High-Density Electroencephalography and Speech Signal Based Deep Framework for Clinical Depression Diagnosis

Abdul Qayyum, Imran Razzak, M. Tanveer, Moona Mazher, and Bandar Alhaqbani

***Abstract*—Depression is a mental disorder characterized by persistent depressed mood or loss of interest in performing activities, causing significant impairment in daily routine. Possible causes include psychological, biological, and social sources of distress. Clinical depression is the more-severe form of depression, also known as major depression or major depressive disorder. Recently, electroencephalography and speech signals have been used for early diagnosis of depression; however, they focus on moderate or severe depression. We have combined audio spectrogram and multiple frequencies of EEG signals to improve diagnostic performance. To do so, we have fused different levels of speech and EEG features to generate descriptive features and applied vision transformers and various pre-trained networks on the speech and EEG spectrum. We have conducted extensive experiments on Multimodal Open Dataset for Mental-disorder Analysis (MODMA) dataset, which showed significant improvement in performance in depression diagnosis (0.972, 0.973 and 0.973 precision, recall and F1 score respectively) for patients at the mild stage. Besides, we provided a web-based framework using Flask and provided the source code publicly.[1]**

***Index Terms*—Depression, EEG, multimodal depression, transfer learning.**

I. Introduction

DEPRESSION, anxiety, and suicide are big threats to any of us that can impact our life any time. While most of us do not know the root cause of depression, there are many factors associated with the development of depression. Though, the exact causes of anxiety and depression are unknown, several things are related to their development. It is not simply a change in brain chemical imbalance (i.e., too little or extra particular chemical in the brain). It is complicated, and many factors impact our brain's working and lead us to depression, such as medication, drugs, stress, genetic vulnerability, and medical conditions. While we all feel moody, sad, or low, some feel intense and for a more extended period, and it may be without any apparent reason. Depression ranges in seriousness from mild symptoms to temporary episodes of sadness and severe persistent depression. Clinical depression is the more-severe form of depression, also known as major depressive disorder or major depression. It is marked by a depressed mood most of the day, sometimes particularly in the morning, and a loss of interest in performing routine activities and relationships.

Depression is a threatening life condition and impacts our mental and physical health. Researchers argue that continuous challenges in employment, illness, abusive relationships, loneliness, or longer working hours may cause depression. Even though the causes of depression have been actively explored, there is still much to be known. Depression is not only the result of a chemical imbalance (not enough or too much) but several other factors that cause depression, such as severe life stress, genetic vulnerability, substances that one takes (drugs, alcohol, or medications,) or any medical conditions. Recent smaller events may trigger depression if we are already facing similar issues and have bad experiences already in our life, or it may be due to personal factors such as

- *Family history* – One can have increased genetic risk if there is a family depression history; however, having a family history does not mean one can face the same experience. Personal factors and life circumstances are critical, and negative aspects can lead to depression.
- *Serious medical illness* – Stress or any medical condition is the leading cause of depression, especially medical conditions that are chronic diseases or long-term diseases.
- *Personality* – Personality plays a considerable role in depression, and some of us may be at higher risk, especially if we are sensitive to personal criticism, perfectionists, self-esteemed, worry a lot, or self-critical and pessimistic.
- *Drug and Alcohol* – Most depression patients consume excess drugs or alcohol.

Everyone is different, and additional factors or combinations of factors could contribute to depression development. Thus, there is no proven approach for depression treatment; it differs from person to person. Even though its treatment

Manuscript received 12 April 2022; revised 20 January 2023; accepted 16 February 2023. Date of publication 14 March 2023; date of current version 8 August 2023. *(Corresponding author: M. Tanveer.)*

Abdul Qayyum is with the Department of Biomedical Engineering, King's College London, WC2R 2LS London, U.K. (e-mail: engr.qayyum@gmail.com).

Imran Razzak is with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: imran.razzak@unsw.edu.au).

Moona Mazher is with the Department of Computer Engineering and Mathematics, University Rovira i Virgili, 43003 Tarragona, Spain (e-mail: moona.mazher@gmail.com).

M. Tanveer is with the Department of Mathematics, Indian Institute of Technology Indore, Indore, Madhya Pradesh 453552, India (e-mail: mtanveer@iiti.ac.in).

Bandar Alhaqbani is with the Technology Control Company, Riyadh 12621, Saudi Arabia (e-mail: haqbanib@gmail.com).

Digital Object Identifier 10.1109/TCBB.2023.3257175

[1]https://github.com/RespectKnowledge/EEG_Speech_Depression_MultiDL

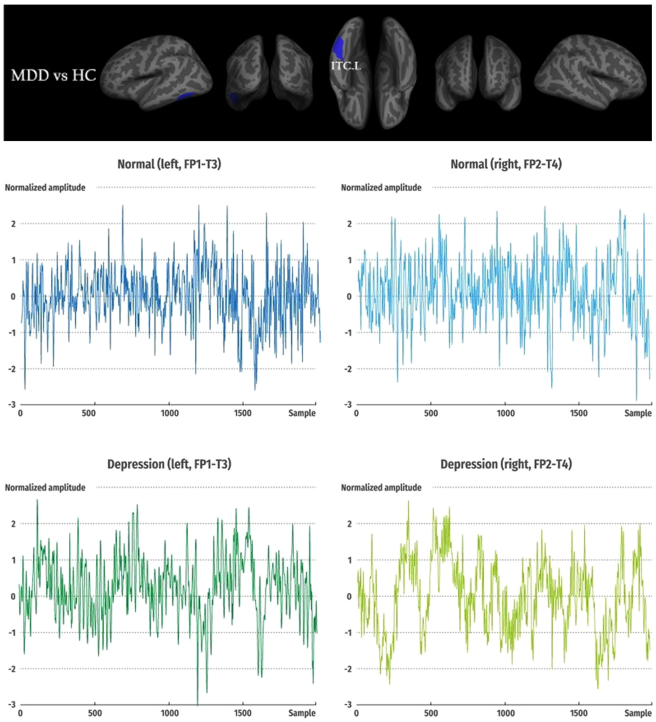


Fig. 1. MDD patients compared with the healthy control group (top) MDD vs healthy changes in MRI [18] (bottom) Normal and depressive EEG signals [1] .

plans are dependent on individual factors. However, there are a range of effective treatment plans and health professionals that can help to recover from severe depression by stimulating the growth of new nerve cells in circuits that regulate mood.

