Acoustic and Text Features Analysis for Adult ADHD Screening: A Data-Driven Approach Utilizing DIVA Interview

Shuanglin Li[1], *Student Member, IEEE,* Rajesh Nair[2], Mohsen Naqvi[1], *Senior Member, IEEE*

***Abstract*—Objective:** Attention Deficit Hyperactivity Disorder (ADHD) is a neurodevelopmental disorder commonly seen in childhood that leads to behavioural changes in social development and communication patterns, often continues into undiagnosed adulthood due to a global shortage of psychiatrists, resulting in delayed diagnoses with lasting consequences on individual's well-being and the societal impact. Recently, machine learning methodologies have been incorporated into healthcare systems to facilitate the diagnosis and enhance the potential prediction of treatment outcomes for mental health conditions. In ADHD detection, the previous research focused on utilizing functional magnetic resonance imaging (fMRI) or Electroencephalography (EEG) signals, which require costly equipment and trained personnel for data collection. In recent years, speech and text modalities have garnered increasing attention due to their cost-effectiveness and non-wearable sensing in data collection. In this research, conducted in collaboration with the Cumbria, Northumberland, Tyne and Wear NHS Foundation Trust, we gathered audio data from both ADHD patients and normal controls based on the clinically popular *Diagnostic Interview for ADHD in adults (DIVA)*. Subsequently, we transformed the speech data into text modalities through the utilization of the Google Cloud Speech API. We extracted both acoustic and text features from the data, encompassing traditional acoustic features (e.g., MFCC), specialized feature sets (e.g., eGeMAPS), as well as deep-learned linguistic and semantic features derived from pre-trained deep learning models. These features are employed in conjunction with a support vector machine for ADHD classification, yielding promising outcomes in the utilization of audio and text data for effective adult ADHD screening.
Clinical impact: This research introduces a transformative approach in ADHD diagnosis, employing speech and text analysis to facilitate early and more accessible detection, particularly beneficial in areas with limited psychiatric resources.
Clinical and Translational Impact Statement: The successful application of machine learning techniques in analyzing audio and text data for ADHD screening represents a significant advancement in mental health diagnostics, paving the way for its integration into clinical settings and potentially improving patient outcomes on a broader scale.

***Keywords*—**Adults ADHD, Speech Modality, Text Modality, Feature Study, Machine learning.

I. Introduction

Mental illness is a kind of health condition that gives rise to shifts in an individual's cognitive functions, emotional reactions, and behavioural tendencies, with empirical evidence underscoring its capacity to exert an impact on the physical welfare of the person. Mental health conditions including depression, schizophrenia, Alzheimer's disease, autism spectrum disorder (ASD), etc., are currently widespread, with an estimated global prevalence of 450 million individuals affected [1]. In addition to these widely publicized mental conditions, Attention Deficit Hyperactivity Disorder (ADHD) has been gaining attention over the years. For example, adults with ADHD are more likely to procrastinate in activities of daily living and cognition processes [2]. ADHD is a mental health condition that has a detrimental impact on the neurodevelopment of the brain and can result in profound impairments in cognitive and social functioning. There is a study suggesting that ADHD patients have a higher divorce rate, criminal conduct, arrests, convictions, imprisonment and decreased life span [3]. Currently, ADHD diagnosis relies on clinical assessments by a specialist, such as a psychiatrist or a paediatrician, to determine if an individual meets DSM-V [4] criteria by displaying five or more relevant symptoms of inattention or impulsivity/hyperactivity. However, the diagnosis of ADHD is often delayed due to the global shortage of specialists in the related areas. In the UK, the ratio of psychiatrists to the general population is 11:100,000 [5] and the British Broadcasting Company (BBC) reported that around 1.5 million adults in the UK have ADHD, but only 120,000 are officially diagnosed and the waiting period for a diagnosis of ADHD is a maximum of 7 years [6].

Moreover, adults with ADHD may exhibit symptoms like irritability, emotional instability, heightened activity, thoughtfulness, paranoia, and worry, which overlap with DSM-V [7] criteria for anxiety and bipolar disorder, adding complexity to the diagnostic process. Additionally, the subjective nature of ADHD criteria and rating scales may contribute to its perceived increasing prevalence. A recent study also suggested that adult ADHD diagnosis is relatively neglected as ADHD affects approximately 7.2% of children and has a lower prevalence rate of 2.58% in adults globally [8], [9]. These elements emphasize the importance of screening ADHD in adults, aiming to mitigate its negative effects on individual well-being and communities.

In contrast to most chronic conditions diagnosed through lab tests, mental disorders rely on self-disclosure through specialized methods, highlighting the intricate nature of mental health data. In the past decades, machine learning (ML) and deep learning (DL) have provided a new paradigm for gaining knowledge from complex data [10], [11] and numerous ML and DL-based techniques have been developed for healthcare applications including mental healthcare and achieved

[1]Intelligent Sensing and Communications Group, School of Engineering, Newcastle University, NE1 7RU, Newcastle Upon Tyne, U.K
[2]Adult ADHD Services, Cumbria, Northumberland, Tyne and Wear, NHS Foundation Trust, NE3 3XT, Newcastle Upon Tyne, U.k

considerable success [12]–[16]. Specifically, studies are using adjunctive data for the detection of ADHD, including magnetic resonance imaging (MRI) [17], electroencephalography (EEG) [18] and electrocardiograms (ECG) [19]. However, such psychological signals are expensive to acquire as they require experienced radiologists and special equipment such as an MRI scanner and wearable EEG headset [20]. Recently, audio and text modalities have drawn more attention for mental health detection due to the characteristics of obtainable, non-invasive and containing abundant information [21], [22]. Different from the psychological signal, human speech conveys not only verbal (linguistic) content like words but also non-verbal (paralinguistic) information, such as speech tone, which is responsive to subtle shifts in the speaker's physiological condition and mental state [23]. For example, adults with ADHD have subtle differences in speech production, increased speech rate, and alternating and sequential motion rates compared to non-ADHD control participants. Moreover, as suggested in DSM-V [7], ADHD patients have speech symptoms such as speaking loud and fast and talking excessively and tangentially [7].

Nevertheless, barriers and limitations persist in applying ML and DL to aid in mental health detection, encompassing challenges such as data scarcity, absence of personality factors, and the extraction of meaningful features from diverse data sources [24]. Motivated by the challenge of ADHD detection and the success of audio and text-based health-conditions detection, we aim to carry out discriminative features in audio and text modalities for ADHD detection. In this research, in collaboration with Cumbria, Northumberland, Tyne and Wear National Health Service (CNTW-NHS) Foundation Trust, we trial a novel multi-modal ADHD dataset from 22 participants (10 ADHD patients and 12 healthy controls). Particularly in the proposed work, we focus on the speech and text modalities. As far as we know, there is an absence of comprehensive prior investigations into audio and text-based ADHD identification. Based on the recorded dataset, we briefly investigate the speech and text features for ADHD detection based on support vector machines (SVM) and logistic regression (LR) and their correlation with ADHD and evidence that speech and text can be used for ADHD early detection.

The main contributions of the proposed work are summarized as follows:

1) The study introduces a novel approach by exploring the feasibility of using speech and text features for early ADHD detection. This approach aims to enable timely diagnosis and intervention, addressing the challenge of delayed ADHD diagnosis.

