CURRENT TRENDS IN INFORMATION COMMUNICATION TECHNOLOGY RESEARCH

BABCOCK UNIVERSITY
KNOWLEDGE TRUTH SERVICE
ILISHAN - REMO, NIGERIA.

CTICTR 2(1): 28 - 46 (June 2023)

ISSN: 2814-1709

Received:00-04-2023

Accepted:00-06-2023

A DEEP LEARNING APPROACH TO SPEECH RECOGNITION FOR DETECTION OF MENTAL DISORDERS

Samuel A. Bindas

Ernest E. Onuiri, PhD.

Bindas1220@pg.babcock.edu.ng
Department of Computer Science
Babcock University
Ilishan-Remo, Ogun State, Nigeria

A Deep Learning Approach to Speech Recognition for Detection of Mental Disorders

Samuel Bindas Amile[a]*, Ernest Onuiri[b]

[a] *Babcock University, Ilishan Remo, Ogun State, 121003, Nigeria*

[b] *Babcock University, Ilishan Remo, Ogun State, 121003, Nigeria*

[a]*bindas1220@pg.babcock.edu.ng*

[b]*onuirie@babcock.edu.ng*

Abstract

Mental disorders are conditions that affect a person's cognitive functions, behavior or thinking, thereby impairing daily functions. The dearth of trained psychologists to the high number of patients living with a mental disorder pose significant challenges in the field of mental health. This study investigates the application of deep learning techniques to speech recognition for the purpose of detecting mental disorders. The main objective of this study is to effectively identify speech patterns associated with various mental disorders and thereafter develop a robust and accurate deep learning model system that can detect risk of a mental disorder in an individual based on their voice. The research methodology involved the collection of a dataset consisting of speech recordings from individuals diagnosed with depression and post-traumatic stress disorder (PTSD). The dataset acquired was carefully curated to include symptom severity levels, and linguistic variations. The results of this study demonstrate the effectiveness of deep learning approaches in speech recognition for mental disorder detection. The trained models achieved 95% and 94% accuracy rates in identifying and differentiating speech patterns associated with depression and PTSD respectively.

The findings of this study have significant implications for the field of mental health. The developed deep learning system offers a promising avenue for the early detection and monitoring of mental disorders. Further research is warranted to validate and refine the developed models using larger and more diverse dataset. Additionally, the integration of multimodal data, such as combining speech analysis with psychological or text-based data, could enhance the diagnostic accuracy and reliability of the system.

Keywords: Deep Learning, Mental Disorder, Post-traumatic Stress Disorder, Speech Recognition, Depression.

1.0 Introduction

According to the World Health Organization (WHO), mental disorders are categorized by a combination of maladaptive thoughts, emotions, behaviours, and interrelationships [1]. Common mental disorders include depression, post-traumatic stress disorder (PTSD), bipolar disorder, schizophrenia, eating disorders, and speech disorders such as attention-deficit hyperactivity disorder (ADHD) and autism [1]. Mental disorders are a significant public health issue that can have profound impacts on an individual's quality of life, relationships, and ability to work or participate in other activities.

---

* Corresponding author.

Recent studies indicate that 12.5% of the total global population are living with a mental disorder, with anxiety and depressive disorders the most common [2]. This number has spiked, especially with the advent of COVID-19, due to a widespread increase in anxiety and major depressive disorders (MDD). In 2020, estimates reveal an almost 30% increase for anxiety and MDDs [2]. Although mental health conditions throughout lifetime are gaining recognition in present times, they remain a public health challenge globally. Mental disorders reduce life expectancy in affected individuals by 10 – 15 years [3]. Timely interventions at the onset of disease increase the chances of better outcomes. Early prevention in susceptible individuals can significantly improve clinical outcomes [4], [5]. For instance, 1 in every 4 adolescents with attenuated symptoms for psychosis accumulate several risk factors and develop the disorder over several years [6]. Treatment for these individuals is often implemented in specialized clinical services and can mitigate or suspend the development of psychosis, however, there is a need for more quality research concerning the efficacy of specific preventive interventions [7]. Routine early surveillance of mental conditions is vital to improving mental health and general well-being throughout a lifetime.

Mental disorders, overall, have a greater impact economically than cancer, diabetes, and cardiovascular diseases, but receive limited focus and funding [8]. Recent approaches to monitoring and evaluation of mental conditions depend on intermittent reports from caregivers or patients. These reports are subjective and involve sufferers' call biases, cognitive limitations. Therefore, there is a pressing need to adequately evaluate, and present evidence-based interventions for mental disorders, especially in individuals with limited access to conventional psychiatric services [9]. While there are several machine learning techniques that can be used to detect mental disorders, deep learning can prove to achieve this better due to its ability to automatically learn complex representations of data without the need for manual feature engineering.

Deep learning is a subset of machine learning that utilizes artificial neural networks with multiple layers to extract complex features from data. This approach has shown significant improvements in various domains, including speech recognition. In speech recognition, deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used to achieve state-of-the-art results in a variety of tasks, including image classification, natural language processing, and speech recognition.

Speech recognition is a subset of deep learning and is the process of detecting and converting spoken language into text or computer-readable format for further analysis. This analysis can be done using various techniques, including acoustic analysis and linguistic analysis. Acoustic analysis is a technique used to examine the physical characteristics of speech sounds, including pitch, volume, duration, and frequency. It involves measuring and analysing the sound waves produced by speech using specialized software. Research has shown that individuals with depression tend to speak in a monotone voice with lower pitch and reduced variation in pitch and volume compared to individuals without depression. In contrast, individuals with bipolar disorder may have a more erratic pitch and volume pattern in their speech [10].

Furthermore, speech analysis is vital to providing more insight into the mental activities of humans, which can be learned by Artificial Intelligence (AI) for timely screening and diagnosis [1]. Studies show a dearth of data regarding mental disorder prevalence in developing countries, compared with developed countries, due to insufficient funding, backing, and facilities [11].

2.0 Literature Review

The application of machine learning technology in data analysis elicited form sensors for mental disorder evaluation has the capacity to screen for individuals susceptible to mental disorders prior to accessing the mental health care system [9]. Furthermore, the potential of machine learning technology is recognized in its capacity to complement clinicians' assessment once patients seek care; and detect symptoms when patients depart the facility. This sub-section details these functions in the following paragraphs.

Firstly, machine learning can address the barriers limiting patients' accessibility to mental health diagnoses and treatments. A correlational analysis of surveys in low-resource setting [12] reports that the major hindrances to treatment for mental disorders include funding, social discrimination, low awareness levels, time limitation, and a lack of resources. There are also the constraints of physical disabilities [13] and individuals in war-torn regions [14]. People with no access to the mental healthcare system can still evaluate their mental well-being with vast technologies [15]. Through cross-sectional data, these resources can access treatment choices by understanding the success rate of various treatments given specific symptomatology [16]. In addition, such technologies can aid screening in educational institutions, military, and primary care settings.

