T5 model

9: **end for**

10: **WaveNet Model Architecture:**

$$\boldsymbol{X}_0 = s$$
$$\boldsymbol{X}_l = ReLU\,(DilatedConv\,(\boldsymbol{X}_{l-1}))$$
$$\boldsymbol{X}_l = \boldsymbol{X}_l + Residual\,(\boldsymbol{X}_{l-1})\ \ for\ l = 1, \ldots, L$$

11: **Feature Extraction from Speech**

12: **for** each speech sample s $\in$ $S$ **do**

13: $\mathbf{Z}_s \leftarrow \text{WaveNet}_{\text{model}}(\text{s})$ // Extract features using WaveNet model

14: **end for**

15: Combine Features

16: **for** each sample $i$ **do**

17: $\boldsymbol{Z}_i \leftarrow [\boldsymbol{Z}_{t_i}, \boldsymbol{Z}_{s_i}]$ // Concatenate text and speech features

18: **end for**

### *3.3.1 Text Based Feature Extraction*

There exist too many deep learning models for text feature extraction but using T5 (Text-to-Text Transfer Transformer) for depression detection from review data is advantageous because of its versatile and robust text-to-text framework, which allows it to handle a wide range of NLP tasks effectively. T5's pre-trained model, built on the powerful Transformer architecture, provides rich contextual embeddings that capture nuanced patterns in text, essential for detecting subtle indicators of depression. Fine-tuning T5 on specific datasets enables it to adapt to the particular linguistic features and sentiment expressions associated with depression, resulting in highly accurate and meaningful feature representations.

After the data preprocessing, each review has been mentioned as labeled with indicators of depression and may be considered as binary (indicating whether depression is present or not). Each review is a sequence of words $\{w_1, w_2, w_3, \ldots, w_n\}$. The normalization process of the preprocessing helps in converting text to lowercase and remove punctuation. Whereas, the tokenization splits the text into tokens, which are then mapped to unique identifiers using a T5-compatible tokenizer. Let $X = \{x_1, x_2, x_3, \ldots, x_n\}$ denote the tokenized sequence for a review. T5 uses the Transformer architecture, consisting of an encoder and a decoder. The encoder processes the input text to generate contextualized embeddings, while the decoder generates output text sequences based on the encoded input. For feature extraction, the focus is on the encoder part.

The tokenized review $X$ is input to the T5 encoder to obtain contextualized embeddings. The encoder's output for each token $x_i$ in the sequence is a hidden state vector $H_i$:

$$H_i = Encoder\,(X)\,[i] \tag{4}$$

where $H_i$ represents the hidden state corresponding to the $i$th token.

To create a fixed-size feature vector representing the entire review, the contextualized embeddings need to be aggregated for which average pooling [31] method has been adopted, where the hidden states of all tokens are averaged by using Eq. (5):

$$F = \frac{1}{n}\sum_{i=1}^{n} H_i \tag{5}$$

The T5 model further fine-tuned on the depression detection task by optimizing its parameters on the specific dataset given in Section 3.1. This involves adjusting the weights of the T5 encoder to better capture the patterns related to depression in the text.

### *3.3.2 Speech Data Feature Extraction*

WaveNet is a deep generative model for raw audio waveforms, which is capable of producing high-quality, natural-sounding speech. It operates on raw audio waveforms directly, without the need for intermediate representations such as mel-spectrograms [32]. This ability makes it an ideal choice for feature extraction from speech data. The process of extracting features from speech data using a WaveNet model involves several steps, each contributing to capturing the essential characteristics of the audio signal. Initially, the input to the WaveNet model is a raw audio waveform, denoted as: $X = \{x_1, x_2, x_3, \ldots, x_T\}$ where, $T$ is the length of the audio signal. The goal is to model the conditional distribution of the waveform sample $x_t$ given all previous samples.

WaveNet employs dilated causal convolutions to process the audio signal. Dilated convolutions allow the model to have a large receptive field with relatively few layers by exponentially increasing the dilation factor at each layer. The output y of a dilated convolution with a dilation factor $d$ is given by Eq. (6) as:

$$y\,[t] = \sum_{k=0}^{K-1} w_k.x[t - d.k] \tag{6}$$

where $w_k$ are the filter coefficients and $K$ is the filter size. This approach ensures that the model captures long-range dependencies in the audio signal. To maintain the temporal order of the data, WaveNet uses causal convolutions, ensuring that the output at time $t$ depends only on the inputs at time $t$ and earlier:

$$y\,[t] = \sum_{k=0}^{K-1} w_k.x[t - k] \tag{7}$$

Each convolutional layer in WaveNet uses gated activation units to enhance the model's capacity:

$$z\,[t] = \tanh(W_{f,k} * x\,[t]) \odot \sigma\,(W_{g,k} * x\,[t]) \tag{8}$$

where $W_{f,k}$ and $W_{g,k}$ are the filter and gate weights, respectively, $*$ denotes convolution, tanh is the hyperbolic tangent function, $\sigma$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. To facilitate training and improve the flow of gradients, WaveNet incorporates residual and skip connections. The output at time t is given by:

$$o\,[t] = x\,[t] + \sum_{l=1}^{L} yl[t] \tag{9}$$

where $o[t]$ is the output of the network at time $t$, $yl[t]$ is the output of the $l$th layer, and $L$ is the total number of layers. The extracted features, or embeddings, from the WaveNet model are used for subsequent tasks which is feature integration based on CCA for correlated projection of text and audio features.

### *3.4 Multimodal Feature Fusion*

To integrate and find correlations between features from multiple modalities, such as text and audio, this research utilizes, CCA which is a statistical method used to understand the relationships between two sets of variables. In this context, CCA is applied to find the correlated projections of text and audio features, allowing us to capture the relationships between the modalities. Given two sets of variables $\mathrm{X} \in \mathbb{R}^{n \times p}$ (text features) and $\mathrm{Y} \in \mathbb{R}^{n \times q}$ (speech features), where $n$ is the number of samples, $p$ is the number of text features, and $q$ is the number of audio features, CCA aims to find linear combinations of the variables in $X'$ and $Y'$ that are maximally correlated.

For a given features let $X'a$ *and* $Y'b$ be zero-mean matrices of text and audio features, respectively. CCA seeks to find projection vectors $\mathrm{a} \in \mathbb{R}^{p}$ and $\mathrm{b} \in \mathbb{R}^{q}$ such that the linear combinations $Xa$ and $Yb$ have maximum correlation.

The correlation between $Xa$ and $Yb$ is given by:

$$\rho = \frac{a^T x^T Yb}{\sqrt{a^T X^T Xa}\sqrt{b^T Y^T Yb}} \tag{10}$$

The goal is to maximize $\rho$ subject to the constraints that the variances of the projections are 1:

$$maximize\ a^T x^T Yb \tag{11}$$

$$Subject\ to\ a^T X^T Xa = 1 \tag{12}$$

$$b^T Y^T Yb = 1 \tag{13}$$

The optimization problem can be solved using Lagrange multipliers. The solutions to this problem are given by the eigenvectors corresponding to the largest eigenvalues of the following matrix pair:

$$(\mathrm{X}^T\mathrm{Y}(\mathrm{Y}^\top\mathrm{Y})^{-1}\mathrm{Y}^\top\mathrm{X})\mathrm{a} = \lambda\mathrm{a} \tag{14}$$

$$(\mathrm{Y}^T\mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{Y})\mathrm{a} = \lambda\mathrm{b} \tag{15}$$

The eigenvectors a and b corresponding to the largest eigenvalues $\lambda$ provide the directions of maximum correlation in the respective feature spaces. The projections of the original data onto the canonical directions are:

$$U = XA \tag{16}$$

$$V = YB \tag{17}$$

where $\mathrm{U} \in \mathbb{R}^{n \times k}$ and $\mathrm{V} \in \mathbb{R}^{n \times p}$ are the canonical variates, and $k$ is the number of canonical correlations ($typically\ k \leq min\,(p, q)$). Before applying CCA, it is essential to standardize the features to have zero mean and unit variance:

$$X_{std} = \frac{X - \mu x}{\sigma x} \tag{18}$$

$$Y_{std} = \frac{Y - \mu y}{\sigma y} \tag{19}$$

where $\mu x$ and $\sigma x$ are the means and standard deviations of the columns of $X$, and $\mu y$ and $\sigma y$ are the means and standard deviations of the columns of $Y$.

The standardized text and audio features are then projected onto the canonical directions to obtain the canonical variates:

$$U = X_{std}A \text{ and } V = Y_{std}B \tag{20}$$

Finally, the canonical variates U and V are concatenated to form a combined feature set:

$$Z = [U, V] \tag{21}$$

where as $Z \in \mathbb{R}^{n \times 2k}$ is the concatenated feature matrix.

### *3.5 Depression Detection*

In this phase as shown in Algorithm 3, the concatenated features from the text and audio data, processed through CCA, are used as input to a neural network designed for depression detection. This neural network leverages the combined feature set to accurately predict the likelihood of depression. The network consists of multiple dense layers with ReLU activation, dropout layers for regularization, and a sigmoid output layer for probability scoring. The input layer receives the concatenated features $Z \in \mathbb{R}^{n \times 2k}$ where n is the number of samples, and 2 $k$ is the dimensionality of the combined feature set from CCA.

