IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 July 2024, accepted 26 July 2024, date of publication 1 August 2024, date of current version 20 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3436895

RESEARCH ARTICLE

A Low-Complexity Combined Encoder-LSTM-Attention Networks for EEG-based Depression Detection

NOOR FARIS ALI[1,2], NABIL ALBASTAKI[1], (Member, IEEE), ABDELKADER NASREDDINE BELKACEM[3], (Senior Member, IEEE), IBRAHIM M. ELFADEL[4,5], (Life Senior Member, IEEE), AND MOHAMED ATEF[1], (Senior Member, IEEE)

[1]Electrical and Communication Engineering Department, College of Engineering, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates
[2]Biomedical Engineering Program, College of Engineering, American University of Sharjah, Sharjah, United Arab Emirates
[3]Department of Computer and Network Engineering, College of IT, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates
[4]Center for Cyber Physical Systems, Khalifa University, Abu Dhabi, United Arab Emirates
[5]Department of Computer and Communication Engineering, Khalifa University, Abu Dhabi, United Arab Emirates

Corresponding author: Mohamed Atef (moh_atef@uaeu.ac.ae)

This work was supported by the Khalifa University (KU)-United Arab Emirates University (UAEU) Joint Research Program under Award KU-UAEU-2023-004 (UAEU grant code G00004680).

ABSTRACT Despite the high performance of existing state-of-the-art deep learning models for depression detection using electroencephalography (EEG), they incur a heavy computational burden. In this paper, we propose an efficient model consisting of a cascade of an encoder, long short-term memory (LSTM), and attention mechanism networks. The encoder compresses data into a lower-dimensional latent space. The LSTM models the temporal variations in brain rhythms. The attention mechanism rectifies the problem of compressed data in sequence-to-sequence models and efficiently leverages parallelism. Compared with recent state-of-the-art, our proposed depression detection model shows better performance and efficiency with a validation accuracy of 99.57% on subject-dependent experiment and a testing accuracy of 84.93% on subject-independent experiment with a total number of 4,355 parameters. The proposed model has resulted in 99.65% reduction in complexity compared with the state-of-the-art EEG-based depression detection models. The results of this study indicate the effectiveness of the proposed model design and the usefulness of the combined encoder, LSTM, and attention modules. These networks serve as mitigating factors for the computational load, which is vital for future research on multi-tasking mental health monitoring using AI-enabled EEG wearables.

INDEX TERMS Depression detection, EEG, LSTM, encoder, attention mechanism, EEG wearables, efficient deep learning.

I. INTRODUCTION

Depression or major depressive disorder (MDD) affects mental and cognitive capabilities and performance, social life, and psychological and emotional stability [1], [2].

Researchers have drawn more attention and made remarkable efforts to extract important information from neurophysiological signals such as the electroencephalogram (EEG) using advanced Deep Learning for depression diagnosis [3], [4], [5]. Using EEG and Deep Learning has led to significant advances in psychiatry towards objective diagnosis and tailored treatment plans based on neural pathologies in the brain detected by EEG [5].

Considerable research has been done on EEG biomarkers for depression diagnosis to identify which EEG frequency bands and channels are more affected in depressed subjects. The temporal and frontal channels are significant for MDD identification [6]. The review presented in [7] reported an

The associate editor coordinating the review of this manuscript and approving it for publication was Venkata Rajesh Pamula.

increase in the absolute power for both *theta* and *beta* bands in depressed people. *Alpha* and *theta* bands are useful differentiators between depressed and healthy controls [8], [9], providing confirmatory evidence that these bands are related to emotional processing [5]. Another review [5] reported that a*lpha* and *beta* were related to anxiety whereas *gamma*, generally, was related to sensory processing and might be related to mood swings. The consensus view of these studies is that the *theta* band plays a significant diagnostic role in depression [5], [7], [8], [9]. *Theta* is believed to reflect activity from the limbic system and hippocampal regions [10] which are known to be impacted by depressive states [1]. *Gamma* could be a potential diagnostic biomarker [5] as well, while *Beta* seems more related to anxiety and ruminating thinking, which can co-exist with depression. However, it cannot provide a differential diagnosis. Studies on the *alpha* band reported conflicting and inconsistent results, but overall, *alpha* asymmetry offers a prognostic biomarker [5].

A. RELATED WORK

CNN-LSTM is one of the most popular and accurate models for depression detection using EEG signals [3]. The study in [11] has proposed a 1D-CNN model for depression diagnosis. The model comprises 5 convolutional layers, 5 batch normalization layers, 5 pooling layers, and 2 fully connected (FC) hidden layers. The model accepts 19-channel EEG-time series of 4-second windows as input. The objective was to extract features from raw EEG signals, specifically spatial information. The total number of parameters (TNP) of the model is 363,882 with achieved accuracy of 99.37%.

In another study [12], the authors proposed a combined 1D-CNN-LSTM model to extract features from a 64-channel EEG. The EEG data was windowed, and FFT was applied to extract time-frequency information. The model consists of 1D-CNN layer, 2 LSTM layers, and 2 FC layers. The TNP is 46,658 with an accuracy of 99.10%.