Traditional methods for depression diagnosis include medical experts diagnosing depression through visualizing EEG signal data based on their expertise, which is time-consuming, and requires expertise [20], [21], [22], [23]. Besides, depression diagnosis is challenging through questions asked by an expert as there could be different causes of depression, hence prone to error. For example, some depressed patients may seem to withdraw into a state of apathy, and some may become agitated or irritable. Furthermore, different patient expresses their feeling differently with variable intensity, i.e. sleeping or eating routine can be exaggerated. To overcome the challenges mentioned above, researchers focus on developing an automated framework for detecting depressed patients. Recently, physiological data such as Electroencephalographic, facial expression, voice etc. have been used for a depression diagnosis. Electroencephalographic or EEG is one of the most heritable bio-markers and may serve as an effective treatment planning tool for diagnosing depressed patients (see Fig. 1). The complex and nonlinear nature of electroencephalography requires an efficient and robust approach to precisely diagnose a disease like depression. Recently, several deep learning-based frameworks have been developed [29], [30], especially for diagnosis using different types of data as input. One Dimensional convolutional neural network (1DCNN) is one of the most famous deep learning architectures applied for EEG signal analysis. 1DCNNs extract the abundant features through various filters, thus, showing considerably better performance than traditional machine learning methods. The recurrent neural network can learn from previous iterations during its training. Long short-term memory (LSTM) is one commonly used RNN architectures time series analysis. Most existing methods utilize EEG signals for depression diagnosis; however, the efficiency and performance of depressed people diagnosis can be improved using multimodal data. Recently, multimodal speech and EEG data have been considered for a depression diagnosis. To automatically discriminate between depressed and healthy controls, in this work, we present a novel end-end framework for diagnosing depressed patient using speech and EEG data. We have performed different level EEG and speech feature fusion and applied vision transformer on speech and EEG spectrum, significantly improving diagnosis performance. The **key contributions** of this work are

- presents an end-to-end multimodal depression diagnosis framework to analyze functional brain network analysis in resting state and speech data.
- combined different levels of features extracted from 1D EEG signals and 2D EEG and audio spectrogram to generate descriptive features.
- fused vision transformer on speech and EEG spectrum, which significantly improves diagnosis performance; we have also used different 2D-pertained networks.
- deployed a lightweight web framework using Flask, which is easy for basic web applications and can be used for various platforms.
- Extensive experiments on multimodal EEG and speech datasets show significant performance gains compared to the state-of-the-art methods.

The rest of the paper is organized as: in the next section, we present the mental disorder background and state-of-the-art methods for depression diagnosis. In the next section, we present multimodal deep transfer learning-based framework. Section IV presents experiment setup, dataset, results, and discussion, followed by finding and future recommendations.

II. Related Work

Depression is a constant state of sadness and feeling low. To be diagnosed with depression, one should be suffering from depression symptoms such as feeling sad, hopeless, tired, irritated, frustrated, change in sleep patterns, difficulty remembering or facing physical problems such as sexual dysfunction, fatigue or headache, or even thinking hurt yourself. The United States Preventive Services Task Force recommended considering depression screening in adult populations, especially pregnant and postpartum. Primary healthcare clinics diagnose and refer to mental health professionals, such as psychiatrists or psychologists.

Recently, automated methods for depression diagnosis at an early stage have been developed using social media data, speech, EEG etc. and recent deep learning methods have achieved state of the art performance in the healthcare sector [6], [7], [10],

[11]. User written text such as social media posts can be used to detect depression. Hamad et al. presented multimodal extractive-abstractive summarization-based framework for automatic detection of social media users suffering from depression [31], [33]. In another work, Hamad et al. presented deep hirerchical attention framework for depression diagnosis and how it has impact during Pandemic [32]. They have utilized social media posts of a user to analyze the depression symptoms by extracting fine-grained and relevant contents through extractive-abstractive summarization on user historical posts. Even though depression detection using social media showed excellent detection performance, however, it is quite challenging and subjective to human mode at the time of post, thus prone to error and not efficient for early diagnosis. However, utilizing social media for depression may alter user based on their behavioral change that they might not be noticed using EEG patterns. Recent studies showed that there is the considerable hemispheric asymmetry in brain signals in depression patient in comparison to healthy people [2], [9], [12]. For example, one of the sign is the decrease in brain functioning.

Traditional methods for depression diagnosis include medical experts diagnosing depression by visualizing EEG signal data based on their expertise, which is time-consuming and requires expertise. To overcome the challenges mentioned earlier, researchers are focusing on developing an automated framework for the detection of the depressed patient. Recently, physiological data such as Electroencephalographic, facial expression, voice etc. have been used for depression diagnosis. Electroencephalographic or EEG is one of the most heritable bio-markers and may serve as an effective treatment planning tool for the diagnosis of depressed patients. The complex and nonlinear nature of electroencephalography requires an efficient and robust approach to precisely diagnose disease like depression. Recently, several deep learning-based frameworks haven been developed for depression diagnosis using different data types as input. One Dimensional convolutional neural network (1DCNN) is one of the most famous deep learning architectures applied for EEG signal analysis. 1DCNNs extract the abundant features through various filters thus, showing considerable better performance than traditional machien learning methods. Acharaya et al. presented CNN based depression detection framework. The CNN network consists of 13 layers of abstraction and showed 93.5% and 96% diagnostic performance for left and right hemispheres on a dataset collected from 30 subjects [1]. One of the major disadvantages of time series data is that it can not retain the memory of earlier time-series patterns, making it unable to learn important and representative time series features that are very important for EEG analysis. Consequently, it is difficult to accurately construct the relationship between the EEG signals and depression-related activation channels.

Recurrent neural networks (RNNs) are powerful deep learning networks designed to handle sequence dependence, typically time series data, by retaining the memory of what has already been processed; hence, are able to learn from previous iterations during its training [28]. Long short-term memory (LSTM) is one commonly used RNN architectures time series analysis. Ay et al. CNN based LSTM framework was developed that showed

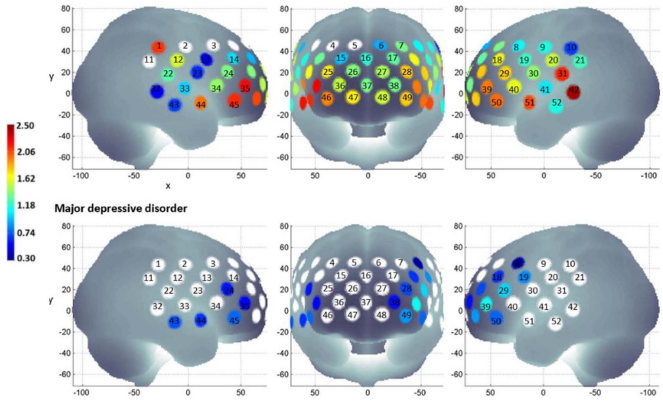


Fig. 2. A comparison of mean oxy-haemoglobin through t-test (Healthy vs Depressed) activation at each channel. Colour gradient indicates the effect size of activation during the VFT. White colored channels did not show differences in oxy-haemoglobin between (pre-task vs task periods) [13].

significant detection performance 99.12% and 97.66% for left and right hemispheres respectively [4]. To diagnose mild depression, Li et al. analyzed various aspects of EEG signals (spectral, spatial, and temporal information) [14] and concluded that spectral information and temporal information showed significant improvement in mild depression diagnosis. To improve the diagnosistic performance, they have used pre-trained ConvNet based framework for classification of EEG-based mental load task and achieved 85.62% accuracy for recognition of patient with mild depression symptoms and normal controls. To extract robust feature set, Liao et al. deployed kernel eigen-filter-bank common spatial pattern and achieved 80% accuracy for severe depression cases [16]. Zhang et al. extracted both nonlinear and linear features EEG data of 25 subjects with closed-eye under resting conditions and achieved 94.2% and 92.9% with BPNN and KNN respectively [26]. Qayyum et al. presented a shallow deep network by combining a CNN with Gated recurrent units for early depression diagnosis [19].