2) The research conducts a comprehensive analysis of various speech and text features to identify the most predictive ones for discriminating ADHD. This analysis enhances our understanding of the disorder's manifestations in these modalities, shedding light on the critical information for ADHD detection.

3) Employing machine learning techniques, particularly SVM and LR, the study provides a practical and data-driven method for ADHD detection. This contribution aligns with the broader field of ML-based healthcare applications and signifies the potential for automated ADHD screening using audio and text data.

The rest of the paper is organized as follows. Section 2 is the related work. Then the recorded multi-modal ADHD dataset and the selected features are introduced in Section 3. The experimental settings and results are discussed in Section 4. Section 5 draws the discussions and conclusions.

II. Related Work

In this section, we delve into contemporary research on mental health detection, focusing on three aspects: ADHD detection utilizing machine learning techniques, and mental health detection through audio and text modalities. It's noteworthy to emphasize that the intersection of audio and text-based ADHD detection remains scarcely explored in existing literature, leading to a limited number of studies available for review in this specific area.

A. ADHD Detection Based on Machine Learning

The existing methods for ADHD detection are mainly based on neuroimaging data and physiological data such as fMRI and EEG signals. The Neuro Bureau ADHD-200 Preprocessed repository (ADHD-200) [25] is the most widely used brain MRI database for automated ADHD detection. The ADHD-200 dataset has more controls than ADHD patients as the challenge's goal is to identify healthy controls, however, in real applications, correctly diagnosing the patient is as important as correctly diagnosing the normal person. Koh et al. use ensembled ML classifiers with entropy features extracted from ECG signals and classify ADHD with a high accuracy of 87.2% [19]. Boroujeni et al. combine the non-linear features from EEG signals to detect ADHD and achieve an overall accuracy of 96.05% [18].

Although those methods have achieved considerable results, there are some limitations. Firstly, it's worth mentioning that both neuroimaging data and EEG measurements are typically confined to hospital settings or controlled laboratories, which necessitates the presence of specialized staff. Furthermore, the collection of data from individuals diagnosed with ADHD is a resource-intensive endeavour, characterized by considerable time and financial investments [26]. Moreover, most of the physiological data requires data preprocessing to reduce the dimensionality or remove the noise and clean the data.

B. Audio Based Mental Health Detection

Speech patterns serve as recognized indicators of mental disorders, and within the field of speech-based mental disorder detection, the quest for dependable acoustic biomarkers has persistently stood as a primary research focus [27], [28]. Existing work can mainly be divided into conventional hand-crafted features and deep-learned features, respectively. Certain correlations exist between conventional hand-crafted features and mental health conditions. A review pointed out that depressive voices have paralinguistic biomarkers such as fundamental frequency variability or jitter, shimmer [29]. Lopez et al. [30] use Mel Frequency Cepstral Coefficients (MFCC) [31] and short-time energy in depression detection based on the Gaussian mixture model and support vector

regression. Nasreen and Rohanian et al. [32], [33] designed AD-specific characteristics such as pause time, and disfluencies for AD detection. Furthermore, various composite low-level descriptors (LLD) feature sets have been employed for mental health detection, e.g. the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [34], IS10˙paraling [35], and ComParE˙2016 [36]. Fasih et al. [37] use the eGeMAPS feature for Alzheimer's Dementia based on spontaneous speech and achieved an accuracy of 71.34%. However, conventional hand-crafted acoustic features cover voice quality but may not capture task-specific symptoms in particular, prompting researchers to extract lengthy feature lists with sub-optimal results. Meanwhile, specially designed features require translating medical knowledge into mathematical expressions, which can be challenging. To overcome such limitations, the utilization of pre-trained language models or ASR (Automatic Speech Recognition) models for the extraction of deep features has become a prevailing practice in the field. Qin et al. fine-tuned Wav2Vec2.0 [38] for speech feature extraction and achieved promising results [39]. Li et al. utilize a convolution neural network as well as an information fusion method to aid in the detection of ADHD based on speech signals and yield considerable results [40], [41].

As speech disruption in ADHD adults has been reported such as disfluencies [42] and poor articulation [43], as well as the notable absence of research in ADHD speech features, it is timely to investigate the audio-based features associated in ADHD detection.

C. Text Based Mental Health Detection

Text-based mental health detection algorithms can be categorized based on their data source into two primary groups: social media posts and clinical notes from interviews. The major social media platforms include Twitter and Reddit. Sinha et al. created a dataset sourced from Twitter which includes manually annotated information to detect instances of suicidal thoughts or ideation among users [44]. Kristen et al. utilize semantic information from Twitter diary entries to predict and screen measures for depression and psychological aggression by an intimate partner [45]. Yates et al. contribute a depression dataset which includes about 9k depressed users and 100k control users [46]. Certain studies have explored the detection of mental illness by conducting interviews and subsequently analyzing linguistic information extracted from transcribed clinical interviews, e.g., Michelle et al. transcribe the audio data into text and provide a comparative analysis of the syntactic structure and semantic content for depression detection [47].

Recently, deep pre-trained language models based on Transformer such as BERT [48] have gained attention for their remarkable effectiveness in learning subtle and complex lexical patterns that are used in the detection of psychiatric disorders based on text modalities. Balagopalan et al [49]. Luz et al. transform speech into text [50], and then use the BERT model for Alzheimer's Disease recognition and show discriminative results; However, there's no ADHD-related text database and therefore no one used text modalities to aid in the detection of ADHD. Here, we intend to investigate whether syntactic and semantic information or text data obtained from ADHD clinical interviews or deep linguistic features extracted by fine-tuned BERT can aid in the detection of ADHD.

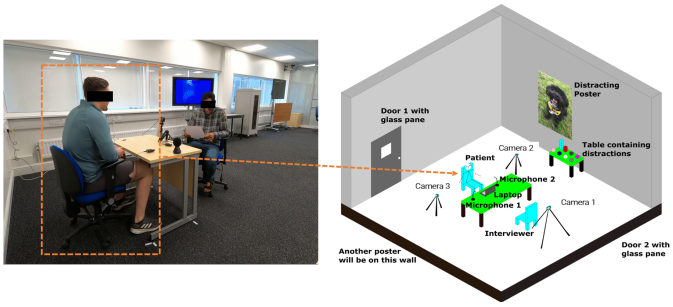


Fig. 1. The intended experimental setting for the dataset recording.

TABLE I
PREPARED QUESTIONNAIRE FROM DIVA

Q1)	How are your focus and concentration?
Q2)	How do you find your thoughts?
Q3)	Do you find yourself switching off and zoning out?
Q4)	Do you leave things to the last minute?
Q5)	Do you make careless mistakes?
Q6)	Are you normally late for an appointment?
Q7)	Do you get bored easily?
Q8)	Do you find it difficult to start a task?
Q9)	Are you forgetful?
Q10)	Do you lose things or misplace things?
Q11)	Are you able to sit still in situations like lectures or meetings?
Q12)	Do you tend to fidget or pace around?
Q13)	How are you in situations like queues and traffic jams?
Q14)	Do you tend to speak loud and fast in daily life?
Q15)	Do you finish sentences for other people?
Q16)	How is your mood?
Q17)	Does your mood tend to change?
Q18)	Do you have a short temper?
Q19)	Do you spend excessively?
Q20)	Are you easily led?
Q21)	Do you make quick and rush decisions?