**Algorithm 3:** Depression detection using CCA with neural network

**Require:** Text data $T = \{t_1, t_2, ..., t_n\}$, Speech data $S = \{s_1, s_2, ..., s_n\}$
**Ensure:** Depression prediction P
1: Initialize T5 model $T5_{model}$ and WaveNet model $WaveNet_{model}$
2: **Load** pre-trained weights for $T5_{model}$ and $WaveNet_{model}$
3: **Feature Extraction from Text**
4: **for** each text sample $t \in T$ **do**
5: $\mathbf{Z}_t \leftarrow T5_{model}(t)$ // Extract features using T5 model
6: **end for**
7: **Feature Extraction from Speech**
8: **for** each speech sample $s \in S$ **do**
9: $\mathbf{Z}_s \leftarrow$ WaveNetmodel(s) // Extract features using WaveNet model
10: **end for**
11: **Combine Features using CCA**
12: **Initialize** CCA model $CCA_{model}$
13: **Fit** CCAmodel on text features $\mathbf{Z}_t$ and speech features $\mathbf{Z}_s$

(Continued)

**Algorithm 3 (continued)**

14: **Transform** features using $\text{CCA}_{\text{model}}$:

$$\boldsymbol{Z}_t^{CCA}, \boldsymbol{Z}_s^{CCA} \leftarrow CCA_{model}.transform\,(\boldsymbol{Z}_t, \boldsymbol{Z}_s)$$
$$\boldsymbol{Z}_{CCA} \leftarrow \left[\boldsymbol{Z}_t^{CCA}, \boldsymbol{Z}_s^{CCA}\right]$$

15: **Depression Prediction**
16: **Initialize** classifier model $Classifier_{\text{model}}$
17: **Load** pre-trained weights for $Classifier_{\text{model}}$
18: $\mathbf{P} \leftarrow Classifier_{\text{model}}(\mathbf{Z}_{\text{CCA}})$ // Predict depression based on combined features
19: **return P**

The multiple dense (fully connected) layers transform the input features. Each dense layer applies a linear transformation followed by a non-linear activation function. The common activation function used is the Rectified Linear Unit (ReLU), which is defined as $ReLU(x) = max(0, x)$. The first dense layer transforms the input $z_i$ for the $i$th sample:

$$h_1 = ReLU(W_{1zi} + b_1) \tag{22}$$

where $W_1$ and $b_1$ are the weights and biases of the first dense layer, and $h_1$ is the output of the first dense layer. Subsequently, dropout layers are implemented to mitigate overfitting by randomly deactivating a portion of the input units during each training update. The dropout rate, which is a hyperparameter, specifies the fraction of units to be dropped. For instance, following the initial dense layer:

$$h_1^{drop} = Dropout(h_1, p) \tag{23}$$

where $p$ is the dropout rate and $h_1^{drop}$ is the output after applying dropout. The process can be repeated for additional dense layers if present. For instance, a second dense layer followed by dropout might be defined as:

$$h_2 = ReLU(W_2 h_1^{drop} + b_2) \tag{24}$$

where $W_2$ and $b_2$ are the weights and biases of the second dense layer, and $h_2$ is the output of the second dense layer. Applying dropout to this layer would be:

$$h_2^{drop} = Dropout(h_2, p) \tag{25}$$

The final layer is a dense layer with a sigmoid activation function, which outputs a probability score between 0 and 1 indicating the likelihood of depression. The sigmoid function is defined as $\sigma\,(x) = \dfrac{1}{1 + e^{-x}}$. The output of the $i$th sample is:

$$y_i' = \sigma(\mathrm{W_o h_{last}} + \mathrm{b_o}) \tag{26}$$

where $\mathrm{W_o}$ and $\mathrm{b_o}$ are the weights and biases of the output layer, $\mathrm{h_{last}}$ is the output of the last hidden layer, and $y_i'$ is the predicted probability of depression for the $i$th sample.

### *3.6 Model Training*

The network is trained using the binary cross-entropy loss function, which is defined as:

$$L = -\frac{1}{n}\sum_{i=1}^{n} (\mathrm{y_i} \log(\mathrm{y_i'}) + (1 - \mathrm{y_i}) \log(1 - \mathrm{y_i'})) \tag{27}$$

where $\mathrm{y_i}$ is the true label (0 or 1) for the $i$th sample, and $\mathrm{y_i'}$ is the predicted probability. The AdamW optimizer is used to update the network weights. AdamW is an extension of the Adam optimizer with

weight decay regularization:

$$w_{t+1} = w_t - \eta\left(\frac{m't}{\sqrt{v't+\epsilon}} + \lambda \mathrm{w_t}\right) \tag{28}$$

where $\eta$ is the learning rate, $m't$ are the bias-corrected first and second moment estimates, respectively, and $\lambda$ is the weight decay coefficient.