Another study [13] proposed a 1D-CNN-LSTM model with 4 CNN layers and 1 Max Pooling layer, 1 LSTM layer, and 1 FC layer. EEG were recorded from the left (Fp1-T3) and right (Fp2-T4) hemispheres. The developed model has a TNP of 1,276,989 with an accuracy of 99.12%.

The work in [14] proposed CNN-GRU-Attention model containing 2 1D-CNN layers, 1 Max Pooling, 1 GRU layer, and 1 Attention (ATTN) layer. The signals from each 16-EEG channels were windowed into 1-second segments. Then, power spectral density (PSD) was computed from each segment using the Welch method to extract the 5 bands. It was shown that Attention enables the accuracy to reach 99.33%.

B. RESEARCH GAPS

Within the past decade, the emergence of low-cost commercial devices and wearables for monitoring mental health has received substantial interest [15]. Currently, wearable EEG systems with AI for depression detection are being investigated for personalized screening at early curable stages [16]. In addition, they are being developed to monitor treatment responses and enable biofeedback therapy in wearables [17].

Embedding AI and deep learning models in microcontrollers is challenging due to the MCU's restricted capabilities [18], [19]. These constraints appear more difficult in multi-tasking cases such as the combination of ECG, EEG, and electrodermal activity for stress, anxiety, and depression identification and for examining comorbidities [15]. This is also the case when scoring depression severity and monitoring treatment progress. In real-time scenarios, multiple signals synchronization, complex models, and large data flow are all problematic factors in terms of the memory and energy requirements of the targeted wearable design.

Several deep learning models have focused on enhancing the baseline accuracy through increasing model complexity and TNP [20]. While this may result in high-accuracy models, it is likely to fall short of latency, throughout, and power targets for real-time deployment. A singular complex model can be accommodated in tiny devices but embedding several models on one single platform will exhaust its resources may be of restricted functionality. According to a review on efficient deep learning [20], efficiency can be formulated using metrics or indicators for model quality (e.g., accuracy), footprint (e.g., memory), and performance (e.g., latency). Efficiency can be tackled through compression, sparsification, quantization, and pruning algorithms, automation techniques (e.g., architecture search and hyperparameter optimization), and streamlined architecture design.

Efficient architecture design is achieved through bottom-up custom design driven by the core efficiency metrics (e.g., TPN, number of submodels, etc.) while searching for optimal architectures under accuracy and F-score constraints. In our work, we follow this custom, bottom-up paradigm.

Despite the competitive performance of the existing state-of-the-art models for depression detection using EEG, they are computationally expensive given their large TNPs, ranging from 40,000 to 1,300,000 [11], [12], [13]. CNN and LSTM are resource-intensive networks as they require a large memory footprint and significant power [21]. It is very challenging to implement such complex networks on tiny hardware like the Arduino Nano 33 BLE Sense (SRAM: 256 KB) [22].

To address the computational cost burden of running large-scale deep learning models, optimization and compression algorithms have been developed to reduce the TNPs of these models using pruning and quantization [20], [21]. However, these methods could cause an imbalance in the size of the subnetworks and thus degrade the overall model's accuracy if not performed properly. For instance, structured pruning leads to an accuracy loss because of the continuous removal of CNN filters [21].

C. THE PAPER CONTRIBUTIONS

To cope with this problem, we propose a novel efficient model under constrained resources (TNP and memory requirements), while retaining competitive accuracy. The model

first uses a pre-trained Encoder to compress the input into a lower-dimensional code offering a compact summary and latent-space representation of the input features.

The encoder is cascaded with a model consisting of two sub-networks: LSTM and Transformer. The LSTM's role is to model the temporal dependencies of the extracted features (related to brain activities) along sequential timesteps. The Transformer's role is to provide a self-sequence temporal attention mechanism to leverage the output of the LSTM compressed sequence by assigning weights to the LSTM hidden states. The goal is to better capture the temporal properties extracted by the LSTM.

For efficiency and performance gains, the encoder bottleneck architecture and attention mechanism have been widely adopted and proven very capable. The encoder efficiently learns important features from input data to best fit in the available minimized space of the bottleneck [23]. The attention mechanism has the potential of rectifying the problem of compressed data in sequence-to-sequence models [20]. The LSTM is recognized as the state-of-the art network in time-series applications (e.g., EEG for depression classification) [24].

In summary the contributions of this paper are as follows:

1) We provide a low-complexity, low-footprint model (TN = 4,355, memory = 126 KB in.h5 format) with competitive validation accuracy (99.57%) and average testing accuracy on 32 subjects (84.93%) compared with the state-of-the-art CNN-LSTM models for MDD detection.
2) The proposed model is shown to be very promising for future research towards multi-modal, AI-enabled mental health monitoring wearables, avoiding model proliferation on a single resource-restricted device.
3) We design, implement, and evaluate an encoder bottleneck architecture that achieves significant input dimensionality reduction using channels selection with no noticeable impact on EEG classification accuracy.
4) We leverage an attention mechanism to better capture temporal information in LSTM latent states.