III. METHODOLOGY

A. Multimodal ADHD Dataset

In our Intelligent Sensing ADHD Trial (ISAT) [51], we create an innovative multimodal dataset comprising audio, video, historical data/questionnaires, transcribed text data, Cambridge Neuropsychological Test Automated Battery (CANTAB) [52] test and keyboard tracking. The recording took place in a quiet, clean, and no electromagnetic room (similar to a consultation room in a hospital) and the experimental environment facilities are arranged as shown in Fig.1. A total of 22 native English speakers, consisting of 10 subjects diagnosed with ADHD (5 males and 5 females) and 12 control participants (8 males and 4 females), were recruited for the study. All the ADHD participants have NHS-certified diagnoses from the Cumbria, Northumberland, Tyne and Wear NHS Foundation Trust, a prominent mental health NHS Trust. Their ADHD condition was confirmed using the ASRSv1.1 [53] symptom checklist. All the health controls are the volunteers recruited through posters on university campuses. Currently, under the medical data protection policy and ethics issues, the dataset can be accessed on request.

TABLE II
DEMOGRAPHICS OF ALL PARTICIPANTS

ID	Diagnosis	Age	Gender	Interview Time[min:sec]
P3	ADHD	35	male	17:50
P7	ADHD	28	male	21:25
P8	ADHD	25	male	4:45
P10	ADHD	32	female	10:38
P11	ADHD	29	female	11:44
P14	ADHD	18	male	6:50
P15	ADHD	53	female	8:20
P19	ADHD	25	female	12:07
P20	ADHD	25	female	7:34
P24	ADHD	26	male	11:28
P2	Control	25	male	9:50
P4	Control	26	male	8:17
P5	Control	28	male	7:45
P6	Control	28	male	8:20
P9	Control	25	male	4:08
P12	Control	25	male	5:30
P16	Control	27	male	3:50
P17	Control	24	female	5:45
P18	Control	25	male	4:50
P21	Control	25	female	4:04
P22	Control	38	female	5:22
P23	Control	28	female	3:24

The data recording process is divided into four parts: 1) The DIVA-based interview task; 2) The performance of CANTAB [52] task; 3) A beep reaction task; 4) Watching normal and boring video tasks; In this research, we only focus on the interview part as the audio and text modality. The *Diagnostic Interview for ADHD in adults (DIVA)* [54] is the primarily used questionnaire for adult ADHD detection which is a semi-structured interview constructed based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-V)* [7]. As shown in Table I, we've chosen 21 questions from *DIVA* focusing on social interactions, hobbies, self-esteem, relationships, and work to compose the questionnaire for the interview part. During the interviews, participants are equipped with portable microphones to capture their speech patterns. Subsequently, we utilize the Google Speech-to-Text API [1] to convert each audio recording into transcript data, thus creating the text modality of our dataset. The demographics of all the participants in the interview part are shown in Table II.

B. Data Preprocessing

All audio recordings are sampled at 16kHz, with an average duration of approximately 8 minutes. However, this duration is lengthy and can compromise the efficiency of feature extraction. Additionally, there may be redundant information present beyond the interview questions, such as sudden laughter and excessively long pauses. Hence we cut all the audio recordings at the interview-question level, hence there are a total of 22x21=462 audio clips corresponding to 462 transcripts. All the audio clips underwent preprocessing, which involved the removal of stationary noise and normalization of audio volume across all audio clips, this was done to mitigate variations arising from recording conditions.

For the transcripts, we exploit the following normalised steps. Firstly, all of the common contractions have been converted to formal writing, e.g., I'll to I will. Secondly, all disfluencies are retained, non-speech phenomena are annotated as $\langle non-speech \rangle$, and punctuation marks are omitted as well. Besides, all the numbers are converted into English words. Finally, all the transcripts are lowercase.

C. Acoustic Features

Speech, a pivotal modality of human communication, bifurcates into two main processes: production and perception. Within production, the intricate anatomy of the vocal tract plays a pivotal role by facilitating various articulatory movements. These movements not only give rise to words and phrases but also generate a diverse range of vocal behaviours. Specifically, vocal behaviour comprises voice and paralinguistic elements encompassing all aspects of speech beyond the core verbal content. Conversely, in the perception phase, as acoustic patterns are intercepted by the auditory apparatus, the acoustic meatus plays a central role in distilling phonetic features, enabling both auditory recognition and deeper cognitive interpretation of the conveyed content. Hence we are motivated to investigate the features to model speech production, perception and as well as paralinguistic pieces of ADHD acoustic information.

All the acoustic features are extracted using the Python library Librosa [55] and openSMILE [56]. All these features were then normalized by subtracting the mean and dividing by the standard deviation of that feature over each recording and passed down to the pipeline as shown in Fig.2.

- Mel-Frequency Cepstral Coefficients (MFCCs)

  The Mel-scale filter bank is crafted to emulate the auditory and physiological aspects of how humans perceive speech signals [57]. MFCC (Mel-frequency cepstral coefficients) quantifies cepstral energies on a non-linear scale known as the Mel scale. This scale mirrors the sensitivity of the human ear, which is more sensitive to low-frequency sounds compared to high-frequency ones. The relationship between the Mel scale and frequency can be approximated by:

  $$\mathrm{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right)$$

  Where f is the frequency measured in Hertz. We extract the initial 13 MFCC bands, along with their respective 13 delta MFCCs and 13 delta-delta MFCCs. These additional features capture the rate of change and the acceleration within the MFCCs.

- Linear Predictive Coding (LPC)

  According to the source-filter model of vocal production, the energy originating from the lungs serves as the excitation source, while the vocal tract acts as the filter, processing the human voice [58]. The information present in the speech signal is shaped by the vocal tract's

[1] https://github.com/Uberi/speech_recognition

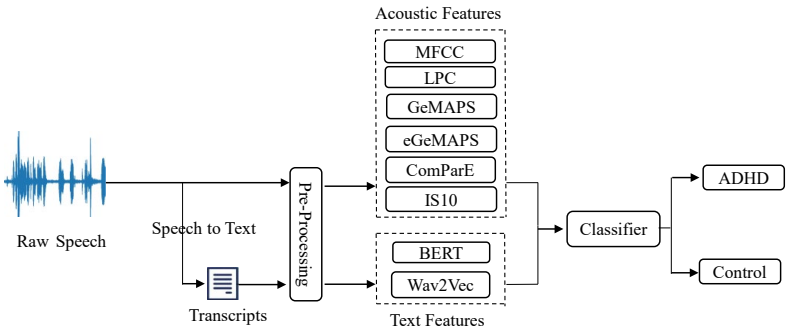


Fig. 2. The flow diagram of the ADHD feature classification process. The transcripts are generated with the speech-to-text technique, the raw speech and transcripts are preprocessed before feature extraction, and the extracted features are fed into the classifier to give the final classification results.

modulation as a dynamic filter, rather than being driven by the energy source. LPC represents the digital filter parameter designed to emulate the vocal tract, capturing the distinct characteristics of speech. The main idea of LPC [59] is that the current speech sample can be closely approximated as a linear combination of past samples, mathematically expressed as:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n)$$

where $s(n)$ is the predicted next speech sample, a_k is the p^{th} order linear predictor coefficients, $s(k)$ is the k past speech sample. Assume the real speech sample as $\widehat{s}(n)$, $e(n)$ is the prediction error calculated as:

$$e_n = \widehat{s}(n) - s(n) = \widehat{s}(n) - \sum_{k=1}^{p} a_k s(n-k)$$

By minimizing the mean square error of $e(n)$, the filter coefficients $a(k)$ can be updated.