Secondly, machine learning can enhance evaluation within the health facility given certain barriers facing clinicians. Following patients' access to the mental healthcare system, there is the potential difficulty in detecting disorders that may be sporadic and have high comorbidity rates. This causes difficulty in separating overlapping symptoms into underlying discrete diagnoses. There is a complication in developing a model to identify specific disorders when individuals exhibit comorbidities or have sporadic symptoms. This is evident in PTSD, where half the cases are accompanied with depression or drug abuse problems [17]. Additionally, suicidality is vital in predictive models and is evident in several disorders. The National Institute of Mental Health (NIMH) established the Research Domain Criteria with the aim to analysing diagnoses with biomarkers to forecast and enhance treatment response [18]. Thus, algorithms based on behavioural descriptors can predict various disorders to assist in differential diagnosis, identify susceptibility for long-term mental disorders, symptomatic episodes, or suicidality; and progressively forecast the most effective treatment given multimodal data [19]. Hence, integrating clinical interviews with machine learning technologies based on the tapes of these interviews can enhance results, save time, promote cost-effectiveness, and increase the efficacy of treatment planning.

Lastly, machine learning models can enhance mental health care by promoting routine real-time monitoring of symptoms. For instance, in cases of non-return of individuals after meeting a clinician, there is the option of remotely observing and examining mental health and subsequent help-seeking behaviour. Additionally, in cases of long-term patients paying regular visits, symptoms may be sporadic in between visits. Sensors can identify emergency episodes or warning signs and adopt online resources or automated therapy to improve outcomes [20]. Individuals and mental health professionals, through real-time monitoring, can observe behaviour, conduct early identification of episodes, schedule assessments, and customize treatment.

The potential discussed in this section has yet to be achieved. Several studies in the area adopt small sample sizes, which underpowers research outcomes. Models based on limited observations of a particular data type (e.g., taped in a quiet corner, white interviewees) may not infer to seemingly similar data. Additionally, algorithms are predisposed to learning biases inherent in the data used in

training them (e.g., wrongly attributing lower disorder prevalence to the ethnic group due to fewer African Americans reporting the disorder in the training set) [21], [22].

Crucially, several high-performing algorithms are complicated, considering it is not yet certain how these models combine features to inform the severity of a disorder. This ambiguity leads to a lack of trust since they have demonstrated proneness to adversarial attacks [23]. It is for this reason that the EU regulation requires a right to obtain an account for crucial decisions from automated algorithms, such as clinical evaluations.

2.1. Speech as a biomarker for mental health

Speech comes to many humans without any prior consideration of how complex it is. Speaking goes beyond mere forming words with the tongue; it is the expression of human communication through articulated thought, resolve, and feeling in a concerted performance. Speech requires a network of muscles and co-ordination of brain regions processing auditory and visual input [24]. Hence, verbal interactions are a pathway to the mind, and present several opportunities for an array of technologies to capture and process speech to detect mental disorders.

There is evidence of the detective function of speech patterns in mental health evaluation. Over a century ago, the famous German psychiatrist, Emil Kraepelin theorized that the speech patterns of depressed patients were by nature, low-pitched, repetitive, quieter, tentative, and with a relatively higher degree of stammering [25]. Compared with other behavioural indicators, speech has a range of benefits including ease of detecting symptoms, direct representation of feelings and thinking via its language content, and an indirect reflection of neural modulation via motor and acoustic variation. Additionally, speech patterns cut across all languages, making it beneficial for low-resource languages when NLP technology is unavailable. It is less cumbersome and cost-effective to elicit speech patterns through smartphones and computers instead of more expensive wearables or invasive neuroimaging techniques, especially as many clinical interviews are already taped. Finally, it is a data form made more accessible given the enhancement in speech detection, as applied in voice-activated interfaces such as Alexa (on Amazon), Siri (Apple), and Cortana (Microsoft), and voice biometrics for security and education [26].

2.2. Related works

Several studies have linked speech recognition and mental disorders [27]–[29]. Related literature in this field have classified these associations into psychiatric disorders, behavioural challenges, and substance-use disorders. Studies such as [30], [31] reported a positive link between speech difficulty and mental disorders. In literature reviewing individual disorders (anxiety, depression, and bipolar disorder), two studies with comparative quality ratings reported dissimilar outcomes. The first study suggested a heightened risk of mood disorders in respondents with speech difficulties [32] while the other study found no association between respondents and their controls because minor to moderate language disorders in early adulthood might not lead to substantial long-term mental health impacts. [33]. Conversely, in studies investigating the subset of personality disorders (chronic, extreme, and sporadic thinking and behavioural patterns), studies reported more personality disorders in respondents with speech learning difficulty than the control group [31], [34]. In a systematic review, seven of the eleven reviewed studies documented a significant association between childhood speech learning difficulties and anxiety [31].

Furthermore, two of these studies suggested higher anxiety levels in their cohort at age 19 [31], [35], while another study indicated no difference at age 31 [33]. The variance could not be explained by change in language abilities over time. This could be due to the demanding environment experienced by respondents at adolescence resulting from their developmental stage, potential academic difficulties, and speech learning difficulties [33]. Studies revealed conflicting evidence in the investigation of the association of speech difficulties and mental disorders. Several studies report an association between depression and speech difficulty [36], [37]. Amongst the five studies with analogous quality ratings, three suggested higher depression in respondents with speech difficulties, whereas the other two studies reported no difference. Likewise, Lindsay [38] opines that the individuals with a history of speech learning difficulties had lower self-esteem than their controls at 16 years old but found no difference at age 17.

In studies investigating behavioural difficulties, authors observed a significant association between early onset speech difficulty and behavioural difficulties. The authors in [39] documented hyperactive behaviour and reactive temperaments amongst participants with a history of speech learning difficulties.

Some research works demonstrated general behavioural challenges; studies that investigated pro-social behaviour (empathic behaviour with the aim to helping others no expectation of reward) demonstrated challenges with this behaviour in adulthood for those with childhood speech learning difficulties [36], [40]–[42]. Three studies examined delinquent behaviour; one study [33] revealed significant relationship with speech learning difficulties whilst the other studies indicated no evidence of association [65].

Emotional problems in respondents with speech learning difficulties are documented by [36]. Similarly, another study indicated a decrease in emotional problems over time; however, their prevalence in participants with childhood speech learning difficulties remained higher than in the general population [43]. Conversely, studies revealed no linkage between early onset speech learning difficulty and maladaptive behaviour [30], [43].

Substance use disorders were examined by five studies; one study reported that adults with difficulty in speech learning were predisposed to alcohol misuse or abuse [32]. However, the other papers reported no variation in rates of substance-use disorders between individuals with a history of speech learning difficulties and their controls [31], [43].

According to [44], speech rate is considered a predictor of depression and it is documented that patients with major depressive disorder spoke slower than patients without. One study, using audio of 7 individuals, found that speech rate revealed a strong negative association with Hamilton Depression Scale Score (HAMD) [45]. In [46], it was found that speech rate quickened after admission of antidepressant for patients with major depressive disorder. Pause time during speaking in individuals has also been reported to be associated with depression. Older research works demonstrate that pause time was longer for depressed patients than for non-depressed patients, and that pause time associated with HAMD scores [47]. Response time was analysed in only a few previous studies [48]. These studies engaged standard reading tasks or close-ended interviews to obtain data and revealed that response time was longer for depressed patients than for controls, and that response time shortened as HAMD scores reduced. Earlier studies have failed to extract the voices of research respondents and determine the speed and / or timing of speech. For instance, [47] interviewed respondents via the phone to obtain speech samples, while others had participants read close-ended statements to obtain only the participant's voice [44]. However, some other gaps exist in speech – mental disorder research. Firstly, there is the potential of speech features being influenced by the personality or speech habits of each respondent. However, there is also implication that speech features suggest depression severity.