II. METHODOLOGY

The main focus of this research is to carry out multiple experiments in an iterative process by gradually decreasing complexity, targeting better allocation of model's parameters under optimal architecture and accuracy, and verifying the proposed Encoder-LSTM-ATTN model's performance and efficiency compared to state-of-art models. First, we use a pre- and post-processing stages for EEG denoising, feature extraction, and channel selection. Second, we create efficient state-of-the-art models, including CNN, LSTM, and CNN-LSTM models as reference models for our experiments. Third, we develop a novel model involving an Encoder bottleneck architecture to compress the input data size, along with an LSTM network having an Attention mechanism to efficiently enhance accuracy.

The first step is to select the dataset to train and test the created models for depression classification. The dataset has been acquired from the MDD Patients and Healthy Controls EEG Data database [25]. The original dataset comprised 34 MDD outpatients (17 males and 17 females, mean age = 40.3 ± 12.9) and 30 healthy controls (HC) (21 males and 9 females, mean age = 38.3 ± 15.6). After conducting a quality examination, corrupted data, such as EEG recordings with less than 5 minutes duration and poor signal quality after pre-processing, were excluded. The eye-closed resting state EEG recordings were acquired using an EEG cap which consists of 19 electro-gel sensors based on the standard international 10-20 system for EEG electrodes placement. Accordingly, the EEG electrodes were distributed on the scalp in different regions: frontal (7 electrodes), central (3 electrodes), parietal (3 electrodes), occipital lobe (2 electrodes) and temporal lobe (4 electrodes). The MDD patients were diagnosed using the internationally recognized diagnostic criteria for depression; Diagnostic and Statistical Manual-IV (DSM-IV).

Fig. 1 displays the workflow of EEG processing and features extraction including the pre- and post-processing steps applied on the recorded 19-channel EEG signals and the deep learning implementation.

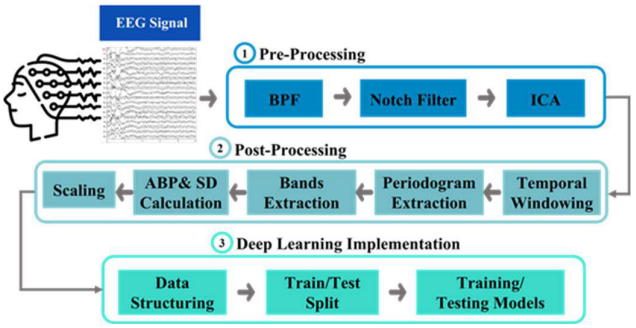

FIGURE 1. EEG Features Extraction and Training.

A. EEG PREPROCESSING STEPS

EEG signals are susceptible to motion artifacts and other noise sources. ECG (0.05 to 150 Hz) and EOG (dc-10 Hz) bandwidths overlap with EEG signal bandwidth (0.5 to 45 Hz). Also, the high frequency components of EMG signals can interfere with the EEG signal.

Therefore, we apply an 8th order Butterworth IIR bandpass filter (BPF) with a low pass cutoff at 0.5 Hz and high pass cutoff at 45 Hz. A notch filter was used to suppress the 50 Hz power line interference.

EOG and EMG interferences remain an issue as they still coincide with the filtered signal frequency band. Therefore, we applied Independent Component Analysis (ICA) to decompose the interfered EEG signal into additive subcomponents and subsequently eliminate artifactual EOG and EMG components [26].

B. POST-PROCESSING

Temporal windowing facilitates the detection of time-varying brain rhythms related to emotional processing and therefore depression identification [27], [28], [29]. On this basis, we partitioned the 5-min (300-second) EEG time-series $x(m)$ into sliding $T = 30$ windows in each channel with a length of 10 seconds for each window.

Welch's method (or averaged periodogram) was employed for estimating the power spectrum in each window. In this method, the window is divided into successive overlapping segments, and the periodogram for each segment is formed through averaging.

For each window $K_i(m)$, we extracted the periodogram using the Welch's method by dividing the window further into smaller overlapping segments using a Hamming window $\omega(m)$ of 1024 samples at a sampling frequency of 256 Hz. Thus, the periodogram $I_i(\omega)$ of each temporal i^{th} window $K_i(m)$ is given by:

$$I_i(\omega) = \frac{1}{u}\left|\sum_{m=0}^{M-1} K_i(m)\omega(m)\, e^{-j\omega m}\right|^2 \tag{1}$$

$$u = \frac{1}{M}\sum_{m=0}^{M-1}\omega^2(m) \tag{2}$$

For each subject, we have 19 channels and $T = 30$ windows, so overall, we have 570 windows and the obtained periodograms $I_i(\omega)$, each is of size 513 with a frequency range 0 to 128 (Nyquist frequency) and a frequency resolution of 0.25 Hz.