- The Geneva Minimalistic Acoustic Parameter Set (GeMAPS)

  GeMAPS [34] is configured with features designed specifically for affective speech tasks, aiming to encompass and encapsulate the speaker's overall characteristics. GeMAPS consists of 18 low-level descriptors (LLD). These include frequency-related parameters such as pitch, jitter, and formants; energy-associated parameters like shimmer, loudness, and the harmonics-to-noise ratio (HNR); along with various spectral metrics like the Alpha ratio, Hammarberg Index, spectral slopes, harmonic differences, and additional formants. Each LLD is subjected to statistical functions, resulting in a total of 62 parameters per 100 frames.

- The extended Geneva Minimalistic Acoustic Parameter (eGeMAPS)

  The eGeMAPS feature set [34] emerged from efforts to streamline the extensive feature into a foundational collection of acoustic attributes. These were chosen based on their capacity to identify physiological variations in voice production, their theoretical importance, and their established utility in prior related research [60], [61]. This set encompasses the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, formant 1, 2 and 3 frequency, alpha ratio, Hammarberg index, and slope V0 attributes, combined with their prevalent statistical functionals, culminating in 88 features for every 100ms frame.

- INTERSPEECH 2010 Paralinguistics Challenge Feature Set (IS10_paraling)

  As mentioned earlier, paralinguistic information provides an immense body of acoustic features that can be used to encode the vocal state of the speaker. IS10_paraling [35] is a feature set that reflects a broad coverage of paralinguistic information assessment. It contains 1582 features for one utterance obtained in total by systematic 'brute-force' feature generation in three steps. Initially, 38 low-level descriptors were extracted from the data at a rate of 100 frames per second. This extraction process involved using varying window types and sizes, specifically a Hamming window for 25 milliseconds for most descriptors and a Gaussian window for 60 milliseconds for pitch. These descriptors were then subjected to smoothing using a simple moving average low-pass filter, which had a window length of 3 frames. Then 21 functionals to each instance in the databases but excluded 16 features with zero information. Finally, two single features F0 number of onsets and turn duration are added.

- INTERSPEECH 2016 Computational Paralinguistics Challenge Feature Set (ComParE_2016)

  The ComParE_2016 feature set [36] has demonstrated effectiveness across a multitude of paralinguistic tasks [62]. The ComParE_2016 feature set is calculated from the computation of various functionals over LLD contours, including energy, spectral, MFCC, and voicing-related LLDs. LLDs include logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness. Statistical functions are also computed, bringing the total to 6,373 features.

D. Text Features

Symptoms of ADHD, including impulsivity, inattention, and hyperactivity, can often be observed in linguistic patterns

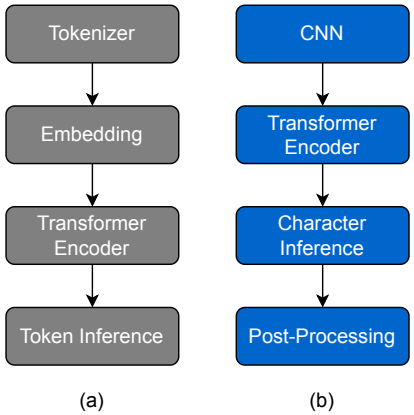


Fig. 3. (a) The typical process of BERT. (b) The typical process of Wav2Vec.

and behaviours. Such linguistic indications may range from inconsistent narrative structures and frequent changes in topic to distinct syntactic formations. Text data offers an avenue to identify and study these nuanced linguistic deviations. Furthermore, during the recording of the experimental data, on certain questions, we found that the responses between ADHD patients and normal subjects were different to some extent, e.g., P7's and P9's answers to the Q4 question are *yeah it i left pretty much everything last minute my dissertation in the first one was like two weeks before i go i still every time i get through something I am confused how I succeeded if that makes sense I like I did not do I've never done like well but i am always confused cos it is like a week or like a day or like you know.* and *sometimes not always depends on how important it is*, respectively. Consequently, we employ deep pre-trained models to derive linguistic and semantic attributes, facilitating the identification of ADHD.

- BERT linguistic Features

  BERT [48] comprises a stack of transformer encoder layers, with its principal advantages encompassing bidirectional pre-training and a unified architecture applicable across various tasks such as modelling the linguistic information. As depicted in Fig.3 (a), the initial step in the BERT process involves preprocessing the transcripts into sub-word-level tokens. Following this, special tokens [CLS] and [SEP] are incorporated, and the text is converted into word embeddings. These word embeddings are subsequently fed into a transformer encoder to produce the final output embeddings.
  In this research, we use the pre-trained BERT-base model with the Hugging Face Transformers library [2] to extract linguistic features from the text. The hyperparameters were set as learning rate to 2e-5, batch size to 4, epochs to 8, and max input length of 256 (sufficient to cover most cases). Two special tokens, [CLS] and [SEP], were added to the beginning and the end of each input.

- Wav2Vec2.0 Semantic Features

  Wav2Vec2.0 [38] is a framework for self-supervised learning of representations from audio data. As presented in Fig.3 (b), in the Wav2Vec model, the first step involves feeding the speech data into a Convolutional Neural Network (CNN) to acquire latent representations. These representations are then passed into a transformer encoder to create context representations. Subsequently, these representations are used in a pre-training task following a self-supervised training strategy. Following pre-training, Wav2Vec undergoes fine-tuning using a character inference component optimized with a Connectionist Temporal Classification (CTC) loss. This component comprises a 1D convolutional layer and a softmax layer, where the convolutional layer operates along the time dimension with a kernel size and stride set to 1, Wav2Vec then processes the output by consolidating consecutive repeated characters, removing blank tokens, and utilizing separator tokens for word separation. The resulting transcript, devoid of punctuation, encapsulates semantic information from the speech data and can be employed as input for transcript-based models.
  In this research, we make use of the huggingface implementation of the wav2vec2.0 base model wav2vec2-base-960h. This base model [3] is pre-trained and fine-tuned on 960 hours of Librispeech [63] on 16kHz sampled speech audio.

IV. Experimental Evaluations and Results

A. Classifiers

Drawing inspiration from a prior study [64] pursuing similar research objectives, we have selected Support Vector Machine (SVM) and Logistic Regression (LR) as our classifiers for distinguishing between ADHD participants and healthy controls. Both SVM and LR, renowned supervised learning algorithms, are particularly effective in binary classification tasks. This choice is reinforced by their demonstrated capability in handling both semi-structured and structured data, a critical factor in our study. Their robustness against overfitting and adaptability in diverse feature classification scenarios, as evidenced in similar studies, make them highly suitable for accurately classifying and differentiating within our target groups. Our experimental approach involves using selected audio and text features as input for the classifier. Specifically, SVM with Gaussian Kernel and 1.0 regularization parameter are employed for feature classification. The leave-one-subject-out (LOSO) cross-validation is used to evaluate the model and all the evaluating data is speaker-independent.