Furthermore, in [44], healthy individuals were markedly older than patients with major depressive disorder. Finally, some drug side-effects such as insomnia are associated with speech rate features. This suggests the crucial need for a detailed side-effect profile to investigate these complex associations.

The detection of mental disorders using self-assessment questionnaires has proven to be the best way of diagnosing patients, these include Patient Health Questionnaire 8 (PHQ-8) for depression, and PTSD Checklist (PCL) for PTSD. [49] in the study of the validity of checklist as a measure of PTSD, compared the PCL to other measures of PTSD symptoms and found that the PCL was the most sensitive to changes in symptoms over time. They concluded that the PCL is a reliable and valid measure of PTSD symptoms, supporting its use in clinical and research settings. Another study also reported good internal consistency, test-retest reliability, and convergent validity for the PCL, supporting its use as a reliable and valid measure of PTSD symptoms [50]. PCL is depicted in the figure 2.2 below. Several studies have clearly backed PHQ as a valid and reliable measure for depression [51], [52]. PHQ, depicted in figure 2.1 below, is a self-report questionnaire that assesses the presence and severity of depressive symptoms based on the DSM-IV or DSM-5 diagnostic criteria. The most recent and suitable versions of the PHQ are the PHQ 8 and 9. PHQ-9 includes nine items, while the PHQ-8 includes eight of the nine items from the PHQ-9, excluding the item on suicidal ideation. While the PHQ-8 may be quicker to administer, the PHQ-9 may be more comprehensive and better validated and may be more appropriate in some settings where detecting suicide risk is a priority.

Table 2.1 PHQ-8 Questionnaire

Over the last 2 weeks, how often have you been bothered by any of the following problems?	**Not at all**	**Several days**	**More than half the days**	**Nearly every day**
Little interest or pleasure in doing things	0	1	2	3
Feeling down, depressed, irritable	0	1	2	3
Trouble falling or staying asleep, or sleeping too much	0	1	2	3
Feeling tired or having little energy	0	1	2	3
Poor appetite or overeating	0	1	2	3
Feeling bad about yourself – or that you are a failure or have let yourself or your family down	0	1	2	3
Trouble concentrating on things, such as schoolwork, reading or watching television	0	1	2	3
Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

Each question on the PHQ-8 is scored on a scale from 0 to 3, with 0 indicating "not at all" and 3 indicating "nearly every day". The total score on the PHQ-8 ranges from 0 to 24, with higher scores indicating more severe depression symptoms.

The scores on can be interpreted thus:

- 0-4: Minimal or no depression
- 5-9: Mild depression
- 10-14: Moderate depression
- 15-24: Severe depression

A cut-off score of 10 or higher can be used to indicate a probable diagnosis of major depressive disorder [52].

Table 2.2 PTSD Checklist

	Not at all	**A little bit**	**Moderately**	**Quite a bit**	**Extremely**
Repeated, disturbing memories, thoughts, or images of a stressful experience from the past?	0	1	2	3	4
Repeated, disturbing dreams of a stressful experience from the past?	0	1	2	3	4
Suddenly acting or feeling as if a stressful experience were happening again (as if you were reliving it)?	0	1	2	3	4
Feeling very upset when something reminded you of a stressful experience from the past?	0	1	2	3	4
Having physical reactions (e.g., heart pounding, trouble breathing, or sweating) when something reminded you of a stressful experience from the past?	0	1	2	3	4
Avoid thinking about or talking about a stressful experience from the past or avoid having feelings related to it?	0	1	2	3	4
Avoid activities or situations because they remind you of a stressful experience from the past?	0	1	2	3	4
Trouble remembering important parts of a stressful experience from the past?	0	1	2	3	4
Loss of interest in things that you used to enjoy?	0	1	2	3	4

Feeling distant or cut off from other people?	0	1	2	3	4
Feeling emotionally numb or being unable to have loving feelings for those close to you?	0	1	2	3	4
Feeling as if your future will somehow be cut short?	0	1	2	3	4
Trouble falling or staying asleep?	0	1	2	3	4
Feeling irritable or having angry outbursts?	0	1	2	3	4
Having difficulty concentrating?	0	1	2	3	4
Being “super alert” or watchful on guard?	0	1	2	3	4
Feeling jumpy or easily startled?	0	1	2	3	4
Having difficulty concentrating?	0	1	2	3	4
Trouble falling or staying asleep	0	1	2	3	4

Each question on the PCL is scored on a scale of 0 to 4, with 0 indicating “not at all” and 4 indicating “extreme”. The total score ranges from 20 to 80, with higher scores indicating more severe PTSD symptoms. A cut-off score of 38 or higher is often used to indicate a probable diagnosis of PTSD [53].

The accuracy of speech-based mental illness detection models has been shown to be promising in some studies, with some models achieving accuracy rates comparable to psychiatric evaluations.

The choice of speech features used in the models is critical for their accuracy, and a combination of multiple features such as speech rate, tone, and vocabulary usage has been shown to be more effective than using a single feature.

However, speech-based mental illness detection models also have several limitations, including the need for large and diverse training datasets, potential for biased results, and limited generalizability to different populations.

These findings suggest that while speech analysis has potential as a tool for mental illness detection, there is still a need for further research to address the limitations and improve the accuracy of speech-based detection models.

3.0 Materials and Methods

The development of the deep learning models involved several stages, including data collection, data pre-processing, feature extraction, model selection, and evaluation. Two models were developed, albeit using the same approach, for both PTSD and depression disorders.

3.1. Methods

The approach for this study used Convolutional Neural Network (CNN) architecture with a 1-dimensional convolutional layer (Conv1D) and Transformer Encoder architecture. This model uses a Conv1D because audio waveforms are one dimensional, as opposed to other patterns like images, which can have 2 or more dimensions. The model was built using Python programming language because of its extensive applicability in scientific computing and extensive libraries like TensorFlow, scikit-learn which were used.

CNN is a type of artificial neural network commonly used for image and audio classification tasks. The key concept behind CNN is to detect features (such as edges, lines, and shapes) in images or audio data by convolving small filters across the input data. In the case of this study where audio data is used, the filters were applied to the spectrogram and Mel-frequency cepstral coefficients (MFCCs) to detect patterns that are important for classification, such as changes in frequency over time. It is useful for audio classification because it can automatically extract relevant features from the audio signal without the need for manual feature engineering. This can lead to better performance and faster development time compared to traditional deep learning approaches that require manual feature extraction. Additionally, CNNs can learn hierarchies of features by combining multiple layers of convolutions, leading to even more accurate and complex classification models.

A transformer encoder is a Sequence-to-Sequence learning model brought forward by Vaswani in 2017 [54] and is responsible for summarizing audio representation. It uses a self-attention mechanism to process sequential input data such as audio and has been used to model speech spectrograms. The resulting representation is passed through a regression module, which outputs a score indicating presence of depression or PTSD. The transformer decoder is responsible for generating the output sequence. It takes the encoded representation of the input sequence from the encoder and produces the output tokens step by step. The decoder consists of several layers of self-attention and feed-forward neural networks, like the encoder. However, by using self-attention and encoder-decoder attention mechanisms along with positional encoding, the decoder can generate output sequences while considering the context of the input sequence encoded by the encoder. This makes the transformer architecture well-suited for audio classification.