Most of the existing methods utilize EEG signals for a depression diagnosis. However, the efficiency and performance of depressed people diagnosis can be improved using multimodal data. Recently, presented multimodal depression diagnosis using speech and EEG data. In order to automatically discriminate depressed and healthy controls, in this work, we present a novel end-end framework for the diagnosis of the depressed patient using speech and EEG data. We have performed different level EEG and speech feature fusion and applied vision transformer on speech and EEG spectrum, which significantly improves diagnosis performance.

III. Multimodal Depression Diagnostic Framework

This section presents a multi-modal EEG and speech signal-based deep framework for the efficient diagnosis of depression. Efficient detection of mild depression disorder is challenging task due to nature and complexity such as significant variability in signals. Recent advancement in EEG makes it a powerful tool to reflect the working status of the human brain and analyze non-invasive investigation on neurological disorder including

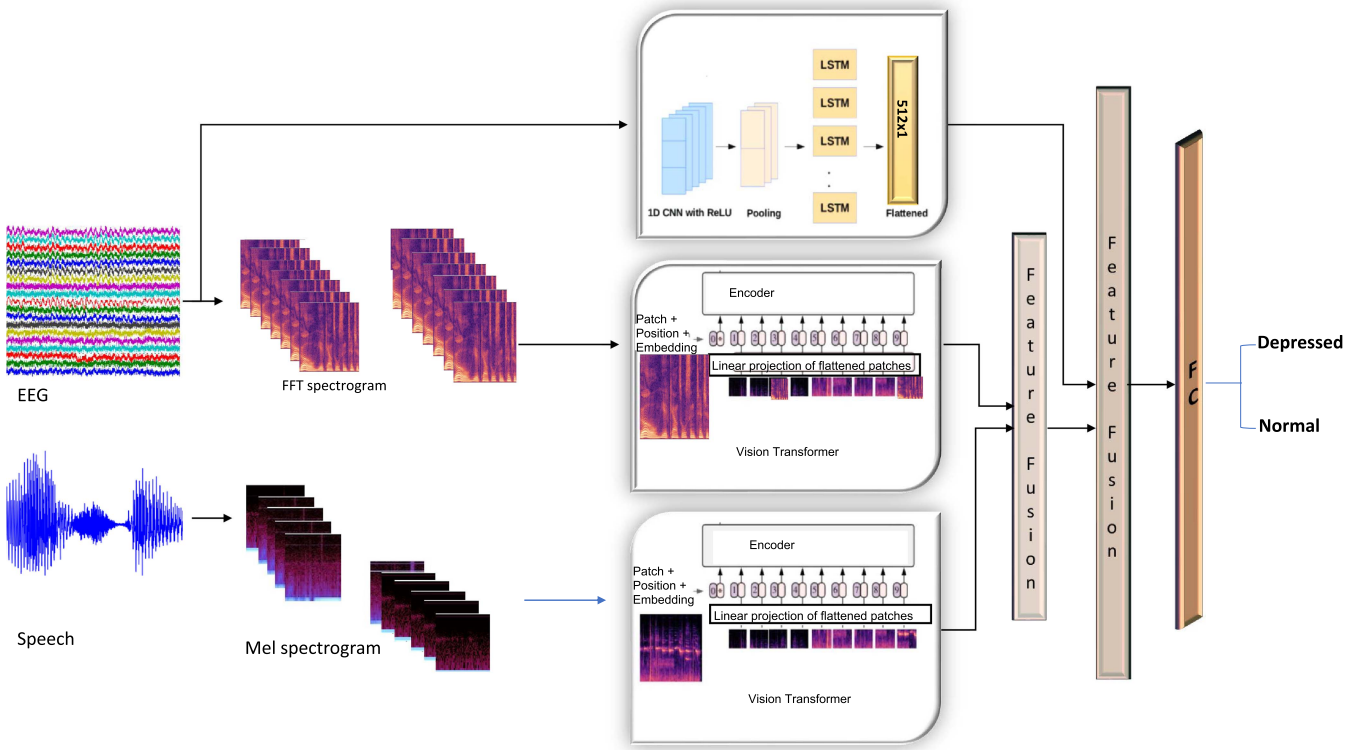


Fig. 3. Hybrid Deep framework for efficient diagnosis of depression.

Alzheimer, Parkinson's and depression. Fig. 3 illustrate the proposed framework. We have fused different types of speech and EEG deep features. As EEG signals are both continuous in time and functionally connected among channels, our model captures both temporal and spatial features. The EEG audio signal is converted to spectrogram by computing the literally tokens aided with positional information, similar to EEG feature extraction. Besides, we have used 1D time series sequential EEG Data to capture temporal relationships. The proposed framework consists of three main components for feature extraction (a) 1DCNN-LSTM, which is based on 1D-EEG signals (2) Speech-Spectrogram, which is a 2D model based on speech signals, and (3). EEG-Spectrogram is a 2D model based on EEG spectrogram. In this experiment, we have proposed two different structures; (a) a 2D-pretrained model and (b) a vision transformer. Features extracted from these components are fused to generate efficient feature representation. The features extracted from each model's last layer are concatenated and fused together to pass a fully-connected network for classification of MDD and normal. The Mel spectrogram and FFT spectrogram were used to convert 29 EEG 1D dataset into 2D images for classification. The source input image would be frequencies (alpha, beta, gamma, theta and so on) of input EEG signal or raw EEG signals. The raw EEG signals or frequencies will be converted into a 2D spectrum using FFT spectrogram for EEG and Mel spectrogram for speech. EEG signals consist of multiple frequency ranges and each frequency range in EEG signals provides task's special information. Specifically, most of the neuronal activities can be reflected by EEG data in the range of 0.5Hz-28 Hz, mainly fall into six frequency bands (Alpha, Beta1, Beta2, Beta3, Theta, and Delta). Hence, most of the information about the intention may lost, if other frequencies are discarded, which may lead to imprecise diagnosis. Though, EEG signals had shown amazing performance, however, there are several challenges present in this task which makes it quite challenging. The key challenges is that EEG is obtained from the scalp of the brain using probes, leading to noise and interruption of many other signals in EEG signals. Besides, the connection between EEG wavelets and multiple intentions recognition is difficult to model, leading to the low performance of the system. Also, the EEG wavelets vary for the same intentions. This is due to the mental state of a person. Aforementioned challenges amy results in poor precise tool. In this work, we have used the most active channel to overcome aforementioned challenges. Fig. 1. illustrate the proposed preprocessing components

A. *Preprocessing*

EEG signal consists of large inter-subject variations with respect to characteristics of brain signals [3]. Besides, signals consist of non-stationary and transitory behavior i.e., outliers measurement artifacts and non-standard noise, making the task very difficult. As we have extracted features from spectrogram as well as from raw signals, thus, to extract features from raw signals, we first applied spatial filtering to reduce the variations, i.e., uncorrelated variations. To determine sample size, We have computed the sample unit as

$$n = \frac{Z^2 p(1-p)}{e^2}$$

where p is an estimated proportion attribute, Z is the standard normal variate, n is the number of sample size, and e is the error margin.

For finite population, we computed the sample size as

$$n = \frac{n}{1 + (n-1)/N}$$

where N is the size of the population.

For unknown p, we used 0.55 to produce the largest sample size. In this work, we set $p = 0.55$ for maximum sample size, and $N = 4097$, $e = 0.01$, and $Z = 2.58$for 99% confidence level.