B. Evaluation Metric

The receiver operating characteristics (ROC) and area under the curve (AUC), confusion matrix, accuracy, precision, recall, and F1-score a common metrics to evaluate a classification model and can be calculated as shown from equations (1) to (4). The *True Positive*, *True Negative*, *False Positive*, *False Negative* are in short as *TP*, *TN*, *FP* and *FN*, respectively. F1 Score is the harmonic mean value of precision and recall and

[2]https://github.com/huggingface/transformers

[3]https://huggingface.co/facebook/wav2vec-base-960h

TABLE III
THE ADHD CLASSIFICATION PERFORMANCE COMPARISON ON DIFFERENT FEATURES WITH LOSO CROSS-VALIDATION AND SVM CLASSIFIER. THE **BOLD** NUMBER SHOWS THE BEST PERFORMANCE.

Modality	Feature Set	Accuracy	Precision	Recall	F1-Score		
					ADHD	Control	AVG
Acoustic	LPC	0.680	0.642	0.667	0.654	0.701	0.678
	MFCC	0.704	0.661	0.714	0.687	0.729	0.703
	GeMAPS	0.716	0.656	0.791	0.717	0.716	0.716
	IS10_paraling	0.725	0.669	0.781	0.721	0.729	0.725
	ComParE_2016	0.736	0.679	0.795	0.733	0.739	0.736
	eGeMAPS	**0.762**	**0.697**	**0.843**	**0.763**	**0.761**	**0.762**
Text	BERT	0.727	0.664	0.810	0.730	0.725	0.727
	Wav2Vec 2.0	**0.768**	**0.702**	**0.852**	**0.770**	**0.767**	**0.768**

TABLE IV
THE ADHD CLASSIFICATION PERFORMANCE COMPARISON ON DIFFERENT FEATURES WITH LOSO CROSS-VALIDATION AND LR CLASSIFIER. THE **BOLD** NUMBER SHOWS THE BEST PERFORMANCE.

Modality	Feature Set	Accuracy	Precision	Recall	F1-Score		
					ADHD	Control	AVG
Acoustic	LPC	0.695	0.656	0.691	0.673	0.714	0.693
	MFCC	0.701	0.653	0.733	0.691	0.711	0.701
	IS10_paraling	0.718	0.661	0.781	0.716	0.721	0.718
	ComParE_2016	0.725	0.665	0.795	0.724	0.725	0.725
	GeMAPS	0.734	0.675	0.801	0.732	0.735	0.734
	eGeMAPS	**0.784**	**0.717**	**0.867**	**0.785**	**0.783**	**0.784**
Text	BERT	0.740	0.678	0.815	0.740	0.740	0.740
	Wav2Vec 2.0	**0.753**	**0.686**	**0.842**	**0.756**	**0.750**	**0.753**

both positive and negative F1 scores are reported to eliminate the impact of unbalanced data. A higher F1 score indicates better discrimination.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

C. Acoustic Feature Analysis

Tables 3 and 4 display the classification results for all feature sets using SVM and LR, respectively. In both cases, both SVM and LR classifiers reveal insightful trends, there is a clear progression in performance metrics as we move from acoustic features like LPC to more specially designed feature sets like eGeMAPS. Specifically, eGeMAPS stands out as the best-performing acoustic feature set in both classifiers, achieving the highest scores in accuracy, precision, recall, and F1-score. This indicates its robustness and effectiveness in distinguishing between ADHD and control groups. The consistent improvement across different feature sets from LPC to eGeMAPS suggests that manually designed hand-crafted acoustic features capture better nuance, the improvement in accuracy from LPC to eGeMAPS in the SVM classifier is 8.2% while the improvement is 8.9% with LR classifier, which is crucial for effective ADHD detection. The similarity in performance by the SVM and LR classifiers also highlights that these findings are likely feature-driven rather than dependent on the choice of the classification algorithm. This consistency emphasizes the importance of feature selection, particularly advanced acoustic features e.g., eGeMAPS, in developing effective ADHD diagnostic tools.

Moreover, it can be found that general acoustic features like LPCs and MFCCs are commonly used in speech-processing tasks and represent the spectral characteristics of the audio signal, such as the distribution of energy in different frequency bands and can capture aspects of voice quality and characteristics of speech, but they may not be specific enough to capture the subtle variations in speech associated with ADHD symptoms. Specially designed manual feature sets, based on domain knowledge, go beyond basic acoustic characteristics and emphasize the paralinguistic aspects of speech. Paralinguistic features, such as frequency-related parameters, energy-related parameters, and spectral parameters, are more informative in capturing the nuances of speech that may indicate ADHD-related symptoms. Generally, a relation between speech abnormalities and ADHD traits seems plausible and should be evaluated in more detail in future research.

D. Text Feature Analysis

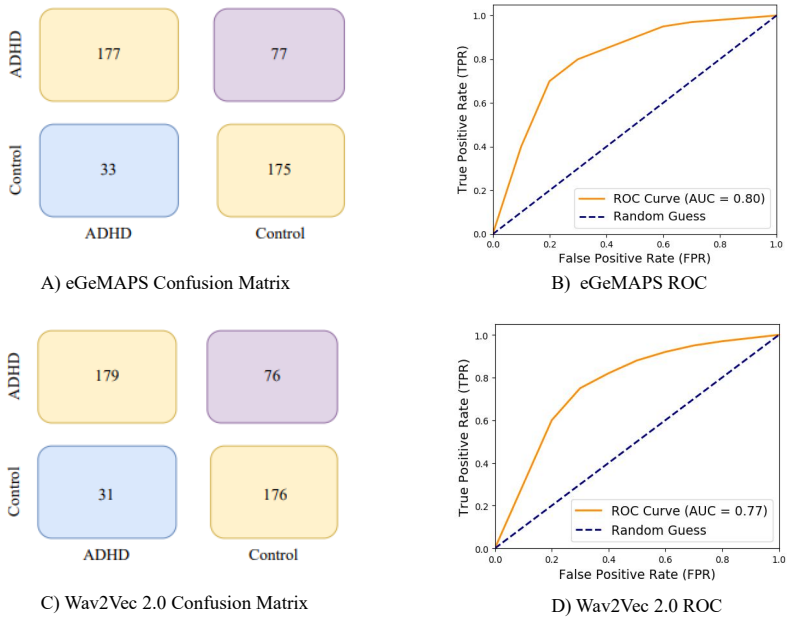
Based on the provided results from Table III, it can be indicated that both linguistic and semantic features contribute effectively to the classification process, yet they exhibit distinct differences in terms of accuracy. Specifically, in the context of the SVM classifier, the linguistic feature set attained an accuracy of 72.7%, while the semantic feature set surpassed this, achieving a higher accuracy of 76.8%. Similarly, when employing the LR classifier, the linguistic features registered an accuracy of 74.0%, whereas the semantic features demonstrated a marginally superior performance with an accuracy of 75.3%. These results underscore the nuanced efficacy of semantic features in comparison to linguistic features across

TABLE V
THE ADHD CLASSIFICATION PERFORMANCE COMPARISON ON INTEGRATED FEATURES WITH LOSO CROSS-VALIDATION AND SVM CLASSIFIER. THE **BOLD** NUMBER SHOWS THE BEST PERFORMANCE.