The EEG signal can be decomposed into five frequency bands, each band has a particular frequency range, *delta D* (0.5 to 4 Hz), *theta T* (4 to 7 Hz), *alpha A* (8 to 13 Hz), *beta B* (14 to 30 Hz), and *gamma G* (>30 Hz). We defined these bands, using the estimated PSD, to be used for computing the absolute band power in each window for each channel and the corresponding standard deviation.

The absolute band power (ABP) is calculated by integrating the power contribution of each identified frequency band in each obtained periodogram using Simpson integration rule [30]:

$$\begin{aligned}\mathrm{ABP} = \int_{\mathrm{fa}}^{\mathrm{fb}} I_i(\mathrm{f})\,\mathrm{df} = &\frac{\Delta f}{3}(I_i(\mathrm{f_a}) + 4I_i(\mathrm{f_1}) + 2I_i(f_2) \\ &+ \ldots 2I_i(\mathrm{f_{n-2}}) + 4I_i(\mathrm{f_{n-1}}) + I_i(\mathrm{f_b}))\end{aligned} \tag{3}$$

We integrate over the produced periodogram I_i of a certain window, taking all the power values with respect to the frequency limits f_a and f_b of a recognized frequency band f.

Another potentially effective feature is the standard deviation (SD) of band power distribution in each periodogram. It conveys information about variations in the power values and hence changes in band-specific brain activity within a particular window. The standard deviation is calculated using:

$$\mathrm{SD} = \sqrt{\frac{\sum_{j=1}^{N}\left(P_j(f) - P(F)\right)^2}{N-1}} \tag{4}$$

where P_j is the power value at a specific frequency f in a recognized band F and $P(F)$ is the average power in that band.

The ABP and SD features are computed for each band in each window and scaled by dividing each feature value V_{ic} of a certain band F in a particular window over the average of that band-specific feature in all windows within the same channel C:

$$V_{scaled}(F) = \frac{V_{ic}(F)}{\frac{1}{L}\sum_{i=0}^{L-1} V_{ic}(F)} \tag{5}$$

C. DEEP LEARNING IMPLEMENTATION

EEG data of 32 subjects (16 healthy and 16 MDD subjects) were selected for training and validating the created models. The scaled values of ABP and SD obtained from each temporal window in each channel were generated. To allow the model to recognize different patterns among healthy and depressed subjects, we configured the input data of each window as a 2D array (channels*features) (19,10). Each array contains the 19 channels and the corresponding ABP values and their SD (10 features) of each of the 5 bands. Therefore, for the 32 subjects, we have 960 independent windows ($T = 30$ windows*32 subjects). Each window was labeled as either 0 (healthy) or 1 (depressed). This strategy was used to increase the number of samples. We used three different splitting schemes for training, validating, and testing the models:

- Experiment #1: we partitioned the windows into training and validation subsets using a conventional 80:20 split.
- Experiment #2: we employed a training and validation strategy involving 20 subjects, with the remaining 12 subjects (6 MDD and 6 HC) reserved exclusively for testing. This approach was designed to ensure that the test set comprised subjects completely excluded from any training or validation procedures.
- Experiment #3: subjects were divided into three distinct groups: a training set comprising 18 subjects, a validation set with 6 subjects, and a test set consisting of 8 entirely unseen subjects. The training process was iteratively conducted over 4 cycles. Each cycle involved randomly selecting a unique subset of 8 subjects from the dataset for testing to assess the model's performance. This systematic rotation ensured that every subject in the dataset had an equal opportunity to be included in the testing set across the iterations. As a result, testing accuracies were obtained for all 32 subjects, providing a comprehensive evaluation of the model's predictive capabilities.

These experimental designs were implemented to address the critical requirement for independent test sets composed of subjects entirely distinct from those used in training. By ensuring such independence, we aimed to enhance the reliability and validity of our model evaluations in distinguishing between MDD and HC.

We applied temporal sequence processing as a prerequisite to feeding the LSTM-based models. Each subject's window was wrapped by its preceding successive sequential 29 sliding windows. This is a common strategy to allow sequential learning, adapt to the spontaneous brain oscillatory patterns and temporal variations from depression to healthy status and vice versa, and prevent feature loss along windows [29]. A single window of one subject may not be sufficient to make a convincing prediction because depression biomarkers may be manifested in instantaneous phases of the recorded EEG.

D. CHANNELS SELECTION

The main practical problem that confronts us is the large number of input data (EEG channels and extracted features) that are fed into the model, which forces us to design a network with a very large TNP. In a real-time setting, streaming raw EEG signals from a high number of channels is impractical and cumbersome EEG headsets are inconvenient in wearables. In addition, it raises serious challenges about memory footprint and energy consumption, given the considerable computational load on the device processor. One approach to solve this problem is to statistically select the significant channels and eliminate the less important ones. This will allow us to reduce the complexity of the network. Consequently, we have investigated the input data (channels and features) using both correlation analysis and two-way Analysis of Variance (ANOVA) scoring.