Recent work suggested that spectrum features showed better performance for speech recognition tasks; thus, in this experiment, we have considered spectrogram-based low-level features from audio and EEG data for a depression diagnosis. In order to extract the spectrogram from the audio signal, we applied short-time Fourier Transform (STFT) and performed this on overlapping windows segments. We converted frequency (y-axis) to log scale and color dimension (amplitude) to generate spectrogram. We map the y-axis onto the mel scale to crate the mel spectrogram. We applied STFT to extract spectrograms from EEG signals, similar to audio data.

B. Multimodal TransformerNet

In this section, we have presented two different framworks (A) Vision Transformer (B) different pre-trained CNN networks such as ResNet, DensNet, EffecieNet. Fig. 3 illustrate the structure of transformer-based framework. We can notice that there are three main components EEG spectrogram, speech spectrogram, and 1D-CNN with LSTM. We have used temporal correlation, or correlation between two time points, that captures the time-domain information. 1D-CNN followed by LSTM to extract features from 1D time series sequential EEG Data. 1D-CNN is used to 1D-CNN is to distillate the features from time-series EEG data and reduce data length. The extracted CNN features are fed to LSTM to learn features from EEG time series data.

Recent work suggested that the spectrum feature showed better performance for the speech recognition task. Besides, we know that the behavior driven by the EEG and speech is a complete process, so our method could be more effective to utilize the relationship speech and EEG for identification of early stage depression. Thus, we have considered spectrogram-based low-level features from audio and EEG data for a depression diagnosis in this experiment. In order to extract spectrogram from the audio signal, we applied short-time Fourier Transform (STFT) and performed this on overlapping windows segments. We converted frequency (y-axis) to log scale and color dimension (amplitude) to generate spectrogram. We map the y-axis onto the mel scale to create the mel spectrogram. We applied STFT to extract spectrograms from EEG signals, similar to audio data. Before feeding the sequence of patches to the network, we projected linearly to vector with dimension d using the learned embedding matrix. We then concatenated the embeddings together along with a linearizable classification token.

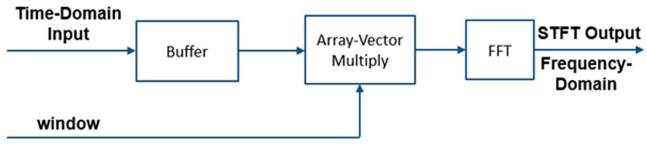


Fig. 4. EEG/Speech signals at Spectrogram conversion.

Unlike CNN, Vision Transformers cannot capture the sequence ordering of input tokens and require position embeddings. A learnable 1D relative-dimension vector, learnable position encoding is adopted in this work [25]. The relative relation is presumably important in EEG signals as relative ordering or of the elements matters significantly. We have used cross method that encodes different relative positions to same embedding based on distance i.e. if one direction is identical, either vertical or horizontal. The cross embedding can be represented as

$$r_{ij} = p^x_{I(i,j)} + P^y_{I(i;j)} \tag{1}$$

$$I^x i, j = g(x'_i - x'_j) \tag{2}$$

$$I^y i, j = g(y'_i - y_j{}') \tag{3}$$

where $p^x_I(i,j), P^y_{I(i,j)}$ are learnable scalars in bias in contextual mode.

Finally, we fed these tokens to the network as a set of patches without the notion of order. In order to maintain the spatial arrangement of patches, the patches arrangements are aided with positional information. We have converted the EEG audio signal to spectrogram and computed the literally tokens aided with positional information, similar to EEG feature extraction. The resultant embedded sequence with a token for EEG and speech can be written as

$$Z_e = [V_e; E_1EE; E_2EE; \ldots E_{nn}EE] + Epos \tag{4}$$

$$Z_s = [V_s; S_1EE; S_2EE; \ldots S_{nn}EE] + Epos \tag{5}$$

Equations (8) and (5) shows that 1D position encoding preserve the position information of flatten patches for EEG and speech signals respectively. We passed the resultant embedding sequence consisting of position information (Z_0) from both speech and EEG signals to the encoder. The encoder consists of L identical layers and each layer has two main components multi-head self-attention and fully-connected block. The fully connected forward block consists of two dense layers with GeLU activation. Notice that both multi-head self-attention and fully-connected block in transformer consist of residual skip connections which are followed by normalization layer and can be written as

$$Z'_\ell = MSA\,(LN(Z_\ell - 1) + (Z_\ell - 1)) \quad where \quad \ell = 1 \ldots L \tag{6}$$

$$Z_\ell = MSA\,(LN(Z'_\ell) + (Z'_\ell)) \quad where \quad \ell = 1 \ldots L \tag{7}$$

Notice that multi-head self-attention (MSA) is the core component that determines the role of a single EEG patch with respect to other patches. MSA consists of four layers i.e., linear, self-attention, concatenation, and final layer. We passed the

TABLE I
ABLATION STUDY

	Accuracy	Percision	Recall	F1 Score	Accuracy (Normal)	Accuracy (Depressed)
1D EEG Signals	0.652	0.652	0.682	0.636	0.682	0.717
EEG Transformer	0.7203	0.728	0.732	0.623	0.705	0.705
1D EEG + EEG Transformer	0.782	0.784	0.692	0.749	0.676	0.722
Speech	0.623	0.591	0.625	0.633	0.631	0.712
1DEEG + Tranformer (EEG+Speech)	**0.9731**	**0.9771**	**0.9734**	**0.9730**	**0.9797**	**0.9724**

TABLE II
DEPRESSION DIAGNOSIS USING PRE-TRAINED ON EEG DATASET

	Accuracy	Percision	Recall	F1 Score	Accuracy (Normal)	Accuracy (Depressed)
DensNet	0.8825	0.8731	0.8128	0.8444	0.8191	0.8321
EFNetV1	0.8803	0.8198	0.7952	0.8194	0.7931	0.7909
effnetv2m	0.7991	0.8496	0.7964	0.7919	0.7694	0.7819
effnetv2s	0.7902	0.7002	0.7053	0.7741	0.8381	0.8174
SENet	0.8215	0.8283	0.7891	0.7858	0.7842	0.7695
mobilNe	0.7156	0.7012	0.7143	0.7106	0.7278	0.7019

external element Z_ℓ to the external multi-head attention model. The high level of attention is represented by attention weight which is computed base don sum of all values of the sequence Z_ℓ. The self-attention head learns the attention weight by computing the scaling of the querying-key value. There are three values key, value, and query for each input by multiplying the element against value, query, and key matrices. The scaling dot product operation in the self-learning attention block is the same as dot product but also incorporates the values of dimension D_k as the scaling vector. Finally, the value of each patch embedding is multiplied by the softmax output to find the patch with the highest score.

$$[Query, Key, Value] = ZU, U \in^{d*3D} \tag{8}$$

Similarly to the transformer framework, we have replaced the transformer with pre-trained networks such as EfficientNet, DenseNet, SENet, and MobiNet. Fig. 4 illustrates the structure of the 2D pre-trained network.