Feature Set	Accuracy	Precision	Recall	F1-Score		
				ADHD	Control	AVG
IS10_paraling+BERT	0.729	0.668	0.805	0.730	0.729	0.729
ComParE_2016+BERT	0.736	0.675	0.810	0.736	0.736	0.736
eGeMAPS+BERT	0.740	0.674	0.829	0.744	0.737	0.740
ComParE_2016+Wav2Vec2.0	0.760	0.694	0.843	0.761	0.758	0.760
IS10_paraling+Wav2Vec2.0	0.768	0.704	0.848	0.769	0.768	0.768
eGeMAPS+Wav2vec2.0	**0.773**	**0.706**	**0.857**	**0.774**	**0.771**	**0.773**

TABLE VI
THE ADHD CLASSIFICATION PERFORMANCE COMPARISON ON INTEGRATED FEATURES WITH LOSO CROSS-VALIDATION AND LR CLASSIFIER. THE **BOLD** NUMBER SHOWS THE BEST PERFORMANCE.

Feature Set	Accuracy	Precision	Recall	F1-Score		
				ADHD	Control	AVG
IS10_paraling+BERT	0.708	0.647	0.786	0.710	0.706	0.708
eGeMAPS+BERT	0.731	0.664	0.829	0.737	0.725	0.731
ComParE_2016+BERT	0.747	0.680	0.834	0.751	0.743	0.747
IS10_paraling+Wav2Vec2.0	0.767	0.694	**0.872**	0.773	0.760	0.767
ComParE_2016+Wav2Vec2.0	0.773	0.709	0.848	0.772	0.773	0.773
eGeMAPS+Wav2vec2.0	**0.784**	**0.722**	0.852	**0.782**	**0.786**	**0.784**



A) eGeMAPS Confusion Matrix

B) eGeMAPS ROC

C) Wav2Vec 2.0 Confusion Matrix

D) Wav2Vec 2.0 ROC

Fig. 4. A) Confusion Matrix for eGeMAPS Feature Using SVM Classifier. B) ROC Curve for eGeMAPS Feature Using SVM Classifier. C) Confusion Matrix for Wav2Vec 2.0 Semantic Feature Using SVM Classifier. D) ROC curve for Wav2Vec 2.0 Semantic Feature Using SVM Classifier. The horizontal is the actual label and the vertical is the predicted label.

different classification models. In conclusion, leveraging state-of-the-art natural language processing and speech processing models can be effective in identifying ADHD-related patterns in text data. The data shows that while linguistic features are valuable for ADHD classification, semantic features are more insightful, capturing finer details crucial for this task. The superior accuracy of the semantic features underscores the importance of grasping the underlying meaning and context of words. This suggests that a focus on semantics yields a more accurate classification compared to solely relying on the structural aspects of language.

However, compared to the speech modality, the text modality requires more computational resources while achieving similar performance. In this experiment, there are two BERT models available, namely BERT-large and BERT-base. However, due to the computational limitations (we only have access to a workstation equipped with 4 Nvidia GTX 1080ti GPUs and 16 GB of RAM), we opted to use the BERT-base model. It can be assumed that using a larger model would likely result in higher accuracy. However, as the text data are collected by using specially designed questionnaires, they cannot meet the general situations.

To represent the classification accuracy more intuitively, Fig.4 displays the confusion matrices and ROC curves for

TABLE VII
COMPARISON OF ADHD WITH DIFFERENT MODALITIES USING MACHINE LEARNING METHOD.

Author [ref]	Dataset	Modality	ML Model	Accuracy (%)
Das et al. [65]	28 ADHD 22 Control	Pupillometric	SVM	76.1%
Ning et al. [66]	118 ADHD 98 Control	MRI	LR	75.8%
Luo et al. [67]	52 ADHD 44 Control	MRI	SVM	76.6%
Kim et al. [68]	34 ADHD 45 Normal	EEG	Ensemble	81.0%
Our Method	10 ADHD 12 Control	Audio	LR	78.4%
Our Method	10 ADHD 12 Control	Text	SVM	76.8%

the two most discriminative features in both the speech and text modalities using the SVM classifier. As shown in Fig.4, both features' performance is far above the random selection, based on the eGeMAPS feature set, 77 out of 252 healthy audio clips are misclassified as ADHD while 33 of normal controls are classified as ADHD. When using the Wav2Vec 2.0 model, there were 76 and 31 misclassified subjects and controls, respectively.

E. Integrated Feature Analysis

In order to investigate whether acoustic features and text features could be complementary to detect ADHD, We simply concatenate the acoustic features (eGeMAPS, IS10_paraling, ComParE_2016) and text features (BERT, Wav2Vec 2.0) as integrated features. The fused feature is then applied to the SVM and LR for ADHD classification. The results of integrated features are shown in Table V and Table VI.

Among all the integrated features, the feature set "eGeMAPS+Wav2Vec2.0" consistently emerges as the top performer in ADHD classification across both SVM and LR classifiers, underscoring its robustness and effectiveness. It achieves the highest scores in all key metrics: Accuracy (0.773 in SVM, 0.784 in LR), Precision (0.706 in SVM, 0.722 in LR), Recall (0.857 in SVM), and F1-Score (0.782 in LR). This superior performance indicates the strong capability of this feature set in accurately classifying ADHD cases. Notably, there is a marked improvement in performance from "IS10˙paraling+BERT" to "eGeMAPS+Wav2Vec2.0", highlighting the effectiveness of combining advanced acoustic features (eGeMAPS) with sophisticated language models (Wav2Vec2.0). Furthermore, the LR classifier generally exhibits a slight enhancement in performance metrics over the SVM classifier for the same feature sets, suggesting that LR may be more adept at handling this specific classification task. This consistency and improvement across different classifiers emphasize the potential of "eGeMAPS+Wav2Vec2.0" as a reliable and powerful tool in the domain of ADHD diagnosis.

In summary, the results indicate that all the feature sets exhibit greater robustness in identifying ADHD patients compared to identifying individuals without ADHD. This observation may be attributed to the data imbalance between the two groups, as individuals with ADHD tend to engage in tangential speech patterns when compared to normal controls. Consequently, the dataset contains a larger volume of ADHD-related data, potentially contributing to the enhanced performance in identifying ADHD cases. Besides, while paralinguistic changes show promise as potential indicators of ADHD, relying solely on the acoustic modality may not provide sufficient information for accurately diagnosing the condition. The integration of linguistic features into the analysis consistently yields substantial enhancements in the predictive accuracy of ADHD, even in cases where the transcripts exhibit relatively high word error rates.

F. Comparison With Traditional Methods

We also compared our method with other traditional data sources such as EEG and MRI in recent years. The results are shown in Table VII. It can be found that the traditional data source reveals high effectiveness, with EEG leading up to 81.0% accuracy. These methods, however, typically involve larger datasets compared to our study using audio and text, where the number of subjects is notably smaller. Despite this, the accessibility and competitive accuracy of audio and text data, even with smaller dataset sizes, offer a practical and promising alternative, particularly in diverse and resource-constrained settings. Our approach could significantly broaden the scope and feasibility of ADHD research and diagnosis.

V. DISCUSSIONS AND CONCLUSIONS

This work shows that using audio and text features to detect ADHD is feasible with considerable results. However, further improvement is possible in various aspects.