3.2 Dataset

To achieve an acceptable model performance, the dataset that was be fed into the model as an input went through several pre-processing steps from collection. Figure 3.1 below shows the data flow diagram of the proposed models and detailed discussion of each step is found further below.

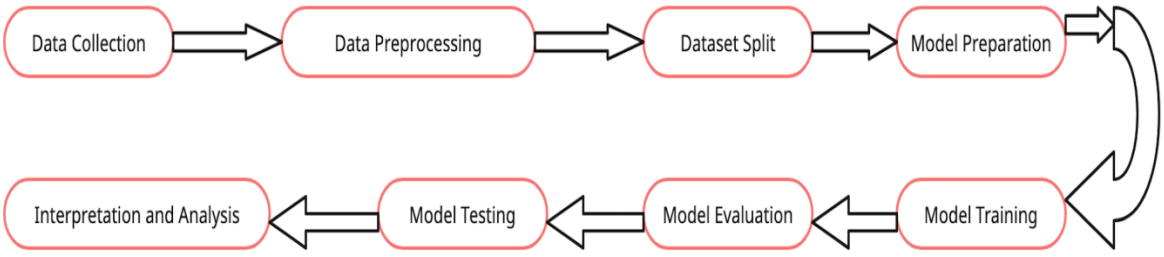

Figure 3.1. Data flow diagram of the proposed model

The model was conducted using Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset compiled by University of Southern Carolina's Institute of Creative Technologies [55]. It

contains clinical interviews that were designed to support the diagnosis of several mental disorders including depression and PTSD. The dataset was available for download upon request to the institute and upon agreeing to End User License Agreement and was approximately 65gb in size.

The dataset consists of interviews between a human interviewer and a participant who was instructed to discuss a recent distressing experience. The interviews were conducted using a standardized protocol, and the participants were asked to rate their level of distress at various points throughout the interview. The dataset includes audio recordings of the interviews, as well as the participants' self-reported distress ratings. It has been used to develop and evaluate deep learning models for detecting and predicting mental health outcomes [56]. The authenticity of this dataset was established through thorough documentation and annotation as well as the corpus underwent peer reviews by other experts [57].

A total of approximately 51 hours of data was gathered from 219 clinical interviews involving 191 patients, which included recorded clinical interviews and transcripts, and facial features. Each participant labelled 300 to 718 was assigned a PHQ-8 score PCL score as labels. Those with a PHQ-8 score of 10 or higher were considered to have depression, while a PCL score of 38 or higher were considered to have PTSD. On average, the duration of audio recordings for the 189 interviews was 974 seconds.

3.3 Data Pre-processing

The dataset contained an extracted question and answer pairs and its timestamps from the audio recording, as it is shown in figure 3.2. Using the answer only responses from the extracted timestamps, the answer audio was segmented from the original interview audio and stored separately.

	A	B	C
1	Start_Time	End_Time	Answer
2	14.3	15.1	so I'm going to
3	20.3	21.1	interview in Spanish
4	23.9	24.3	okay
5	62.1	62.7	good
6	68.8	69.8	Atlanta Georgia
7	74.8	77.1	my parents are from here
8	83.4	84.3	I love it
9	88.1	92.9	I like the weather I like the opportunities
10	104.2	105.3	at the minute
11	107.5	108.4	someone easy
12	113.8	115.1	congestion
13	120.2	120.9	that's it
14	128.2	131.8	I took up business and administration
15	136.6	143.8	yeah I am here and there I'm on a break right now but I plan on going back in the next semester
16	149	151.1	probably to open up my own business
17	159.3	161.4	no
18	165.5	168.5	no specific reason I just

Figure 3.2 Timestamps of audio extracted

3.3.1 Noise cleaning

This process involved removing any noise or artifacts that may have been picked up during the recording. This was achieved using a python library called Librosa.

3.3.2 Segmentation

The audio files were divided into smaller segments of 10 second per segment as compared to the original audio files which are approximately 16 minutes per interview. This was done to ensure

computational resources required to perform tasks on the data are reduced. This segmentation was done using python libraries Librosa and Soundfile.

3.3.3 Normalization & Resampling

Normalization is done to ensure that the input data is consistent and comparable across different audio files. Audio signals can have varying amplitudes, which can lead to difficulties in training learning models. Normalizing the audio data can help to standardize the amplitude levels, making it easier for the algorithm to better extract meaningful features and patterns from the data. The audio data was resampled to a standard rate of 20KHz. These were be done using python libraries Librosa and Soundfile.

3.3.4 Feature Selection and Extraction

This study is focused on extracting prosodic acoustic features. Prosodic features refer to various acoustic characteristics of speech such as pitch, tone, rhythm, stress, voice quality, which a listener can generally perceive. CNNs typically require a visual image as input. To represent speech stimuli, Mel spectrograms are utilized in this study as it is evident that CNN models using Mel spectrograms outperform the models using MFCC [91]. A spectrogram provides a visual representation of sound by showing the amplitude of frequency components of a signal over time. What distinguishes lower-level feature representations like MFCCs from spectrograms is that spectrograms maintain a high level of detail. While a spectrogram displays the distribution over time and frequency, a Mel spectrogram considers the fact that human hearing is less sensitive to changes in the high-frequency range than the low-frequency range. This is done by scaling the frequency axis using a Mel scale, which is a nonlinear transformation of the frequency scale, which makes it a more accurate representation of the way humans perceive sound. Overall, Mel Spectrogram, MFCC and Chroma features are extracted in this study. An example of a spectrogram from a 10 second segment and an audio file utilized in this study can be seen in figure 3.3 below. The x-axis represents time, and the y-axis represents frequency. The intensity or colour of each point in the spectrogram indicates the strength or amplitude of the frequency component at a specific time. Darker areas usually represent lower amplitudes, while brighter areas represent higher amplitudes.