Finally, we combined the features extracted from all three components (EEG spectrogram, Speech Spectrogram, and 1DCNN-LSTM. The output of the EEG transformer encoder, speech spectrogram encoder and 1DCNN-LSTM yield a better feature set containing both space and time domain features. one-dimensional CNN basead LSTM efficiently learns the time-domain information at different sampling points. Besides, the transformers learn the spatial information among different channels. Finally, we have fused the global information in the representation follwed by fully-connected layer to obtain the final classification and cross-entropy and sigmoid are used as loss function and activation function.

$$L = -\frac{1}{n}\sum_{n=1}^{N}\sum_{c=1}^{C} y_n^c log(\hat{y}_n^c) \tag{9}$$

where C and N denotes the number of categories and the number of batch sizes respectively. y_n^c is the true one hot label and $\hat{y}c_n$ is the predicted probability of the corresponding category.

IV. EXPERIMENT

In this section, we presented in depth analysis of the proposed framework and compared the performance with state-of-the-art networks. As described in Section III, we have combined speech and EEG features extracted from the spectrogram using the vision transformer with 1D learnable position encoding. We performed 5-fold cross-validation on the benchmark multimodal speech and EEG dataset for further analysis. Furthermore, we have also performed ablation study on both pre-trained networks and vision tranformer (see Table I and Table II). In the following discussion, we first describe the dataset detail and network parameters, followed by results, and finally compare the performance with SOTA. The proposed models were tested using 20 percent of the available EEG dataset. Several training hyperparameters need to be optimized, such as optimizers, learning rate, and loss functions. We have used Adam optimizer with a learning rate of 0.0004 and used 70 epochs for training with batch size 100. We have applied binary cross-entropy loss function and Sigmoid activation function for the classification layer.

A. Dataset

EEG provides a non-invasive analysis which can assist in diagnosis of neuropsychiatric and psychiatric conditions. We have performed experiments on benchmark Multi-modal Open Dataset for Mental-disorder Analysis (MODMA). MODMA was collected in a quiet, clean, soundproof, and no electromagnetic interference room from 52 subjects (29 matching normal controls people- 16 males and seven females and 23 clinically depressed patients - 20 males and nine females) which were selected by professional psychiatrists and carefully diagnosed. The speech data was recorded while reading, picture description, and interviewing, and EEG data were collected in resting state and under-stimulation. Further description of dataset can be found at [5], [17].

B. Results

In this section, we evaluate the performance of the proposed framework under different parameters. As mentioned earlier,

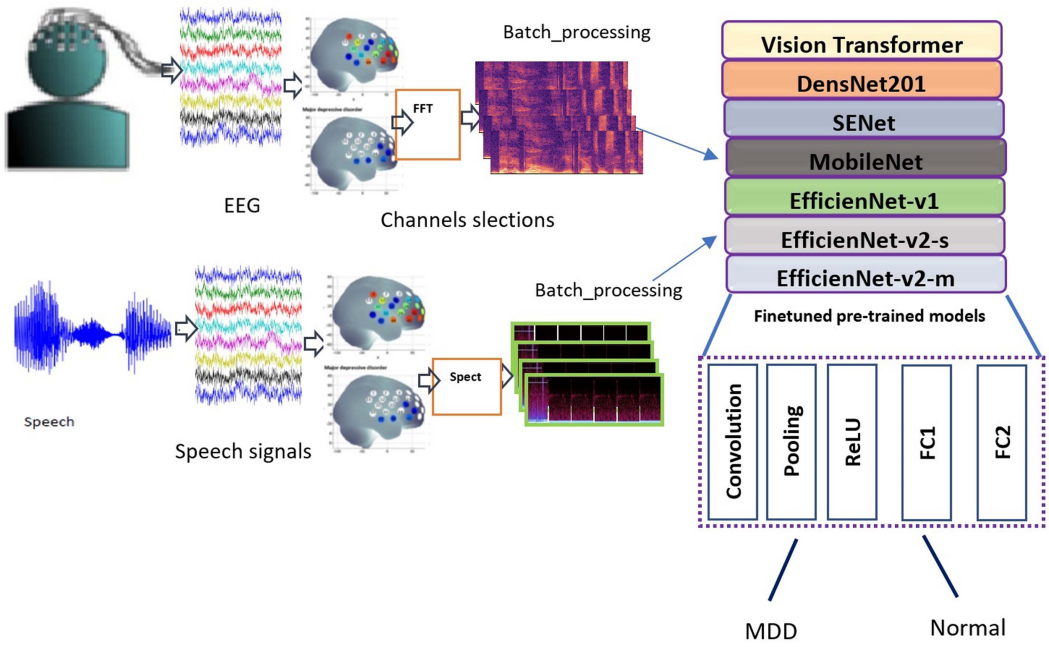


Fig. 5. EEG/Speech Spectrogram based 2D Nework.

TABLE III
DEPRESSION DIAGNOSIS USING PRE-TRAINED ON SPEECH DATASET

	Accuracy	Percision	Recall	F1 Score	Accuracy (Normal)	Accuracy (Depressed)
DensNet	0.8476	0.8318	0.8652	0.8308	0.8397	0.8446
EFNetV1	0.753642	0.81676	0.8177	0.7924	0.74911	0.7944
Effnetv2m	0.8141	0.82683	0.8014	0.7912	0.7958	0.6999
Effnetv2s	0.9031	0.8192	0.7936	0.7222	0.7449	0.6974
SENet	0.7942	0.8110	0.7129	0.7846	0.7916	0.7698
MobileNets	0.7889	0.7881	0.7190	0.7807	0.7871	0.7067

we have performed two different experiments using different pretrained networks and vision transformer. To perform an ablation study, we first performed the experiments on individual networks in our first experiment. To do so, we, first, have developed a pre-trained multimodal network based on state-of-the-art networks such as DenseNet, EffNetV2M, EffNetVS, SeNet and MobileNet. Fig. 5 depicts the architecture of network. We have fed these models with a spectrogram and fused the extracted features to generate descriptive features, which are then fed to fully connected layer. Table III exhibits the results of 2-dimension EEG-based depression diagnosis using different networks. Notice that, DesneNet showed slightly better overall performance (88.25%, 87.31%, 81.28%, and 84.44% accuracy, precision, recall, and F1 score respectively) and significantly better performance than other networks (EffNetV2M, EffNetVS, SeNet, and MobileNet). We can observe that DenseNet showed slightly poor performance in differentiating depressed people. Similarly, we have used a pretrained approach using a different famous network such as DenseNEt, EffNetV2M, EffNetVS, SeNet, and MobileNet on speech modality only. Table III exhibits the results of 2-dimension speech-based depression diagnosis using different pretrained networks on speech modality only. Similar to EEG modality, DesneNet showed slightly better overall performance (84.76%, 83.18%, 86.52 and 83.08% accuracy, precision, recall, and F1 score respectively) and significantly better performance than another network (EffNetV2M, EffNetVS, SeNet, and MobileNet). We can observe that DenseNet showed slightly poor performance in differentiating depressed people.

In our second experiment, we have developed different pretrained networks, as illustrated in Fig. 5. We also applied a vision transformer instead of a pretrained network. We have fed extracted features to generate descriptive features fed to a fully connected layer. Table IV exhibits the results of transformer-based depression diagnosis. We can notice that the transformer showed slightly better performance than the pre-trained network. Furthermore, we can notice that the pre-trained network showed better performance than the basis network.