In conclusion, by integrating these components, the neural network can effectively leverage the combined text and audio features to detect depression with high accuracy. The use of CCA ensures that the features are optimally correlated, enhancing the overall performance of the neural network in this multimodal analysis task.

## 4 Experimental Results and Evaluation

This section presents the experimental results and assesses the effectiveness of the proposed method. Various experiments were conducted to evaluate the accuracy and efficiency of the developed system. The experimental evaluation demonstrated that the proposed method significantly outperforms current state-of-the-art techniques.

### *4.1 Performance Measures*

To evaluate the performance of MDD the following benchmark matrices has been used:

- **Accuracy:** Accuracy is the ratio of correctly predicted instances to the total instances. It is a measure of the overall effectiveness of a classification model.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Sample} \tag{29}$$

- **Precision:** Precision, or Positive Predictive Value, is the measure of correctly identified positive observations in relation to all observations predicted as positive. It reflects the proportion of true positive cases among the predicted positive instances, indicating the accuracy of positive predictions.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{30}$$

- **Recall:** It may also be known as Sensitivity or True Positive Rate, is the ratio of correctly predicted positive observations to all observations in the actual class.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{31}$$

- **Precision-Recall AUC** (**Area under the Curve**): is an evaluation metric used to assess the performance of a binary classification model. The Precision-Recall curve plots Precision ($y$-axis) against Recall ($x$-axis) for different threshold values.
- **Log Loss:** Also known as Logistic Loss or Cross-Entropy Loss, this performance metric evaluates a classification model's prediction accuracy. It measures how closely the predicted probabilities match the actual outcomes in binary or multiclass classification scenarios.

$$Log\ Los = -\frac{1}{N}\sum_{i=1}^{n}(\mathrm{y_i}\log(\mathrm{p_i}) + (1-\mathrm{y_i})\log(1-\mathrm{p_i})) \tag{32}$$

Here, $\mathrm{y_i}$ is the true label (0 or 1) and $\mathrm{p_i}$ is the predicted probability of the sample being in Class 1.

### *4.2 Results*

In the initial phase, the experiment evaluated the efficacy of MDD by measuring its precision, accuracy, and recall. The experimental results, graphically presented in Fig. 2, highlight the performance of the proposed approach across different datasets. Fig. 2 shows that the technique achieved high scores on all datasets, indicating impressive precision, accuracy, and recall. The DAIC-WOZ model demonstrates a strong performance with an accuracy of 92.75%, precision of 92.05%, and recall of 92.22%. This indicates a highly reliable model for identifying and classifying relevant cases. The E-DAIC model, while slightly less accurate with an accuracy of 91.74%, maintains respectable precision and recall scores of 90.35% and 90.95%, respectively, suggesting consistent performance, though with room for improvement compared to DAIC-WOZ. The CD-III model exhibits the highest performance among the three, achieving an accuracy of 93.05%, precision of 92.12%, and recall of 92.85%.

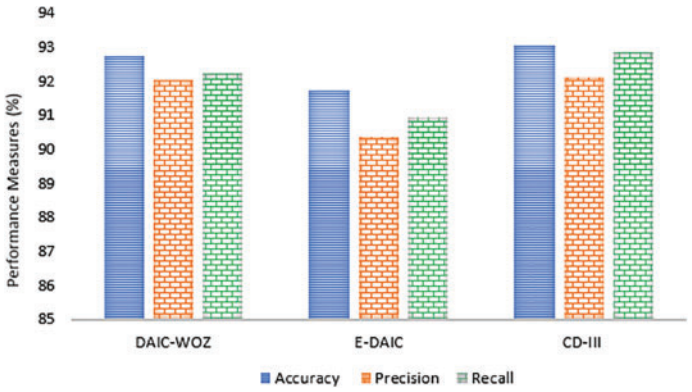
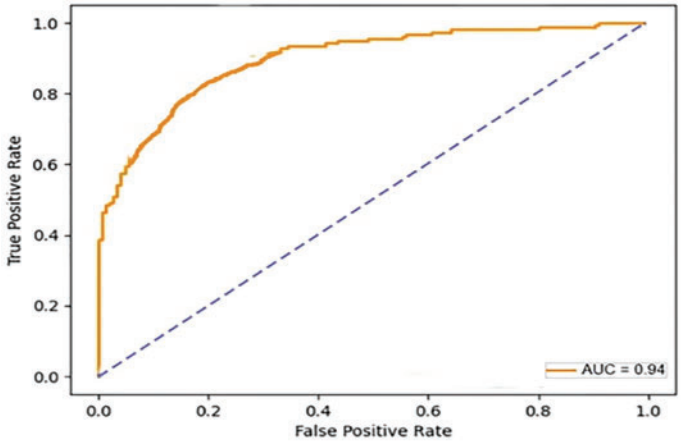