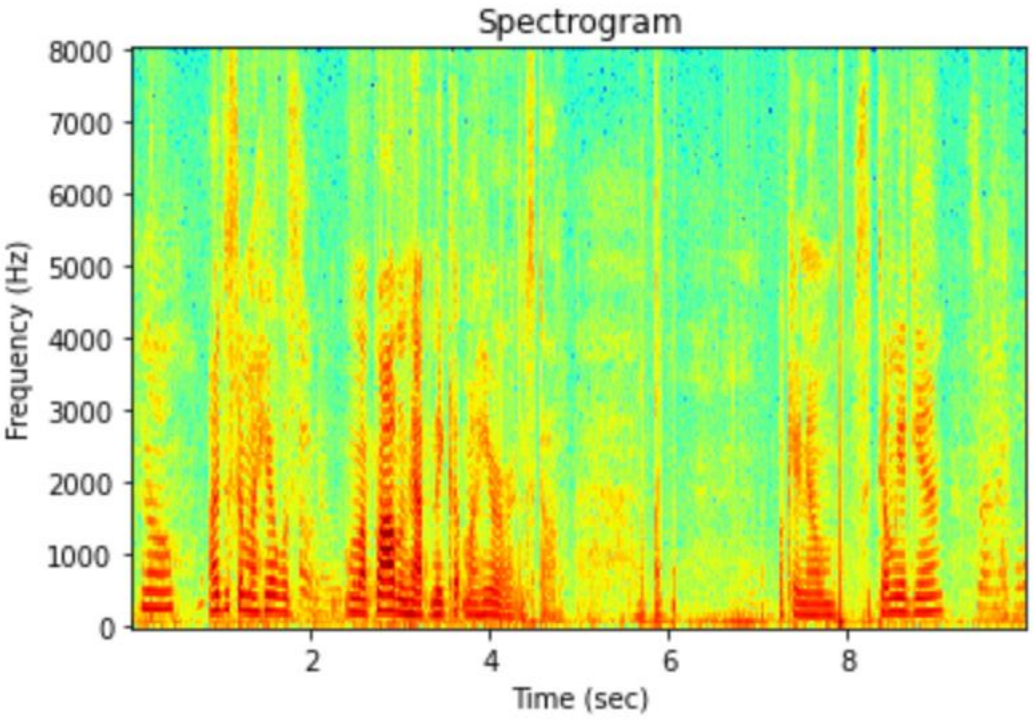

Figure 3.3: Spectrogram of one of the depressed segments

3.3.5 Data Splitting

The dataset was split into 2 sets; training, testing and development/validation set with a ratio of 70:30 respectively.

3.4 Training

The labelling criteria adhered to a class of depressed (1) or not depressed (0) and +ptsd (1) or -ptsd (0). The model was trained for 50 epochs using this data set.

3.5 Model

The model was based off the CNN algorithm, which included a transformer encoder, three hidden layers and one output layer. It was implemented using the Keras Sequential model as seen in figure 3.4 below.

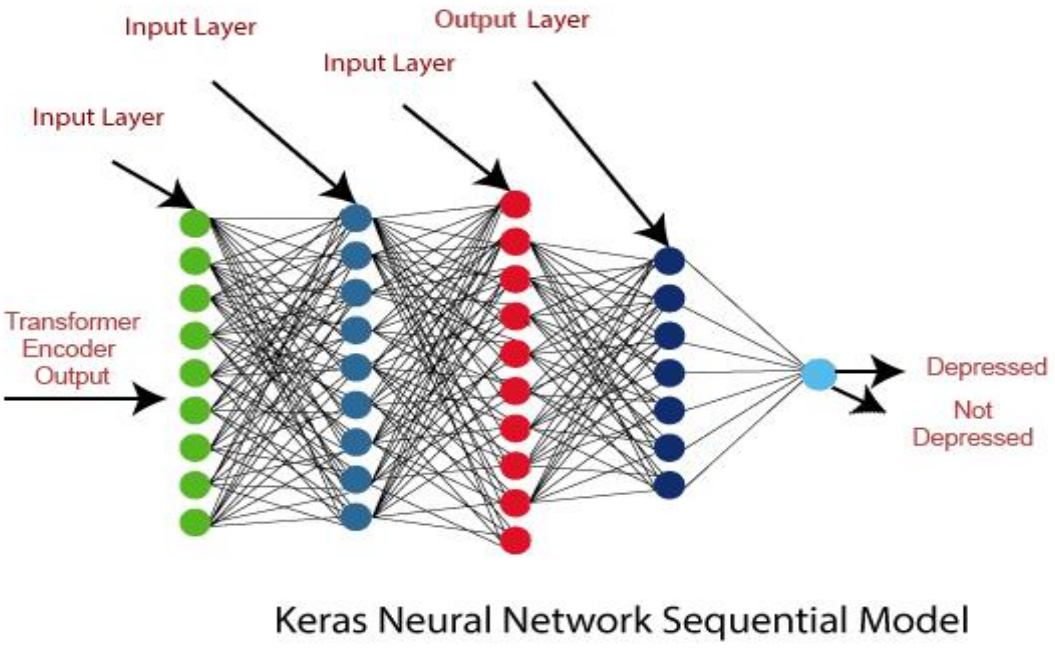

Figure 3.4 Keras Sequential Model

A Transformer Encoder layer with 2 layers, 2 attention heads, and a 128-dimensional model size was used to encode the input sequence. The output of the Transformer Encoder layer is then flattened and passed through three dense layers with the rectified linear unit (ReLU) activations and dropout layers. ReLU is a non-linear function that is commonly used in artificial neural networks and has several advantages over other activation functions, such as the sigmoid and tanh functions. It is faster to train, it is more robust to overfitting, and it can learn more complex patterns in the data.

The first hidden layer has 100 neurons, and it uses the ReLU activation function. ReLU is a common activation function in deep learning because it is computationally efficient and can help to avoid the vanishing gradient problem. This layer also includes a dropout layer with a dropout rate of 0.5. Dropout is a regularization technique that randomly drops out (sets to zero) a fraction of the neurons during training. This can help to prevent overfitting and improve the generalization performance of the model.

The second hidden layer has 200 neurons, and it also uses the ReLU activation function. It also includes a dropout layer with a dropout rate of 0.5.

The third hidden layer has 100 neurons and uses the ReLU activation function. It also includes a dropout layer with a dropout rate of 0.5.

The output layer uses the ReLU activation function.

4.0 Results

From the first to 50^{th} epoch, there was an improvement of validation loss from 0.65317 in the first epoch to 0.63103 in the 50^{th} epoch which is desirable since the goal of the training process is to minimize the loss function or report its decrease over time.

For the testing phase, a separate set of 560 spectrograms from 14 participants (40 spectrograms per participant, total of 160 seconds of audio) was used to evaluate the model's performance. Initially, the model made predictions on each 10-second Mel spectrogram to determine if depression or PTSD can be detected from short audio segments. Subsequently, the majority vote of the 40 spectrogram predictions per participant was utilized to classify the participant as either depressed or not depressed.

The tables below show how both depression and PTSD models performed based on Accuracy, Kappa Statistic and AUC metrics for both training and testing sets.

Table 4.1 Performance of both models based on training data

	Accuracy	Kappa Statistic	AUC
Depression	0.955	0.91	0.95
PTSD	0.94	0.87	0.94

Table 4.2 Performance of both models based on testing data

	Accuracy	Kappa Statistic	AUC
Depression	0.912	0.85	0.93
PTSD	0.896	0.79	0.88

Figures 4.1 and 4.2 below shows the Area Under the ROC Curve (AUC-ROC) curves of both models. The AUC-ROC score measures the performance of a model by examining the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity).