The correlation analysis was carried out separately on the band features (ABP and SD) and the channels to remove the highly correlated ones. The correlation analysis had revealed that frequency band features are not highly correlated. However, channels are highly correlated, and redundant channels can be safely dropped to reduce network complexity.

ANOVA scoring on pairs of (channel, band feature), showed that some pairs were having high scores, and some have low scores. Fig. 2 shows the first 7 highest scores. The results revealed that (*Cz*, *theta*) was the one with the highest score. Also, *gamma* and *theta* bands were dominating in the first most important features. These results coincide with the literature results which reinforced the diagnostic potentials of these two bands as they are related to emotional processing (*theta*) and mood swings (*gamma*) [5].

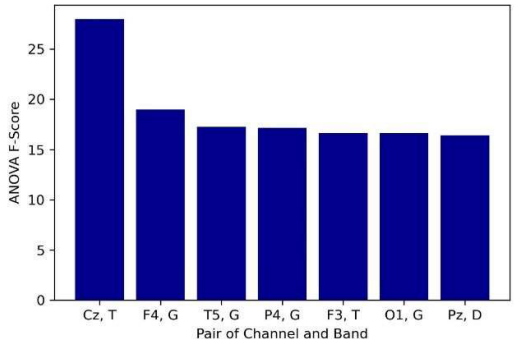

FIGURE 2. Channel-feature pairs with the highest ANOVA scores.

TABLE 1. Summary of channels selection.

Channel (ANOVA Score)	Correlated Channels (ANOVA Score)	Selected Channels
Fp1 (59.6)	*FP2* (59.04)	*Fp1*
F3 (60.5)	*F4* (56)	*F3*
F7 (66.86)	*F8* (55.2) and *Fz* (48)	*F7*
C3 (89.67)	*Cz* (83.2) and *C4* (62.67)	*C3* and *Cz*
P3 (47.5)	*Pz* (65.9) and *P4* (53.9)	*Pz*
O1 (40.63)	*O2* (66.3)	*O2*
T3 (47.2)	*T4* (49.3)	*T4*
T5 (60.15)	*T6* (66.09)	*T6*

The selection of channels was based on the correlation between channels and ANOVA scoring of the correlated channels. The channels with the highest total ANOVA score are selected. Consequently, 9 channels out of 19 channels were selected. The summary of channel selection is shown in TABLE 1, which demonstrates the correlated channels, the corresponding total ANOVA score in each channel, and the selected channels. Note here that *C3* and *Cz* were both selected as the *Cz* and *Theta* combination has the highest ANOVA score.

III. REFERENCE DEEP LEARNING MODELS

We have trained and tested various Deep Learning models. First, the reference models (CNN, LSTM, CNN-LSTM) were designed efficiently (low TNP). Second, the proposed novel Encoder-LSTM-ATTN model was created to overcome the shortcomings of reference models and the enormously complex state-of-the-art models.

Despite the competitive accuracy of the CNN-LSTM model, it requires extensive local memory and exhibits excessive power consumption. To improve the latter metrics, the model is redesigned to reduce the number of parameters compared to state-of-the-art models.

First, we tested the conventional CNN and LSTM models, each designed with one hidden layer for maximum efficiency. Then, the CNN-LSTM model is designed and trained. As shown in Fig. 3, a typical 1D-CNN is first constructed to receive the input X_t in a temporal sequence of 2D windows with their corresponding features (timestamp, channels, features). The entire sequence of CNN layers including the convolutional layer, Max Pooling layers, and flatten layers is wrapped in Time-Distributed layers. Then, these layers are stacked to the LSTM layer.

$$X_t = [x_0, x_1, x_2, \ldots, x_T]\, \varepsilon R^{T*m*n} \tag{6}$$

The pre-extracted features of ABP and SD are further processed by CNN to identify complex and high-level representations and patterns about the relative band power and variation in the brain's activity among different bands. Referring to Fig. 3, at a given timestep, each 1D-CNN in a Time-Distributed layer accepts the 2D spatial input features of each window to adaptively learn meaningful spatial characteristics and hierarchies. Window's features are inputted into a CNN and get convolved with the kernels. Then, Max Pooling is applied to reduce the dimensionality of produced

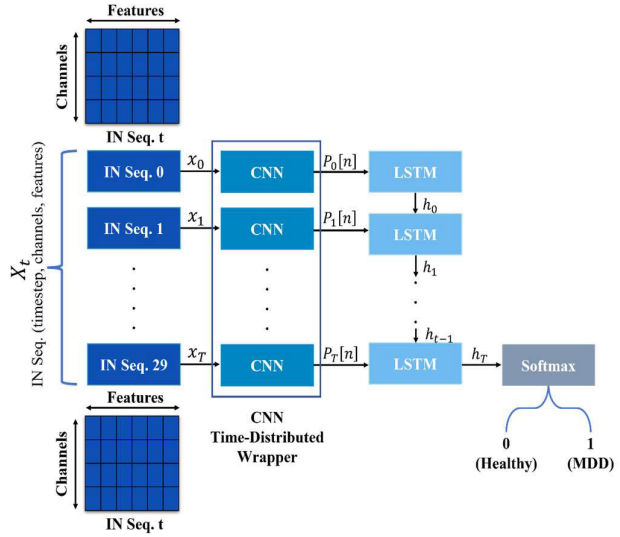

FIGURE 3. Typical CNN-LSTM.

features maps $P_t[n]$. The procedure continues in a time-sequential manner.