In the course of our research, we encountered a notable disparity in the gender distribution among the participants included in our dataset. Specifically, our dataset was composed of an approximately equal split between individuals diagnosed with ADHD, with 50% being males and 50% females, while the control group consisted of 66.7% males and 33.3% females. This observed imbalance was particularly pronounced in terms of the representation of individuals diagnosed with ADHD across different genders. It's worth highlighting that this gender imbalance within our dataset surpasses the expected male-to-female incidence ratio of 1.6:1, which is commonly reported in the DSM-V [7]. This manual serves as a reference for the prevalence of ADHD among different genders in clinical and research settings. Importantly, it is crucial to acknowledge that our dataset, although derived from real-world recordings and clinical data, may not provide a completely accurate reflection of the broader population's gender distribution to ADHD. The skew in gender representation within our dataset underscores the complexities of studying and interpreting the real-world prevalence and characteristics of ADHD, which may vary across different demographic groups. Therefore, any findings or conclusions drawn from this dataset should be interpreted in light of this inherent gender bias.

Secondly, It's important to recognize the impact of inherent factors in data collection, particularly the limited number of

participants, which can lead to reliance on individual-level outcomes and affect the robustness and generalizability of the findings. Remarkably, despite the challenges posed by this limited dataset, it is noteworthy that several features exhibited notably high levels of accuracy. Besides, when unimodal data is limited, ADHD's heterogeneity necessitates a multi-perspective approach, making the use of multimodal data, including brain MRI, physiological signals, and behavioural tests, essential for comprehensive study. By employing Deep Learning models, which excel in integrating varied data through their hierarchical structure, each modality is effectively analyzed and combined. Both of the two points prove that the expansion of datasets in future research is crucial for enhancing ADHD classification, as larger datasets significantly improve model performance. With more comprehensive training data, models are more likely to detect broader patterns and nuances, leading to results that are not only more accurate but also more reliable. This underscores the substantial potential of expanded datasets in advancing our understanding and capabilities in ADHD research.

A third noteworthy observation pertains to the disparity in data availability between speech and text modalities concerning individuals with ADHD and normal controls. This disparity arises partly due to the characteristic symptomatology of ADHD, including hyperactivity, resulting in a higher volume of data in speech and text modalities from individuals with ADHD. An alternative way to address the data imbalance issue is to use transfer learning where we can transfer the knowledge from a related task that has already been learned. Transfer learning offers a pathway to leverage the accumulated knowledge from related tasks to improve the accuracy and reliability of ADHD classification models. Transfer learning allows us to benefit from insights and features learned from related tasks. Significantly improving the performance and generalization of models even in the face of unbalanced datasets. In addition, migration learning typically requires fewer labelled examples, making it particularly valuable in situations where access to data is limited.

Lastly, the automatic extraction of features from text patterns using large-scale pre-trained models has demonstrated encouraging outcomes. However, it's worth noting that in the domain of acoustic patterns, the automatic extraction of features using neural networks remains an underexplored area. This presents a compelling avenue for future research and exploration. Besides, the integration of audio and text features demonstrates promising results in our study. In essence, it suggests that the interplay between acoustic and linguistic elements is crucial for a more comprehensive understanding and prediction of ADHD, transcending the limitations of each modality when considered in isolation. This finding emphasizes the importance of a multimodal approach that harnesses both paralinguistic and linguistic cues for a more robust and accurate assessment of ADHD.

Overall, we hold the perspective that leveraging audio and text data for ADHD classification represents a worthwhile pursuit. This is due to the ease and cost-effectiveness of collecting such data and the potential for generalizable results. Consequently, this approach holds promise for potential application in large-scale ADHD screening efforts. Nevertheless, it is crucial to emphasize that, in its current state, this approach does not supplant the significance of clinical diagnosis conducted by seasoned experts, which continues to represent the benchmark for ADHD assessment.

REFERENCES

[1] J. Campion, A. Javed, C. Lund, *et al.*, "Public mental health: required actions to address implementation failure in the context of COVID-19," *The Lancet Psychiatry*, vol. 9, no. 2, pp. 169–182, 2022.

[2] K. O. Plowden, T. Legg, and D. Wiley, "Attention deficit/hyperactivity disorder in adults: A case study," *Psychiatric Nursing*, vol. 38, pp. 29–35, 2022.

[3] C. Mohr-Jensen and H. C. Steinhausen, "A meta-analysis and systematic review of the risks associated with childhood attention-deficit hyperactivity disorder on long-term outcome of arrests, convictions, and incarcerations.," *Clinical Psychology Review*, vol. 48, pp. 32–42, 2016.

[4] J. Posner, G. V. Polanczyk, and E. J. S. Sonuga-Barke, "Attention-deficit hyperactivity disorder," *The Lancet*, vol. 395, pp. 450–462, 2020.

[5] S. A. Chong, "Mental health in singapore: a quiet revolution?," *Annals-Academy of Medicine Singapore*, vol. 36, no. 10, p. 795, 2007.

[6] BBC, "ADHD diagnosis for adult 'can take seven years'." Accessed Aug. 2023. https://www.bbc.co.uk/news/uk-england-44956540, 2018.

[7] G. Arbanas, "Diagnostic and statistical manual of mental disorders (DSM-V)," *Alcoholism and Psychiatry Research*, pp. 61–64, 2015.

[8] J. A. Fayyad, N. A. Sampson, I. H. Hwang, *et al.*, "The descriptive epidemiology of DSM-IV adult ADHD in the World Health Organization World Mental Health surveys," *ADHD Attention Deficit and Hyperactivity Disorders*, vol. 9, pp. 47–65, 2016.

[9] P. Song, M. Zha, Q. Yang, *et al.*, "The prevalence of adult attention-deficit hyperactivity disorder: A global systematic review and meta-analysis," *Journal of Global Health*, vol. 11, 2021.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] H. Duan, Y. Long, S. Wang, H. Zhang, C. G. Willcocks, and L. Shao, "Dynamic unary convolution in transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[12] J. Guo, W. Cao, B. Nie, and Q. Qin, "Unsupervised learning composite network to reduce training cost of deep learning model for colorectal cancer diagnosis," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 54–59, 2022.

[13] Y. Zhang, T. Liu, V. Lanfranchi, and P. Yang, "Explainable tensor multi-task ensemble learning based on brain structure variation for Alzheimer's disease dynamic prediction," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 1–12, 2022.

[14] M. A. Ottom, H. A. Rahman, and I. D. Dinov, "Znet: deep learning approach for 2d mri brain tumor segmentation," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–8, 2022.

[15] S. R. Chetupalli, P. Krishnan, N. Sharma, *et al.*, "Multi-modal point-of-care diagnostics for COVID-19 based on acoustics and symptoms," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 11, pp. 199–210, 2023.

[16] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: opportunities and challenges," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 3, no. 3, pp. 223–230, 2018.

[17] Y. Tang, J. Sun, C. Wang, *et al.*, "ADHD classification using auto-encoding neural network and binary hypothesis testing," *Artificial Intelligence in Medicine*, vol. 123, p. 102209, 2022.

[18] Y. K. Boroujeni, A. A. Rastegari, and H. Khodadadi, "Diagnosis of attention deficit hyperactivity disorder using non-linear analysis of the EEG signal.," *IET Systems Biology*, vol. 135, pp. 260–266, 2019.

[19] J. E. Koh, C. P. Ooi, N. S. Lim-Ashworth, *et al.*, "Automated classification of attention deficit hyperactivity disorder and conduct disorder using entropy features with ECG signals," *Computers in Biology and Medicine*, vol. 140, p. 105120, 2022.