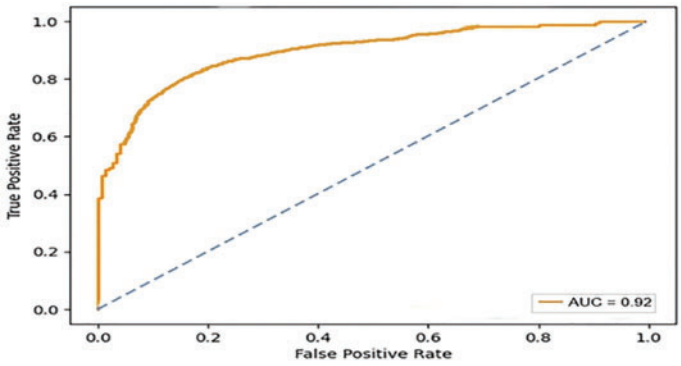

**Figure 2:** Experimental results on multiple datasets

In another experiment, ROC curves were employed to evaluate the effectiveness of the proposed approach in distinguishing between True and False instances, as illustrated in Fig. 3. The model demonstrated a notable average AUC of 0.93 across all datasets, indicating a high true positive rate while maintaining a low false positive rate across various classification thresholds.

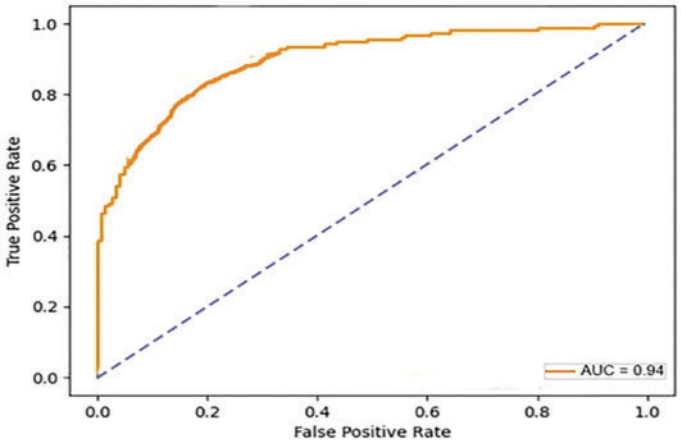



(a): The True Positive and False Positive rate of proposed mode on DAIC-WOZ

**Figure 3:** (Continued)




(b): The True Positive and False Positive rate of proposed model on E-DAI




(c): The True Positive and False Positive rate of proposed model on CCD-III

**Figure 3:** ROC Curves on DAIC-WOZ, E-DAI and CCD-III

The experimental results after the log loss base evaluation has been summarized in Table 3. The results demonstrate the effectiveness of MDD compared to all three baseline models. The results of Xia et al., serving as a foundational reference, showed the highest Log Loss values across all datasets, indicating the least effective performance [27]. Li et al. introduced more sophisticated feature extraction methods, resulting in a noticeable reduction in Log Loss values [28]. Das et al., with even more advanced techniques, achieved further improvements, showcasing a more robust model with better classification accuracy [29]. Ye et al. proposed a model for depression and achieved better results [30]. However, the proposed model significantly outperformed all baseline models, achieving the lowest Log Loss values across all datasets. Specifically, the proposed model attained a Log Loss of 0.320 for DAIC-WOZ dataset, 0.330 for E-DAIC, and 0.325 for CD-III. These results highlight the superior performance of the proposed model in accurately detecting depression, attributable to its advanced feature extraction methods and effective multimodal fusion techniques.

**Table 3:** Log loss comparison of depression detection models

| Dataset | Xia et al. [27] | Li et al. [28] | Das et al. [29] | Ye et al. [30] | MDD (Proposed model) |
|---|---|---|---|---|---|
| DAIC-WOZ | 0.450 | 0.420 | 0.390 | 0.48 | 0.320 |
| E-DAIC | 0.460 | 0.430 | 0.400 | 0.46 | 0.330 |
| CD-III | 0.455 | 0.425 | 0.395 | 0.45 | 0.325 |

The substantial reduction in Log Loss values for the proposed model underscores its capability to capture and classify depression-related features more precisely, thereby providing more reliable and accurate depression detection compared to the baseline models.