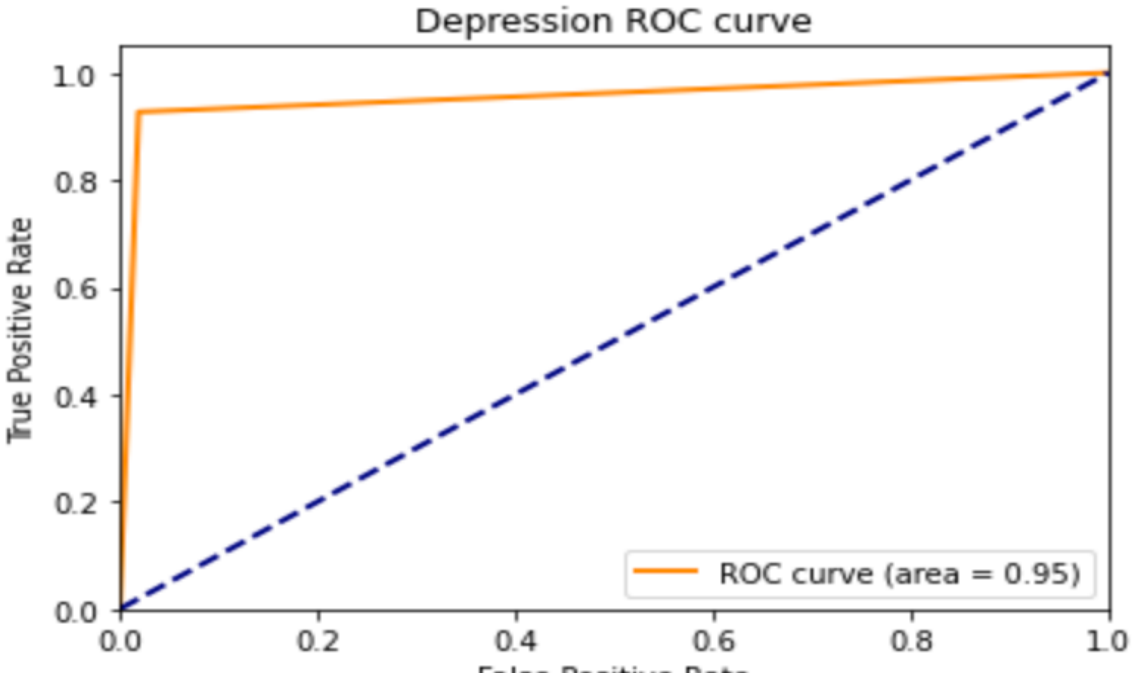

Figure 4.1 ROC Curve for the Depression Model

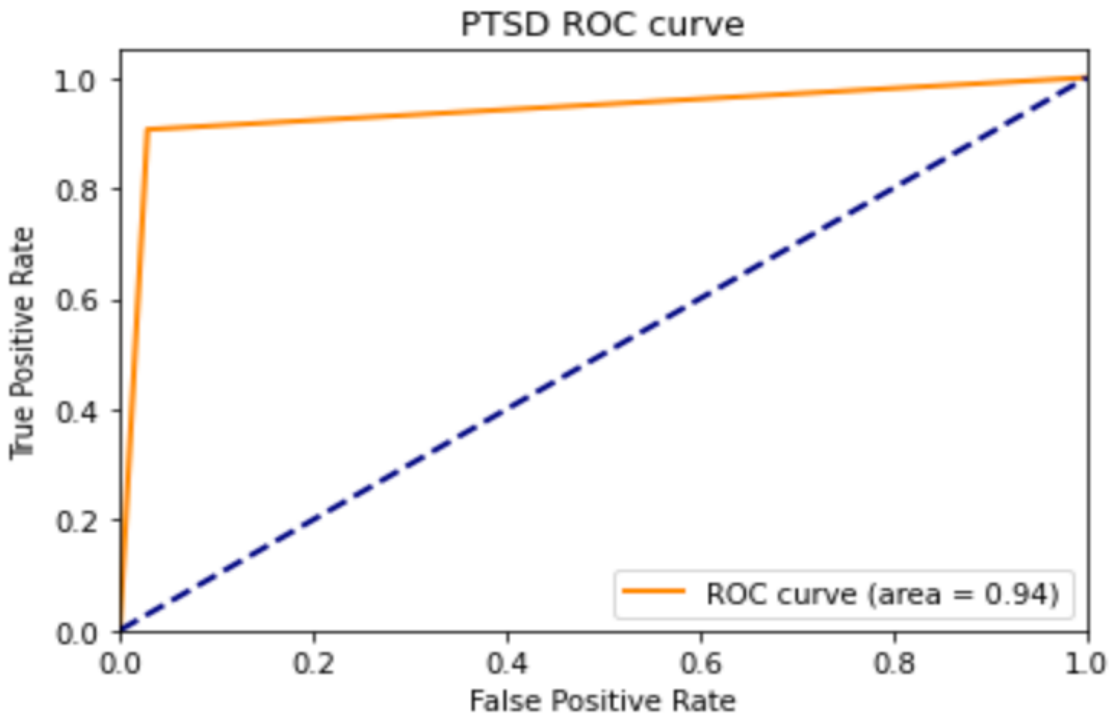

Figure 4.2 ROC Curve for the PTSD Model

Looking carefully at the results, an accuracy of 0.955 suggests that the model was able to correctly classify 95.5% of the cases. This is a high accuracy rate and indicates that the model is performing well in distinguishing between depression and non-depression cases.

Cohen's kappa measures the level of agreement between the model's predictions and the true labels, considering the possibility of random agreement. A value of 0.91 indicates a high level of agreement between the model's predictions and the true labels, which means that the model is making predictions that are consistent with the actual outcomes, however on the test set, a Kappa score of 0.79 is noted for PTSD model, this is not a perfect agreement, and it indicates a substantial agreement of the correctness of the prediction. Overall, these performance results are a good indication of both models' reliability and robustness.

Carefully examining these results, an AUC of 0.95 on the training set and 0.93 suggests that the model has a high true positive rate (sensitivity) and a low false positive rate (1 - specificity), which means that the model is effective in detecting depression cases while minimizing the number of false positives. This is important in a clinical setting where false positives can lead to unnecessary interventions and treatments. It also performed better than the adopted model, which reported an AUC score of 0.9457 on the train set and 0.9290 on the test set.

5.0 Conclusion

The analysis of speech signals for the detection of mental illness has shown promising results. The use of deep learning techniques and deep learning algorithms, such as convolutional neural networks and transformers, have enabled accurate classification of depression and PTSD based on speech patterns. Importantly, this research contributes to the field of mental health by offering a non-invasive and objective approach to detect mental disorders, highlighting features important for optimal performance. Mel Spectrograms have been found to be effective in representing speech stimuli, and while there is still a long way to go before this technique can be used in practical settings, the potential

benefits for early detection and intervention of mental disorders make this a promising area for future research.

6.0 Recommendation for future work

There are several exciting avenues for future research emerge. Firstly, it is imperative to focus on the ethical implications surrounding the use of AI in mental health diagnosis, ensuring patient privacy, consent, and data security. Secondly, expanding the scope of this technology beyond mental health, such as applying similar speech recognition models in the field of education for identifying learning disabilities or language disorders, could significantly benefit diverse populations. Moreover, exploring real-time applications in human-computer interaction, such as voice-controlled virtual assistants tailored for individuals with cognitive impairments, holds immense potential.

References

[1] Y. Li, Y. Lin, H. Ding, and C. Li, "Speech databases for mental disorders: A systematic review," *Gen. Psychiatry*, vol. 32, no. 3, p. e100022, Jul. 2019, doi: 10.1136/gpsych-2018-100022.

[2] "Mental disorders." https://www.who.int/news-room/fact-sheets/detail/mental-disorders (accessed Sep. 04, 2022).