[20] C. L. Nash, R. Nair, and S. M. Naqvi, "Machine learning and ADHD mental health detection - a short survey," *International Conference on Information Fusion*, pp. 1–8, 2022.

[21] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *NPJ Digital Medicine*, vol. 5, no. 1, p. 46, 2022.

[22] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, pp. 96 – 116, 2020.

[23] B. Stasak, J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect.," *Interspeech*, vol. 17, pp. 834–838, 2017.

[24] J. Han, Z. Zhang, C. Mascolo, *et al.*, "Deep learning for mobile mental health: Challenges and recent advances," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 96–105, 2021.

[25] P. Bellec, C. Chu, F. Chouinard-Decorte, *et al.*, "The neuro bureau ADHD-200 preprocessed repository," *Neuroimage*, vol. 144, pp. 275–286, 2017.

[26] V. Pereira-Sanchez and F. X. Castellanos, "Neuroimaging in attention-deficit/hyperactivity disorder," *Current Opinion in Psychiatry*, vol. 34, no. 2, p. 105, 2021.

[27] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bio-acoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, p. 105020, 2023.

[28] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli, "Voice patterns in schizophrenia: A systematic review and bayesian meta-analysis," *Schizophrenia Research*, vol. 216, pp. 24–40, 2020.

[29] N. Cummins, S. Scherer, J. Krajewski, *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[30] P. Lopez-Otero, L. Dacia-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for depression detection," *2nd International Workshop on Biometrics and Forensics*, pp. 1–6, 2014.

[31] T. Alhanai, R. Au, and J. Glass, "Spoken language biomarkers for detecting cognitive impairment," *Automatic Speech Recognition and Understanding Workshop*, pp. 409–416, 2017.

[32] S. Nasreen, J. Hough, M. Purver, *et al.*, "Detecting Alzheimer's disease using interactional and acoustic features from spontaneous speech," *Interspeech*, 2021.

[33] M. Rohanian, J. Hough, and M. Purver, "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs," *ArXiv preprint arXiv:2106.15684*, 2021.

[34] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[35] B. Schuller, S. Steidl, A. Batliner, *et al.*, "The Interspeech 2010 paralinguistic challenge.," *Interspeech*, pp. 2794–2797, 2010.

[36] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, *et al.*, "The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," *Interspeech*, vol. 8, pp. 2001–2005, 2016.

[37] F. Haider, S. De La Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2019.

[38] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[39] Y. Qin, W. Liu, Z. Peng, S.-I. Ng, J. Li, H. Hu, and T. Lee, "Exploiting pre-trained ASR models for alzheimer's disease recognition through spontaneous speech," *ArXiv preprint arXiv:2110.01493*, 2021.

[40] S. Li, Y. Sun, R. Nair, and S. M. Naqvi, "Enhancing ADHD detection using DIVA interview-based audio signals and a two-stream network," *2023 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pp. 291–296, 2023.

[41] S. Li, R. Nair, and S. M. Naqvi, "Detecting ADHD from speech using full-band and sub-band convolution fusion network," *2023 IEEE SENSORS*, pp. 1–4, 2023.

[42] P. E. Engelhardt, M. Corley, J. T. Nigg, and F. Ferreira, "The role of inhibition in the production of disfluencies," *Memory & Cognition*, vol. 38, no. 5, pp. 617–628, 2010.

[43] M. Kamath, C. Dahm, J. Tucker, *et al.*, "Sensory profiles in adults with and without ADHD," *Research in developmental disabilities*, vol. 104, p. 103696, 2020.

[44] P. P. Sinha, R. Mishra, R. Sawhney, *et al.*, "Suicidal: A multipronged approach to identify and explore suicidal ideation in Twitter," *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 941–950, 2019.

[45] K. Allen, A. L. Davis, and T. Krishnamurti, "Indirect identification of perinatal psychosocial risks from natural language," *IEEE Transactions on Affective Computing*, 2021.

[46] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," *ArXiv preprint arXiv:1709.01848*, 2017.

[47] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 136–143, 2016.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[49] A. Balagopalan, B. Eyre, F. Rudzicz, and J. Novikova, "To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection," *arXiv preprint arXiv:2008.01551*, 2020.

[50] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The Adresso challenge," *arXiv preprint arXiv:2104.09356*, 2021.

[51] Y. Li, S. Li, C. Nash, S. M. Naqvi, and R. Nair, "24 intelligent sensing in ADHD trial ISAT–pilot study," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 94, p. 2, 2023.

[52] M. Sandberg, "Cambridge neuropsychological testing automated battery," *Encyclopedia of clinical neuropsychology. Springer*, pp. 480–482, 2011.

[53] R. C. Kessler, L. Adler, M. Ames, *et al.*, "The world health organization adult ADHD self-report scale ASRS: a short screening scale for use in the general population," *Psychological Medicine*, vol. 35, no. 2, pp. 245–256, 2005.

[54] J. J. S. Kooij and M. Francken, "Diagnostic Interview for ADHD in adults (DIVA 2.0)," 2010.

[55] B. McFee, C. Raffel, D. Liang, *et al.*, "librosa: Audio and music signal analysis in python," *Proceedings of the 14th python in science conference*, vol. 8, pp. 18–25, 2015.

[56] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

[57] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[58] M. Guzman, T. Bertucci, C. Pacheco, *et al.*, "Effectiveness of a physiologic voice therapy program based on different semioccluded vocal tract exercises in subjects with behavioural dysphonia: a randomized controlled trial," *Journal of Communication Disorders*, vol. 87, p. 106023, 2020.

[59] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[60] J. De Boer, A. Voppel, S. Brederoo, *et al.*, "Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom recognition tool," *Psychological Medicine*, vol. 53, no. 4, pp. 1302–1312, 2023.

[61] M. Godel, F. Robain, F. Journal, *et al.*, "Prosodic signatures of ASD severity and developmental delay in preschoolers," *NPJ Digital Medicine*, vol. 6, no. 1, p. 99, 2023.

[62] N. Cummins, A. Baird, and B. Schuller, "The increasing impact of deep learning on speech analysis for health: Challenges and opportunities," *Methods, Special Issue on Translational Data Analytics and Health Informatics*, vol. 151, pp. 41–54, 2018.

[63] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[64] H. Tanaka, H. Adachi, N. Ukita, *et al.*, "Detecting dementia through interactive computer avatars," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 5, pp. 1–11, 2017.

[65] W. Das and S. Khanna, "A robust machine learning based framework for the automated detection of ADHD using pupillometric biomarkers and time series analysis," *Scientific reports*, vol. 11, no. 1, p. 16370, 2021.

[66] N. Qiang, Q. Dong, H. Liang, *et al.*, "A novel ADHD classification method based on resting state temporal templates (rstt) using spatiotemporal attention auto-encoder," *Neural Computing and Applications*, vol. 34, no. 10, pp. 7815–7833, 2022.

[67] Y. Luo, T. L. Alvarez, J. M. Halperin, and X. Li, "Multimodal neuroimaging-based prediction of adult outcomes in childhood-onset ADHD using ensemble learning techniques," *NeuroImage: Clinical*, vol. 26, p. 102238, 2020.

[68] S. Kim, J. H. Baek, Y. J. Kwon, *et al.*, "Machine-learning-based diagnosis of drug-naive adult patients with attention-deficit hyperactivity disorder using mismatch negativity," *Translational Psychiatry*, vol. 11, no. 1, p. 484, 2021.