Next, an LSTM layer with a sequential length of 30 is placed to model the temporal dependencies of information extracted by CNN from each time window and its 29 predecessor windows. The LSTM takes the entrant features maps obtained from the pooling layer, $P_t[n]$, for each sequential window, adjusts the hidden state h_t one at a time, and passes the information to other LSTM cells in an ordered fashion to model the temporal variations among the sequential time windows. The LSTM uses a gating mechanism to regulate the flow of incoming sequential information and keeps relevant time-dependent information. The final hidden state carries information about the latest 30^{th} window to predict its state value based on its prior time windows.

Finally, a *Softmax* layer is used to perform the classification prediction based on the final hidden state. The *Softmax* function is applied to the final hidden state h_{29} of the last LSTM cell to obtain a vector of probabilities Pr distribution of possible classes:

$$ClassPr = Softmax(h_{29} \times w_{softmax} \times b_{softmax}) \quad (7)$$

It is then mapped to one output class (the one with the highest probability) using the argmax function (8).

$$PredictedClass = argmax(ClassPr) \quad (8)$$

For the traditional CNN, LSTM, and CNN-LSTM topologies, multiple architectures were investigated with different design parameters within each topology. The goal is to facilitate the analysis and trace the effect of reducing the cost of CNN and LSTM operations on the model's performance and improve TNP utilization (the best result for each model is shown in TABLE 2). It is worth noting that this reduction in complexity could result in poor generalizability on unseen data because these restricted-design models may not be able to capture and learn crucial spatial features. Our proposed model will address this issue efficiently, as discussed in the next section.

IV. PROPOSED ENCODER-LSTM-ATTN MODEL

Two important factors have influenced the architecture of the proposed model. The first is the sufficiency and effectiveness of the extracted features (ABP and SD). The features hold viable clues about the brain activity rhythms and their variance. The second is the role of LSTM with its internal structure capable of modelling the short-term and long-term behaviors of brain activity and representing time-dependent properties and temporal variations patterns. To develop our proposed model, we followed the following methodology to address the computational load problem. First, we paid particular attention to reducing the input size using an encoder bottleneck architecture to boost efficiency, accuracy, and generalizability.

The encoder has the potential to compress the input and alleviate the effects of noise and irrelevant bands/features [23], [31], resulting in more efficient training. Furthermore, the encode's compressed output allows us to design a tiny model without being susceptible to generalizability issues, as it will keep only relevant features in its bottleneck structure. Second, we fully harnessed the LSTM because of its decisive influence on accuracy enhancement. Finally, we utilized a transformer network with a self-sequence temporal attention mechanism network to align with the compressed input data and potential information loss during chronological long-sequence processing. The general topology of the proposed model is depicted in Fig. 4. The detailed architectural design of the model is demonstrated in Fig. 5. CNN was removed as it increases TNP without a significant accuracy improvement. Also, CNN may not be valuable for compressed input structure, owing to the spatial information loss caused by the encoder.

A. ENCODER BOTTLENECK ARCHITECTURE

The encoder bottleneck architecture used in the proposed model is shown in Fig. 5. It is adapted from an Autoencoder (AE) network. The Encoder layer receives the input data and adapts its weights to reconstruct and recover the data at the output of the decoder. The goal of an autoencoder is to regenerate the input data while providing a more suitable internal representation. The coder or bottleneck serves as an intermediate medium holding a compressed version of the data into a compact latent representation.

Unlike usual AE trained in an unsupervised manner, we trained our encoder with a *Softmax* layer included. We set cross-entropy as the loss function for depression classification instead of the decoder loss function. We found that it is more convenient to optimize the weights assigned to the input data according to the final target prediction. This representation carried by the encoder encompasses the most important characteristics needed to identify depression. Also, removing the decoder part boosted efficiency.