[3] C. Hjorthøj, A. E. Stürup, J. McGrath, and M. Nordentoft, "SA57. Life Expectancy and Years of Potential Life Lost in Schizophrenia: A Systematic Review and Meta-Analysis," *Schizophr. Bull.*, vol. 43, no. suppl_1, pp. S133–S134, Mar. 2017, doi: 10.1093/schbul/sbx023.056.

[4] P. Fusar-Poli, C. U. Correll, C. Arango, M. Berk, V. Patel, and J. P. A. Ioannidis, "Preventive psychiatry: a blueprint for improving the mental health of young people," *World Psychiatry*, vol. 20, no. 2, pp. 200–221, Jun. 2021, doi: 10.1002/wps.20869.

[5] M. J. Millan *et al.*, "Altering the course of schizophrenia: progress and perspectives," *Nat. Rev. Drug Discov.*, vol. 15, no. 7, pp. 485–515, Jul. 2016, doi: 10.1038/nrd.2016.28.

[6] G. S. de Pablo *et al.*, "P.0178 Transition to psychosis in adolescents at clinical high risk for psychosis: a meta-analysis," *Eur. Neuropsychopharmacol.*, vol. 53, pp. S129–S130, Dec. 2021, doi: 10.1016/j.euroneuro.2021.10.172.

[7] C. Davies *et al.*, "Lack of evidence to favor specific preventive interventions in psychosis: a network meta-analysis," *World Psychiatry*, vol. 17, no. 2, pp. 196–209, Jun. 2018, doi: 10.1002/wps.20526.

[8] S. Trautmann, J. Rehm, and H. Wittchen, "The economic costs of mental disorders," *EMBO Rep.*, vol. 17, no. 9, pp. 1245–1249, Sep. 2016, doi: 10.15252/embr.201642951.

[9] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investig. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, Feb. 2020, doi: 10.1002/lio2.354.

[10] C. Wanderley Espinola, J. C. Gomes, J. Mônica Silva Pereira, and W. P. dos Santos, "Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study," *Res. Biomed. Eng.*, vol. 38, no. 3, pp. 813–829, Jun. 2022, doi: 10.1007/s42600-022-00222-2.

[11] A. J. Baxter, G. Patton, K. M. Scott, L. Degenhardt, and H. A. Whiteford, "Global Epidemiology of Mental Disorders: What Are We Missing?," *PLoS One*, vol. 8, no. 6, p. e65514, Jun. 2013, doi: 10.1371/journal.pone.0065514.

[12] D. C. Mohr *et al.*, "Perceived barriers to psychological treatments and their relationship to depression," *J. Clin. Psychol.*, p. n/a-n/a, 2010, doi: 10.1002/jclp.20659.

[13] R. J. Turner, D. A. Lloyd, and J. Taylor, "Physical disability and mental health: An epidemiology of psychiatric and substance disorders.," *Rehabil. Psychol.*, vol. 51, no. 3, pp. 214–223, 2006, doi: 10.1037/0090-5550.51.3.214.

[14] A. Shalev, I. Liberzon, and C. Marmar, "Post-Traumatic Stress Disorder," *N. Engl. J. Med.*, vol. 376, no. 25, pp. 2459–2469, Jun. 2017, doi: 10.1056/NEJMra1612499.

[15] A. L. Rathbone, L. Clarry, and J. Prescott, "Assessing the Efficacy of Mobile Health Apps Using the Basic Principles of Cognitive Behavioral Therapy: Systematic Review," *J. Med. Internet Res.*, vol. 19, no. 11, p. e399, Nov. 2017, doi: 10.2196/jmir.8598.

[16] A. M. Chekroud *et al.*, "Cross-trial prediction of treatment outcome in depression: a machine learning approach," *The Lancet Psychiatry*, vol. 3, no. 3, pp. 243–250, Mar. 2016, doi: 10.1016/S2215-0366(15)00471-X.

[17] D. A. Lewis, R. Michels, D. S. Pine, S. K. Schultz, C. A. Tamminga, and R. Freedman, "Conflict of Interest," *Am. J. Psychiatry*, vol. 163, no. 4, pp. 571–573, Apr. 2006, doi: 10.1176/ajp.2006.163.4.571.

[18] D. A. Regier *et al.*, "Response to Hasin et al. Letter," *Am. J. Psychiatry*, vol. 170, no. 4, pp. 443–444, Apr. 2013, doi: 10.1176/appi.ajp.2013.13010032r.

[19] J. Gideon, H. T. Schatten, M. G. McInnis, and E. M. Provost, "Emotion Recognition from Natural Phone

Conversations in Individuals with and without Recent Suicidal Ideation," in *Interspeech 2019*, ISCA: ISCA, Sep. 2019, pp. 3282–3286. doi: 10.21437/Interspeech.2019-1830.

[20] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015, doi: 10.1016/j.specom.2015.03.004.

[21] P. W. Koh and P. Liang, "Understanding Black-box Predictions via Influence Functions." PMLR, pp. 1885–1894, Jul. 17, 2017. Accessed: Jun. 04, 2023. [Online]. Available: https://proceedings.mlr.press/v70/koh17a.html

[22] D. S. Char, N. H. Shah, and D. Magnus, "Implementing Machine Learning in Health Care — Addressing Ethical Challenges," *N. Engl. J. Med.*, vol. 378, no. 11, pp. 981–983, Mar. 2018, doi: 10.1056/NEJMp1714229.

[23] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018, doi: 10.1109/ACCESS.2018.2807385.

[24] J. M. Hamilton, "Review: *The Speech Chain* , by Peter B. Denes and Elliot N. Pinson," *Am. Biol. Teach.*, vol. 27, no. 4, pp. 291–292, Apr. 1965, doi: 10.2307/4440952.

[25] E. Kraepelin, "Manic Depressive Insanity and Paranoia," *J. Nerv. Ment. Dis.*, vol. 53, no. 4, p. 350, Apr. 1921, doi: 10.1097/00005053-192104000-00057.

[26] M. B. Hoy, "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants," *Med. Ref. Serv. Q.*, vol. 37, no. 1, pp. 81–88, Jan. 2018, doi: 10.1080/02763869.2018.1404391.

[27] I. Schoon, S. Parsons, R. Rush, and J. Law, "Children's Language Ability and Psychosocial Development: A 29-Year Follow-up Study," *Pediatrics*, vol. 126, no. 1, pp. e73–e80, Jul. 2010, doi: 10.1542/peds.2009-3282.

[28] E. Arkkila, P. Räsänen, R. P. Roine, H. Sintonen, V. Saar, and E. Vilkman, "Health-related quality of life of adolescents with childhood diagnosis of specific language impairment," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 73, no. 9, pp. 1288–1296, Sep. 2009, doi: 10.1016/j.ijporl.2009.05.023.

[29] J. Law, R. Rush, I. Schoon, and S. Parsons, "Modeling Developmental Language Difficulties From School Entry Into Adulthood: Literacy, Mental Health, and Employment Outcomes," *J. Speech, Lang. Hear. Res.*, vol. 52, no. 6, pp. 1401–1416, Dec. 2009, doi: 10.1044/1092-4388(2009/08-0142).