C. Discussion

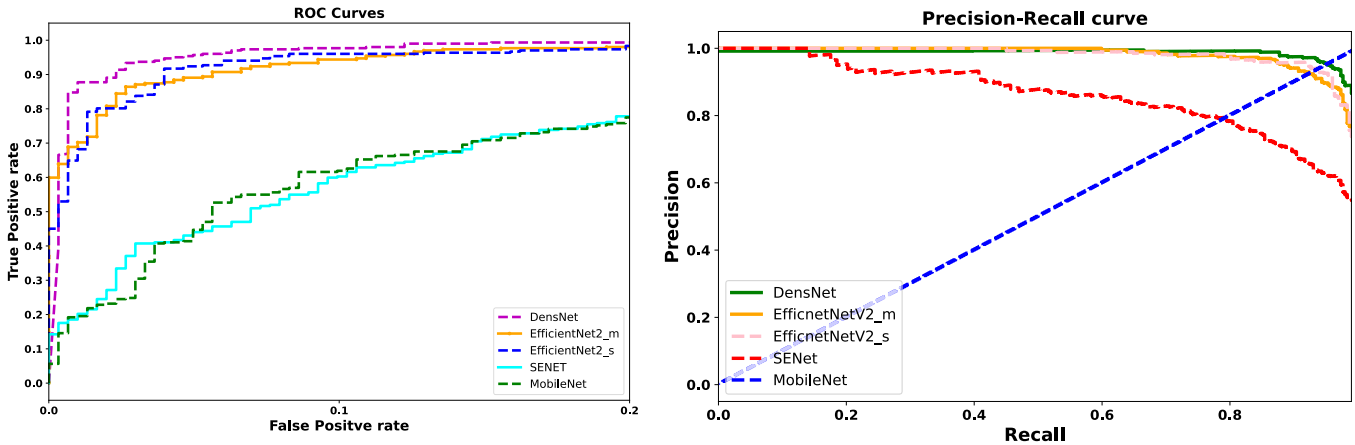
This section analyze and compare the effectiveness of proposed framework over SOTA methods such as 1D plus 2D CNN LSTM [27], Self attention + multitask learning [15], Temporal Convolution Network [24] and decision tree [8]. We have used precision, recall, and F1 score for evaluation as standards for comparison. Zhao et al. extracted emotionally salient regions, and frame-level features were extracted to store temporal information. The extracted temporal feature set is then forwarded to

TABLE IV
DIAGNOSTIC PERFORMANCE ON PROPOSED FRAMEWORK USING DIFFERENT NETWORK STRUCTURE ON MULITMODAL SPEECH AND EEG DATASET

	Accuracy	Percision	Recall	F1 Score	Accuracy (Normal)	Accuracy (Depressed)
Proposed (Transformer)	**0.9731**	**0.9771**	**0.9734**	**0.9730**	**0.9797**	**0.9724**
DensNet (Pre-trained)	0.9636	0.9516	0.9561	0.9624	0.9411	0.9674
effnetv2m (Pre-trained)	0.9621	0.9283	0.9114	0.9112	0.958	0.8799
effnetv2s (Pre-trained)	0.9307	0.9292	0.9176	0.9392	0.9644	0.8997
SENet (Pre-trained)	0.81942	0.8066	0.7729	0.7846	0.9176	0.8098
mobilNe (Pre-trained)	0.83889	0.7881	0.779419	0.7807	0.8071	0.8167
DensNet	0.9542	0.8516	0.87577	0.91324	0.8511	0.8444
effnetv2m	0.9241	0.8983	0.8114	0.8912	0.8158	0.7699
effnetv2s	0.9497	0.8992	0.9007	0.8222	0.79497	0.7774
SENet	0.7902	0.8116	0.7781	0.7046	0.7716	0.7098
mobilNe	0.7879	0.7781	0.7794	0.7217	0.7520	0.7777

TABLE V
PROPOSED

	Approach	Precision	Recall	F1 Score
Proposed	Vision Transformer	**0.972**	**0.973**	**0.973**
Proposed	DenseNet-pretrained	0.964	0.952	0.961
Zhao et al. [28]	1D + 2D CNN LSTM	0.929	0.935	0.932
Li et al. [15]	Self attention + multitask learning	0.935	0.901	0.918
Wang et al. [24]	Temporal Convolution Network	0.916	0.911	0.905
Chen et al. [8]	Decision Tree	0.834	0.819	0.805



(a) (b)

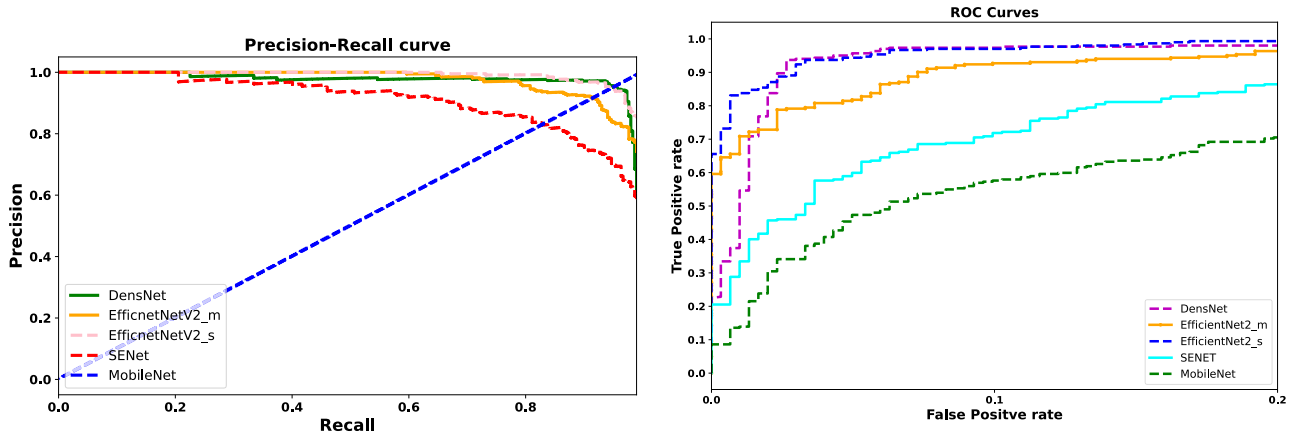
Fig. 6. Diagnostic performance (a) Precision Recall (b) ROC plot on EEG dataset.

multihead time-dimension attention LSTM. Lin et al. applied ID CNN and bidirectional LSTM with an attention layer to deal with linguistic contents. Besides, audio and text features are utilized simultaneously, which reduces the impact of patients' misleading information.

Table V describes the comparative results of the proposed framework with state of the art network. Notice that our proposed framework (both transformer and pre-trained network-based) showed significantly better performance. We have achieved 0.972 (Precision), 0.972 (Recall) and 0.973 (F1-Score). It may be due to the use of both 2D and 1D features, which proves that fusion of 2D-based speech and spectrogram-based features aided with position information significantly impacts diagnostic performance. It is noteworthy that our method achieved significantly better performance and we have deployed it as a tool that clinicians can use to diagnose depression. Figs. 6, 7, and 8 describes the comparative evaluation of proposed framework with benchmark methods.