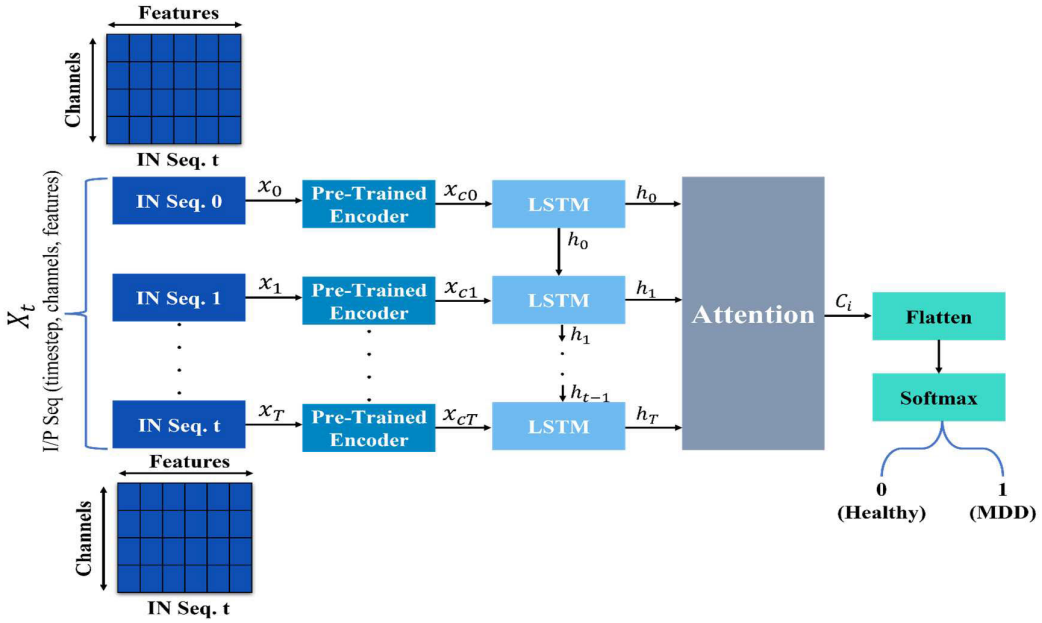

FIGURE 4. Proposed Model General Topology.

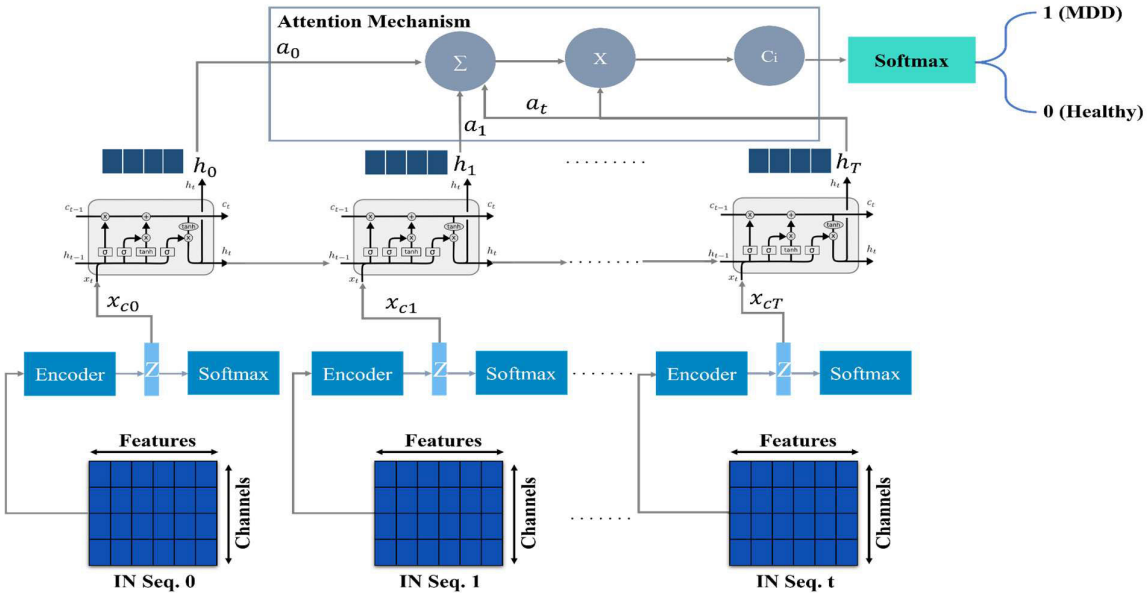

FIGURE 5. Proposed model detailed architecture.

In our model, the encoder takes each temporal window x_t feature vector and processes it to obtain the coded output.

$$x_{ci} = [x_{co}, x_{c1}, \ldots, x_{cT}] \tag{9}$$

which will be processed by the LSTM, on the one hand, and mapped component-wise using the nonlinear map:

$$Z\,(x_{ci}) = \sigma(w_{x_i} + b) \tag{10}$$

At the input of the Encoder, the feature vector is of size 90, which accounts for 9 channels and their corresponding 10 bands-related features. This is followed by a hidden layer containing 30 neurons and a bottleneck layer with 20 neurons. The Encoder results in a compact representation of the data using a vector of size 20, i.e., a compression ratio of 77.8%.

It is important to note that we are taking the most significant information learnt by the coder or bottleneck. However, the spatial information of the input window is lost.

B. LSTM

LSTM can model the transient non-periodic brain activities and states and grasp interactions and temporal dynamics among the sequential temporal windows. The generated

sequential temporal windows x_{ci} obtained from the coder Z are passed through the LSTM layer with a sequential length of 30-timesteps.

Each window (to perform a prediction on) is packed with its preceding 29 windows as we did in the CNN-LSTM model. The LSTM takes each window in the sequence, passes it through an LSTM cell, and outputs the hidden state h_t at a given timestep. The same LSTM operation we articulated in the CNN-LSTM model section is accomplished.