[30] A. J. O. Whitehouse, M. Robinson, and S. R. Zubrick, "Late Talking and the Risk for Psychosocial Problems During Childhood and Adolescence," *Pediatrics*, vol. 128, no. 2, pp. e324–e332, Aug. 2011, doi: 10.1542/peds.2010-2782.

[31] J. H. BEITCHMAN *et al.*, "Fourteen-Year Follow-up of Speech/Language-Impaired and Control Children: Psychiatric Outcome," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 40, no. 1, pp. 75–82, Jan. 2001, doi: 10.1097/00004583-200101000-00019.

[32] R. Armstrong *et al.*, "Change in receptive vocabulary from childhood to adulthood: associated mental health, education and employment outcomes," *Int. J. Lang. Commun. Disord.*, vol. 52, no. 5, pp. 561–572, Sep. 2017, doi: 10.1111/1460-6984.12301.

[33] J. H. Beitchman, E. B. Brownlie, and L. Bao, "Age 31 Mental Health Outcomes of Childhood Language and Speech Disorders," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 53, no. 10, pp. 1102-1110.e8, Oct. 2014, doi: 10.1016/j.jaac.2014.07.006.

[34] S. E. Mouridsen and K.-M. Hauschild, "A longitudinal study of personality disorders in individuals with and without a history of developmental language disorder," *Logop. Phoniatr. Vocology*, vol. 34, no. 3, pp. 135–141, Jan. 2009, doi: 10.1080/14015430903117441.

[35] J. H. Beitchman *et al.*, "Adolescent Substance Use Disorders: Findings From a 14-Year Follow-Up of Speech/Language-Impaired and Control Children," *J. Clin. Child Psychol.*, vol. 28, no. 3, pp. 312–321, Aug. 1999, doi: 10.1207/S15374424jccp280303.

[36] G. Conti-Ramsden and N. Botting, "Emotional health in adolescents with and without a history of specific language impairment (SLI)," *J. Child Psychol. Psychiatry*, vol. 49, no. 5, pp. 516–525, May 2008, doi: 10.1111/j.1469-7610.2007.01858.x.

[37] N. Botting and G. Conti-Ramsden, "The role of language, social cognition, and social skill in the functional social outcomes of young adolescents with and without a history of SLI," *Br. J. Dev. Psychol.*, vol. 26, no. 2, pp. 281–300, Jun. 2008, doi: 10.1348/026151007X235891.

[38] G. Lindsay, J. Dockrell, and O. Palikara, "Self-esteem of adolescents with specific language impairment as they move from compulsory education," *Int. J. Lang. Commun. Disord.*, vol. 45, no. 5, pp. 561–571, Sep. 2010, doi: 10.3109/13682820903324910.

[39] S. Goh Kok Yew and R. O'Kearney, "Early language impairments and developmental pathways of emotional problems across childhood," *Int. J. Lang. Commun. Disord.*, vol. 50, no. 3, pp. 358–373, Apr. 2015, doi: 10.1111/1460-6984.12142.

[40] K. Durkin, G. Conti-Ramsden, and Z. Simkin, "Functional Outcomes of Adolescents with a History of Specific Language Impairment (SLI) with and without Autistic Symptomatology," *J. Autism Dev. Disord.*, vol. 42, no. 1, pp. 123–138, Jan. 2012, doi: 10.1007/s10803-011-1224-y.

[41] B. A. Lewis *et al.*, "Psychosocial co-morbidities in adolescents and adults with histories of communication disorders," *J. Commun. Disord.*, vol. 61, pp. 60–70, May 2016, doi: 10.1016/j.jcomdis.2016.03.004.

[42] R. Wadman, N. Botting, K. Durkin, and G. Conti-Ramsden, "Changes in emotional health symptoms in adolescents with specific language impairment," *Int. J. Lang. Commun. Disord.*, vol. 46, no. 6, pp. 641–656, Nov. 2011, doi: 10.1111/j.1460-6984.2011.00033.x.

[43] M. C. St Clair, A. Pickles, K. Durkin, and G. Conti-Ramsden, "A longitudinal study of behavioral, emotional and social difficulties in individuals with a history of specific language impairment (SLI)," *J. Commun. Disord.*, vol. 44, no. 2, pp. 186–199, Mar. 2011, doi: 10.1016/j.jcomdis.2010.09.004.

[44] M. Yamamoto *et al.*, "Using speech recognition technology to investigate the association between timing-related speech features and depression severity," *PLoS One*, vol. 15, no. 9, p. e0238726, Sep. 2020, doi: 10.1371/journal.pone.0238726.

[45] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain Cogn.*, vol. 56, no. 1, pp. 30–35, Oct. 2004, doi: 10.1016/j.bandc.2004.05.003.

[46] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biol. Psychiatry*, vol. 72, no. 7, pp. 580–587, Oct. 2012, doi: 10.1016/j.biopsych.2012.03.015.

[47] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolinguistics*, vol. 20, no. 1, pp. 50–64, Jan. 2007, doi: 10.1016/j.jneuroling.2006.04.001.

[48] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting Depression Severity from Vocal Prosody," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 142–150, Apr. 2013, doi: 10.1109/T-AFFC.2012.38.

[49] C. A. Blevins, F. W. Weathers, M. T. Davis, T. K. Witte, and J. L. Domino, "The Posttraumatic Stress Disorder Checklist for *DSM-5* (PCL-5): Development and Initial Psychometric Evaluation," *J. Trauma. Stress*, vol. 28, no. 6, pp. 489–498, Dec. 2015, doi: 10.1002/jts.22059.

[50] D. Forbes, M. Creamer, and D. Biddle, "The validity of the PTSD checklist as a measure of symptomatic change in combat-related PTSD," *Behav. Res. Ther.*, vol. 39, no. 8, pp. 977–986, Aug. 2001, doi: 10.1016/S0005-7967(00)00084-X.

[51] K. Kroenke and R. L. Spitzer, "The PHQ-9: A New Depression Diagnostic and Severity Measure," *Psychiatr. Ann.*, vol. 32, no. 9, pp. 509–515, Sep. 2002, doi: 10.3928/0048-5713-20020901-06.

[52] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[53] Jeffrey M. Cohen, Dustin Kieschnick, Christine M Blasey, and Nitya Kanuri, "(PDF) Preliminary Evaluation of the Psychometric Properties of the PTSD Checklist for DSM – 5." https://www.researchgate.net/publication/270905350_Preliminary_Evaluation_of_the_Psychometric_Properties_of_the_PTSD_Checklist_for_DSM_-_5 (accessed Apr. 04, 2023).

[54] A. Vaswani *et al.*, "Attention Is All You Need."

[55] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews." pp. 3123–3128, 2014. Accessed: Jun. 04, 2023. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

[56] "GitHub - AmirHoseein99/Depression-Engine: Detecting depressed Patient based on Speech Activity, Pauses in Speech and Using Deep learning Approach." https://github.com/AmirHoseein99/Depression-Engine (accessed Mar. 09, 2023).

[57] "DAIC---WOZ Depression Database", Accessed: Aug. 2, 2023. [Online]. Available: https://dcapswoz.ict.usc.edu/wp-content/uploads/2022/02/DAICWOZDepression_Documentation.pdf