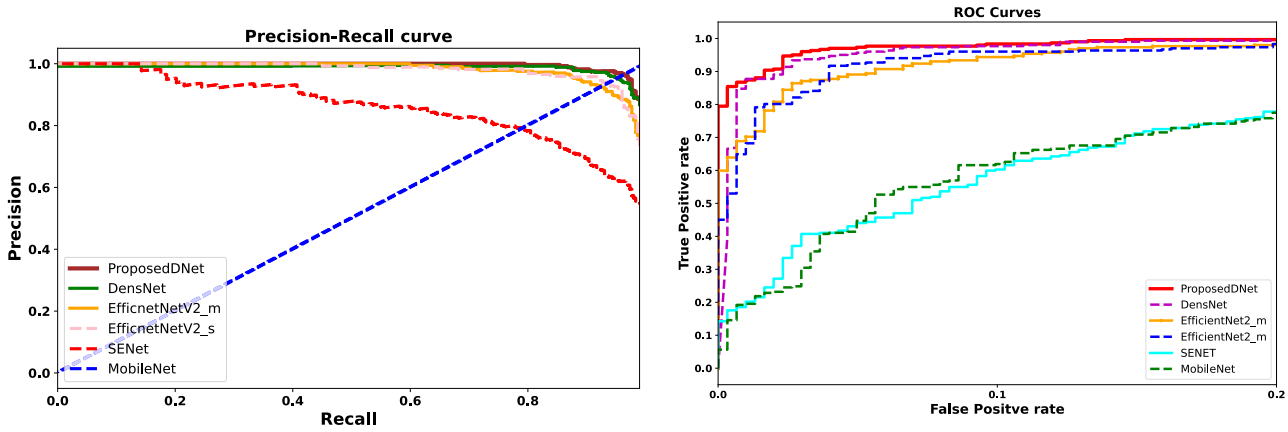
D. Deployment

To facilitate the clinicians, we have implemented and deployed as a depression diagnostic tool. The proposed deep learning models are trained well on depression and healthy training datasets. Beside, we have provide the weights of trained models which can be integrated in tool to build a tool for depression detection. The following number of steps are already performed for tool development and deployment.1) loading the saved the weight,2) pre-process the testing data 3) perform prediction on test set and the prediction response. Further steps are required to



(a) (b)

Fig. 7. Diagnostic performance (a) Precision Recall (b) ROC plot on Speech dataset.



(a) (b)

Fig. 8. Diagnostic performance (a) Precision Recall (b) ROC plot on Multimodal dataset.

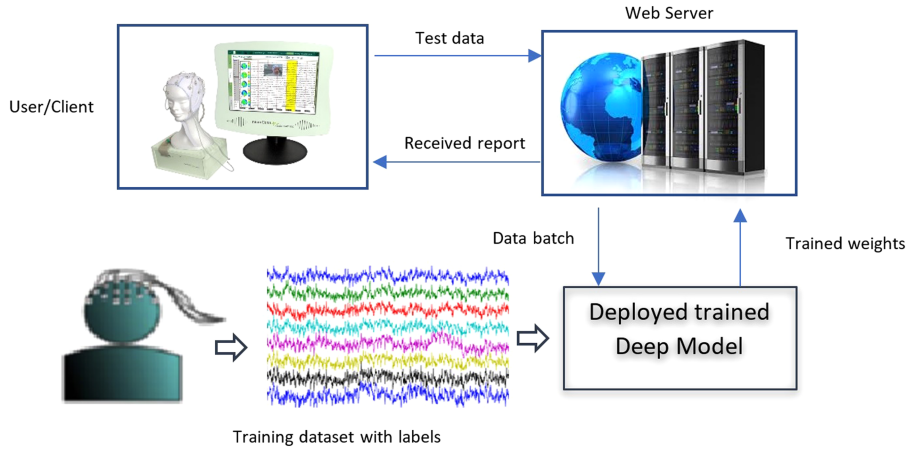


Fig. 9. EEG/Speech signals at Spectrogram conversion.

be taken to obtain the depression assessment results on the test dataset remotely. The trained weights are loaded as well as all model related dependencies are included in the web server using light-weight web framework. The framework can be used or integrated with vairous platforms. The manual method to check depression is based on the questionnaires and some physical scales used inside the hospital. The framework can be deployed remotely through a web server. The testing EEG signals obtained from the patients will be sent to the webserver deployed at remote location. We have deployed the trained model on web server which can be used for prediction of depressed patient. The system will take input of data, predict it and will send the

patient report back to the clinic. Fig. 8 illustrate the protocol for web based depression detection.

V. Conclusion

In this work, we presented an end-to-end deep framework for the effective detection of depression using speech and Electroencephalograph signals. We use multiple frequencies of EEG signals along with speech signals to recognize motion intentions. Thus, we are not only able to classify the specific intention but also cater for multiple tasks recognition. We combine different level features extracted from EEG and audio spectrogram and raw EEG to generate descriptive features and applied vision transformer on speech and EEG spectrum, which significantly improves diagnosis performance. Experiments were conducted on a benchmark MODMA dataset consisting of EEG and audio data collected from 52 subjects (both clinically depressed patients and normal control). Results showed that the proposed framework achieved significantly better performance in discriminating mild depression disorder from normal compared to state-of-the-art methods. Finally, We deployed our trained model using Flask, a lightweight web framework.

References

[1] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, 2018.

[2] J. J. Allen, W. G. Iacono, R. A. Depue, and P. Arbisi, "Regional electroencephalographic asymmetries in bipolar seasonal affective disorder before and after exposure to bright light," *Biol. Psychiatry*, vol. 33, no. 8/9, pp. 642–646, 1993.

[3] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Front. Neurosci.*, vol. 6, 2012, Art. no. 39.

[4] B. Ay et al., "Automated depression detection using deep representation and sequence learning with EEG signals," *J. Med. Syst.*, vol. 43, no. 7, 2019, Art. no. 205.

[5] H. Cai et al., "Modma dataset: A multi-modal open dataset for mental-disorder analysis," 2020, *arXiv:2002.09283*.

[6] J. Chen et al., "A transfer learning based super-resolution microscopy for biopsy slice images: The joint methods perspective," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 103–113, Jan./Feb. 2021.

[7] T. Chen et al., "Discriminative cervical lesion detection in colposcopic images with global class activation and local bin excitation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1411–1421, Apr. 2022.

[8] X. Chen and Z. Pan, "A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health," *Int. J. Environ. Res. Public Health* vol. 18, no. 12, 2021, Art. no. 6441.

[9] R. J. Davidson, "Anterior electrophysiological asymmetries, emotion, and depression: Conceptual and methodological conundrums," *Psychophysiol.*, vol. 35, no. 5, pp. 607–614, 1998.

[10] R. Feng, X. Liu, J. Chen, D. Z. Chen, H. Gao, and J. Wu, "A deep learning approach for colonoscopy pathology WSI analysis: Accurate segmentation and classification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3700–3708, Oct. 2021.

[11] H. Gao, K. Xu, M. Cao, J. Xiao, Q. Xu, and Y. Yin, "The deep features and attention mechanism-based method to dish healthcare under social IoT systems: An empirical study with a hand-deep local-global net," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 1, pp. 336–347, Feb. 2022.

[12] I. H. Gotlib, "EEG alpha asymmetry, depression, and cognitive functioning," *Cogn. Emotion*, vol. 12, no. 3, pp. 449–478, 1998.

[13] S. F. Husain et al., "Cortical haemodynamic response measured by functional near infrared spectroscopy during a verbal fluency task in patients with major depression and borderline personality disorder," *EBioMedicine*, vol. 51, 2020, Art. no. 102586.