The comparative analysis of the proposed model, illustrated in Fig. 4, demonstrates incremental improvements over three baseline models, culminating in the superior performance of the proposed approach. Baseline 1 establishes a moderate foundation, with accuracy, precision, and recall metrics ranging from 84.56% to 85.96%, indicating adequate but improvable performance in emotion detection. Baseline 2 shows slight improvements, with accuracy rising to 87.56%, precision to 86.14%, and recall to 85.5%, reflecting subtle refinements in the model's capacity to accurately identify and classify emotional expressions. Baseline 3 marks a significant enhancement, achieving 90% across all metrics, suggesting a more robust model balancing the identification of relevant cases with classification accuracy. The proposed model, however, surpasses all baselines, achieving an accuracy of 92.51%, precision of 91.50%, and recall of 92.03%. This superior performance indicates that modifications in the model's algorithm or underlying technologies, potentially including advanced feature extraction methods and improved classification algorithms, have significantly boosted its efficacy.

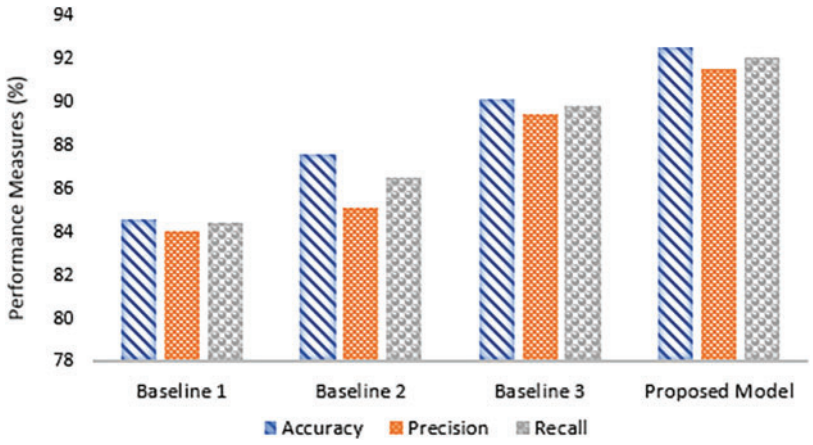



**Figure 4:** Comparative analysis of proposed model with baseline approaches

This superior performance of the proposed model as compared to baseline approaches can be attributed to several key factors. First, the proposed model likely utilizes a more advanced multimodal fusion strategy, effectively integrating text, audio, and potentially other features in a way that captures more nuanced patterns than the pre-fusion strategy used in Baseline 1. Additionally, the model may employ a more sophisticated feature extraction process, possibly involving deeper neural networks, optimized convolutional operations, or better architectural design, leading to enhanced precision and recall compared to Baseline 2's Spatial-Temporal Network and Relation Global Attention

mechanisms. In terms of audio processing, the proposed model appears to exceed the capabilities of Baseline 3, which focuses primarily on audio-based depression detection. By incorporating additional modalities alongside audio, the proposed model likely benefits from a more holistic approach, leading to better overall performance. Furthermore, the model's improved generalization capabilities suggest that it was trained with more effective regularization techniques, a diverse dataset, or more refined training processes, reducing overfitting and enhancing its ability to perform well on unseen data.

Table 4 presents feature extraction metrics for various approaches in a depression detection task. The CNN achieved an accuracy of 82.32%, with a precision of 81.12% and recall of 81.89%. While the CNN provides a solid foundation for feature extraction, its performance is relatively modest compared to more advanced methods. The Recurrent Neural Network (RNN) shows an improvement, with an accuracy of 85.45%, precision of 84.43%, and recall of 85.03%. This enhancement reflects the RNN's ability to better handle sequential data, capturing temporal dependencies more effectively than the CNN. The Bidirectional Long Short-Term Memory (BiLSTM) network performs even better, with an accuracy of 88.23%, precision of 87.34%, and recall of 87.89%. The BiLSTM's bidirectional architecture allows it to capture context from both past and future inputs, which significantly improves its ability to recognize depression-related features. The T5 + WaveNet model outperforms the others, achieving an accuracy of 92.42%, precision of 91.12%, and recall of 91.78%. This approach leverages the T5 model for advanced text feature extraction and the WaveNet model for detailed speech feature analysis. The combination of these models provides superior performance in accurately detecting depression specific features. This demonstrates the effectiveness of integrating multimodal data and advanced feature extraction techniques to achieve high diagnostic accuracy.

**Table 4:** Comparison of proposed feature extraction with CNN, RNN and BiLSTM

| Algorithms | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| CNN | 82.32 | 81.12 | 81.89 |
| RNN | 85.45 | 84.43 | 85.03 |
| BiLSTM | 88.23 | 87.34 | 87.89 |
| T5 + WaveNet (Proposed approach) | 94.42 | 93.12 | 93.78 |

## 5 MDD Recommendations

Given the significant challenges in diagnosing depression and the limitations of traditional methods, the proposed Multimodal Depression Detection (MDD) system leverages advanced AI technologies to enhance the accuracy and robustness of depression diagnosis. Below are key recommendations to ensure the successful implementation and operation of the MDD system:

- Utilizing advanced AI models like T5 and WaveNet ensures a more accurate diagnosis of depression, reducing the chances of misdiagnosis.
- Early detection through automated systems allows for quicker intervention, potentially mitigating the severity of depression in individuals.
- The MDD system can be accessed remotely, providing diagnostic support to individuals in underserved or remote areas where mental health services are limited.
- The system offers a reliable second opinion, supporting healthcare providers in making informed diagnostic decisions.

- Promote awareness and acceptance through educational initiatives and collaboration with mental health organizations.
- Combining text and speech data provides a holistic view of an individual's mental state, leading to a more thorough understanding and diagnosis.

## 6 Conclusion and Future Work

Depression is a widespread mental health issue, challenging to diagnose due to societal stigma and lack of awareness. Traditional methods like in-person therapies and questionnaire-based assessments are time-consuming and often inaccurate, highlighting the need for AI assistance. Our research proposed the MDD approach, utilizing multimodal data collection, advanced feature extraction models (T5 for text and WaveNet for speech), CCA for feature projection, and a neural network for detection. The results showed significant improvements in accuracy and reliability over traditional methods. Future research directions include enhancing the proposed MDD framework by integrating additional modalities such as facial expressions and physiological signals to further improve the accuracy and robustness of depression detection. Additionally, exploring real-time application and deployment in clinical settings will be crucial.

**Acknowledgement:** The authors acknowledged the support of King Saud University, Saudi Arabia.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** The authors confirm contribution to the paper as follows: study conception and design: Farah Mohammad, data collection, analysis and interpretation of results: Khulood Mohammed Al Mansoor. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The dataset is available on CD-III, https://github.com/AQR315/depression_dataset (accessed on 24 October 2024).

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

[1] L. Squarcina, F. M. Villa, M. Nobile, E. Grisan, and P. Brambilla, "Deep learning for the prediction of treatment response in depression," *J. Affect. Disord.*, vol. 281, no. 1, pp. 618–622, 2021. doi: 10.1016/j.jad.2020.11.104.

[2] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," in *Proc. Fifth Workshop Comput. Linguist. Clin. Psychol.: Keyboard Clinic.*, 2018, pp. 88–97. doi: 10.18653/v1/W18-0610.

[3] A. Amanat *et al.*, "Deep learning for depression detection from textual data," *Electronics*, vol. 11, no. 5, 2022, Art. no. 676. doi: 10.3390/electronics11050676.

[4] A. Sarkar, A. Singh, and R. Chakraborty, "A deep learning-based comparative study to track mental depression from EEG data," *Neurosci. Inform.*, vol. 2, no. 4, 2022, Art. no. 100039. doi: 10.1016/j.neuri.2022.100039.

[5] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: A deep learning approach," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1465–1474, 2021. doi: 10.1109/TCSS.2021.3084154.

[6] H. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. A. Hameed, "Applying deep learning technique for depression classification in social media text," *J. Med. Imaging Health Inform.*, vol. 10, no. 10, pp. 2446–2451, 2020. doi: 10.1166/jmihi.2020.3169.

[7] M. A. Wani *et al.*, "Depression screening in humans with AI and deep learning techniques," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 2074–2089, 2022. doi: 10.1109/TCSS.2022.3200213.

[8] A. Safayari and H. Bolhasani, "Depression diagnosis by deep learning using EEG signals: A systematic review," *Med. Novel Technol. Dev.*, vol. 12, no. 2, 2021, Art. no. 100102. doi: 10.1016/j.medntd.2021.100102.

[9] P. Meshram and R. K. Rambola, "Diagnosis of depression level using multimodal approaches using deep learning techniques with multiple selective features," *Expert. Syst.*, vol. 40, no. 4, 2023, Art. no. e12933. doi: 10.1111/exsy.12933.

[10] N. Marriwala and D. Chaudhary, "A hybrid model for depression detection using deep learning," *Meas.: Sens.*, vol. 25, 2023, Art. no. 100587. doi: 10.1016/j.measen.2022.100587.

[11] P. P. Thoduparambil, A. Dominic, and S. M. Varghese, "EEG-based deep learning model for the automatic detection of clinical depression," *Phys. Eng. Sci. Med.*, vol. 43, no. 4, pp. 1349–1360, 2020. doi: 10.1007/s13246-020-00938-4.

[12] G. Lam, D. Huang, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *ICASSP 2019–2019 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 3946–3950. doi: 10.1109/ICASSP.2019.8683027.

[13] K. M. Hasib *et al.*, "Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey," *IEEE Trans. Comput. Soc. Syst.*, vol. 10, no. 4, pp. 1568–1586, 2023. doi: 10.1109/TCSS.2023.3263128.