[14] X. Li et al., "EEG-based mild depression recognition using convolutional neural network," *Med. Biol. Eng. Comput.*, vol. 57, no. 6, pp. 1341–1352, 2019.

[15] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. Interspeech Conf*, 2019, pp. 2803–2807.

[16] S. C. Liao, C. T. Wu, H. C. Huang, W. T. Cheng, and Y. H. Liu, "Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns," *Sensors*, vol. 17, no. 6, 2017, Art. no. 1385.

[17] Z. Liu, D. Wang, L. Zhang, and B. Hu, "A novel decision tree for depression recognition in speech," 2020, *arXiv:2002.12759*.

[18] M. Niu et al., "Common and specific abnormalities in cortical thickness in patients with major depressive and bipolar disorders," *EBioMedicine*, vol. 16, pp. 162–171, 2017.

[19] A. Qayyum, I. Razzak, and W. Mumtaz, "Hybrid deep shallow network for assessment of depression using electroencephalogram signals," in *Proc. Int. Conf. Neural Inf. Process.*, 2020, pp. 245–257.

[20] I. Razzak, M. Blumenstein, and G. Xu, "Multiclass support matrix machines by maximizing the inter-class margin for single trial EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1117–1127, Jun. 2019.

[21] I. Razzak, I. A. Hameed, and G. Xu, "Robust sparse representation and multiclass support matrix machines for the classification of motor imagery EEG signals," *IEEE J. Transl. Eng. health Med.*, vol. 7, pp. 1–8, 2019.

[22] M. I. Razzak, M. Imran, and G. Xu, "Big data analytics for preventive medicine," *Neural Comput. Appl.*, vol. 32, pp. 4417–4451, 2020.

[23] A. Rehman, S. Naz, and I. Razzak, "Leveraging Big Data analytics in healthcare enhancement: Trends, challenges and opportunities," *Multimedia Syst.*, vol. 28, pp. 1339–1371, 2022.

[24] Y. Wang, F. Liu, and L. Yang, "EEG-based depression recognition using intrinsic time-scale decomposition and temporal convolution network," in *Proc. 5th Int. Conf. Biol. Inf. Biomed. Eng.*, 2021, pp. 1–6.

[25] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10033–10041.

[26] X. Zhang, B. Hu, L. Zhou, P. Moore, and J. Chen, "An EEG based pervasive depression detection for females," in *Proc. Joint Int. Conf. Pervasive Comput. Networked World*, 2012, pp. 848–861.

[27] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1 D & 2 D CNN LStM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, 2019.

[28] X. Zhou, Y. Li, and W. Liang, "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 912–921, May-Jun. 2020.

[29] X. Zhou, W. Liang, K. I. K. Wang, and S. Shimizu, "Multi-modality behavioral influence analysis for personalized recommendations in health social media environment," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 5, pp. 888–897, Oct. 2019.

[30] X. Zhou, X. Xu, W. Liang, Z. Zeng, and Z. Yan, "Deep learning enhanced multi-target detection for end-edge-cloud surveillance in smart IoT," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12588–12596, Aug. 2021.

[31] H. Zogan, I. Razzak, and S. Jameel, and G. Xu, "Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 133–142.

[32] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "Hierarchical convolutional attention network for depression detection on social media and its impact during pandemic," *IEEE J. Biomed. Health Informat.*, early access, Feb. 9, 2023, doi: 10.1109/JBHI.2023.3243249.

[33] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-modalities using a hybrid deep learning model on social media," 2020, *arXiv:2007.02847*.

Abdul Qayyum received the PhD degree in electrical & electronics engineering with specializaion in deep learning and image processing from Universiti Teknologi PETRONAS, Malaysia, in 2017. He is currently working with the Department of Biomedical Engineering, Imaging Sciences Kings College London, U.K. Previosuly, he was joined as lecturer with the University of Bourgogne Franche-Comté France. His research interests include machine learning and deep learning for signal processing e.g., EEG, ECG, biomedical images (Cardiac MRI,CT, NCCT, Liver CT, Prostate, Kidney Tumor) segmentation, and classification.

Imran Razzak is a senior lecturer in human-centered machine learning with the School of Computer Science and Engineering, University of New South Wales, Sydney, Australia. Previously, he was a senior lecturer in computer science with Deakin University, Geelong and assistant professor with King Saud bin Abdulaziz University for Health Sciences (2013-2017). His research interests include focus on connecting language and vision for better interpretation and spans over three broad areas: machine learning, computer vision, and natural language processing with special emphasis on healthcare.

M. Tanveer is associate professor and ramanujan fellow with the Discipline of Mathematics of the IIT Indore. Prior to that, he worked as a postdoctoral research fellow with the Rolls-Royce @NTU Corporate Lab, NTU, Singapore. His research interests include support vector machines, optimization, machine learning, deep learning, applications to Alzheimer's disease and dementias. He has published more than 120 referred journal papers of international repute. His publications have more than 3550 citations with h index 31 (Google Scholar, March 2023). Recently, he has been listed in the world's top 2% scientists with the study carried out by Stanford University, USA. He is the recipient of the 2023 IIT Indore Best Research Paper Award, 2022 Asia Pacific Neural Network Society Young Researcher Award, 29th ICONIP 2022 Best Research Paper Award, 2017 SERB-Early Career Research Award in Engineering Sciences and the only recipient of 2016 DST-Ramanujan Fellowship in Mathematical Sciences. He is currently the Associate Editor in several prestigious journals including *IEEE Transactions on Neural Networks and Learning Systems*, *Pattern Recognition*, *Elsevier, Neural Networks*, *Elsevier, Engineering Applications of Artificial Intelligence, Elsevier, Neurocomputing, Elsevier, Cognitive Computation, Applied Soft Computing, Elsevier*. Amongst other distinguished, international conference chairing roles, he is the General Chair for 29th ICONIP2022. He is currently the Principal Investigator (PI) or Co-PI of 12 major research projects funded by Government of India.

Moona Mazher received the master's degree in electrical and electronics engineering with specialization in biomedical sciences (neuroscience) from Universiti Teknologi PETRONAS, Malaysia, in 2017. She is currently working toward the PhD degree in 'brain tumor segmentation & prognosis analysis' with the Department of Computer Engineering and Mathematics, University Rovira i Virgili, Spain. She has been actively participating with the MICCAI (Medical Image Computing and Computer Assisted Interventions) and has secured top five and even 1st position in many MICCAI challenges. Her research interests include machine learning, deep learning, medical imaging and signal processing, computer vision, and explainable AI.

Bandar Alhaqbani received the BS degree with top honor in computer engineering, the master's degree in information technology (E-Business), and the PhD degree in information technology (information security). Currently, he is the CEO in Technology Control Company (TCC). Prior to that, he used to be the COO in Technology Control Company (TCC), executive director of Digital Security within Technology TCC, general director of IT Services within KSAU-HS, and head of Advanced Computing and Technologies within KAIMRC. In addition to his industrial vast experiences, he has served as the chairman of health informatics department within the College of Public Health and Health Informatics in KSAU-HS and is an adjunct assistant professor where he teaches IT and information security courses. He is a co-founder of the Health Information System Bachelor program. In recognition to his health informatics experiences, he has been elected as the president for Saudi Association for Health Informatics.