[14] A. Haque, V. Reddi, and T. Giallanza, "Deep learning for suicide and depression identification with unsupervised label correction," in *Artif. Neural Netw. Mach. Learn.-ICANN 2021: 30th Int. Conf. Artif. Neural Netw.*, Bratislava, Slovakia, 2021, pp. 436–447. doi: 10.1007/978-3-030-86368-2_35.

[15] M. Kang *et al.*, "Deep-asymmetry: Asymmetry matrix image for deep learning method in pre-screening depression," *Sensors*, vol. 20, no. 22, 2020, Art. no. 6526. doi: 10.3390/s20226526.

[16] N. P. Shetty *et al.*, "Predicting depression using deep learning and ensemble algorithms on raw twitter data," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 4, pp. 3751–3756, 2020. doi: 10.11591/ijece.v10i4.pp3751-3756.

[17] I. A. Anik *et al.*, "A robust deep-learning model to detect major depressive disorder utilising EEG signals," *IEEE Trans. Artif. Intell.*, vol. 5, no. 10, pp. 4938–4947, 2024. doi: 10.1109/TAI.2024.3394792.

[18] F. Rehmani *et al.*, "Depression detection with machine learning of structural and non-structural dual languages," *Healthc. Technol. Lett.*, vol. 11, no. 4, pp. 218–226, 2024. doi: 10.1049/htl2.12088.

[19] B. Manjulatha and S. Pabboju, "Multimodal depression detection using deep learning in the workplace," in *Proc. 2024 Fourth Int. Conf. Adv. Electr., Comput., Commun. Sustainable Technol. (ICAECT)*, 2024, pp. 1–8. doi: 10.1109/ICAECT57636.2024.9827582.

[20] K. Katiyar, H. Fatma, and S. Singh, "Predicting anxiety, depression and stress in women using machine learning algorithms," in *Combating Women's Health Issues with Machine Learning*. Boca Raton, FL, USA: CRC Press, 2024, pp. 22–40. doi: 10.1201/9781003225946-3.

[21] U. K. Lilhore *et al.*, "Unveiling the prevalence and risk factors of early stage postpartum depression: A hybrid deep learning approach," *Multimed. Tools Appl.*, vol. 83, no. 26, pp. 1–35, 2024. doi: 10.1007/s11042-024-18182-3.

[22] T. T. Prama *et al.*, "AI-enabled deep depression detection and evaluation informed by DSM-5-TR," *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 5, 2024. doi: 10.1109/TCSS.2023.3249682.

[23] S. Dalal, S. Jain, and M. Dave, "Convolution neural network having multiple channels with own attention layer for depression detection from social data," *New Gener. Comput.*, vol. 42, no. 1, pp. 135–155, 2024. doi: 10.1007/s00354-023-00263-2.

[24] R. Dou and X. Kang, "TAM-SenticNet: A neuro-symbolic AI approach for early depression detection via social media analysis," *Comput. Electr. Eng.*, vol. 114, no. 6, 2024, Art. no. 109071. doi: 10.1016/j.compeleceng.2023.109071.

[25] W. Zhang *et al.*, "Depression detection using digital traces on social media: A knowledge-aware deep learning approach," *J. Manag. Inf. Syst.*, vol. 41, no. 2, pp. 546–580, 2024. doi: 10.1080/07421222.2024.2340822.
[26] S. Khan and S. Alqahtani, "Hybrid machine learning models to detect signs of depression," *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 38819–38837, 2024. doi: 10.1007/s11042-023-16221-z.
[27] Y. Xia *et al.*, "A depression detection model based on multimodal graph neural network," *Multimed. Tools Appl.*, vol. 83, no. 23, pp. 1–17, 2024. doi: 10.1007/s11042-023-18079-7.
[28] J. Li *et al.*, "Image encoding and fusion of multi-modal data enhance depression diagnosis in Parkinson's disease patients," *IEEE Trans. Affect. Comput.*, pp. 1–16, 2024. doi: 10.1109/TAFFC.2024.3418415.
[29] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomed. Signal Process. Control*, vol. 90, no. 1–3, 2024, Art. no. 105898. doi: 10.1016/j.bspc.2023.105898.
[30] J. Ye, Y. Yu, Q. Wang, W. Li, and H. Liang, "Multi-modal depression detection based on emotional audio and evaluation text," *J. Affect. Disord.*, vol. 295, no. 1, pp. 904–913, 2021. doi: 10.1016/j.jad.2021.08.090.
[31] A. Mastropaolo *et al.*, "Studying the usage of text-to-text transfer transformer to support code-related tasks," in *Proc. 2021 IEEE/ACM 43rd Int. Conf. Softw. Eng. (ICSE)*, 2021, pp. 336–347. doi: 10.1109/ICSE.2021.00044.
[32] J. Chorowski *et al.*, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, 2019. doi: 10.1109/TASLP.2019.2938863.