To persist in having an efficient model, we fixed the LSTM hidden units to only 4. Consequently, the temporal characteristics extracted and stored in the hidden units are very concise. Accordingly, it is expected to lose information along the sequential LSTM cells.

C. TEMPORAL ATTENTION MECHANISM

Each hidden state produced by an LSTM cell feeds the consecutive one with condensed information in a compressed hidden units of 4 hence, we might lose sequential contextual information. The main challenge here is the connection between hidden states feeding each other about previous timesteps sequences.

Through a long sequence of 30 windows (thus 30 hidden states with very small hidden units), a gradient vanishing problem is encountered because the data is compressed by the Encoder and then further compacted in a very small number of hidden units.

To rectify this problem, we applied temporal sequential self-attention mechanism to selectively focus on important input sequences. That is to guarantee that the model will not underfit due to insufficient, lost, and noisy information because of compression [32].

The attention mechanism layer is intended to assign temporal Attention weight to each hidden state h_j in the current time. At any given timestep, sequence self-attention takes all existing sequential hidden states (with their embedding information in the hidden units) and compare them with each other considering the context for each timestamp.

It searches for the input sequence of highest significance, assigns weight a_{ij} to the current and all previous hidden states according to their relevancy and impact on each other based on the context of the current position/time. Then, it adds these weights to the output sequence to generate a Context Vector

$$C_i = \sum_{j}^{T} a_{ij} h_j \tag{11}$$

as illustrated in Fig. 5. The output sequence positions are modified according to these new assigned weights and are stored in a context vector C_i of the same size obtained by the LSTM layer.

The utilization of this network permits the model to pay more attention to the most relevant tokens of information in the sequence and thus create a more meaningful representation. This helps in obtaining the influential temporal features to improve the prediction accuracy of the model without increasing TNP massively. The new sequence produced by the attention layer is then passed into a flattened layer and finally through a *Softmax* layer to perform the classification, as shown in Fig. 5.

V. RESULTS

A. EXPERIMENT 1: SUBJECT-DEPENDENT (80:20 WINDOWS SPLIT)

In this section, we will illustrate some of experimental results generated from the tests we implemented on the state-of-the-art efficient CNN, LSTM, and CNN-LSTM models and the proposed Encoder-LSTM-ATTN model. TABLE 2 summarizes these results into three parts.

As shown in the first part of TABLE 2, we carried out planned comparisons between the reference models trained on the 9 selected channels and the corresponding 10 features. The approach was to decrease the number of parameters, mainly the number of CNN filters and LSTM units to reduce complexity. The best performance results by the models are presented in TABLE 2. First, we tested the CNN model with 64 filters (TNP = 3,394), it achieved 95.89% validation accuracy, which is less than the accuracy of current state-of-the-art (99.37%) [11]. This result is generally accepted based on accuracy-cost trade-off concept. We reduced the filter size significantly compared to other studies [11], [13], [14] which used a high number of filters e.g., 128 and 256. The standalone LSTM with 32 hidden units led to better validation accuracy (96.92%) but higher TNP (15,810). Then, we reduced the network size further to TNP =5,554 using the CNN-LSTM model, which yielded the same validation accuracy of CNN (95.89%).

Looking at the second part of TABLE 2 which outlines the models' performance when trained on the compressed input by the encoder bottleneck. At first glance, it is explicitly observed that the same reference models (CNN and CNN-LSTM) delivered better results with the encoder. A general trend was identified in the reference models (combined with encoder) tests outcomes, that is a better generalizability in the learning curves and reduced complexity. This reinforces our primary premise that encoder boosts efficiency and performance by filtering the input from noise and irrelevant data and providing a compressed refined representation.

On the other hand, LSTM with the encoder (95.00%) showed a noticeable performance contrast compared to LSTM alone (96.92%). Presumably, this is a result of compressed information in 4 hidden units, and it can be rectified by the attention module. Evidently, Encoder-CNN-LSTM showed superior performance (98.88%) with the attention module. Nevertheless, it required 7,731 TNP.

The last row of TABLE 2 presents the results of the proposed Encoder-LSTM-ATTN model. As clearly shown, the Encoder-LSTM model's accuracy was 95.00% without the Attention network, but in the proposed model, it reached the highest validation accuracy of 99.57% with the Attention module. The accuracy curves are shown in Fig. 6 and Fig. 7.

The proposed model (99.57%) outperformed the Encoder-CNN-LSTM-ATTN model (98.88%) and reduced TNP to half (4,355) of that in the reference Encoder-CNN-LSTM-ATTN model (7,731).

B. EXPERIMENT 2: SUBJECT-INDEPENDENT (TESTING ON 12 UNSEEN SUBJECTS)

Initially, our proposed model achieved a high validation accuracy of approximately 99.57%, which raised concerns about potential overfitting and data leakage. To address this, we conducted a rigorous subject-independent experiment where the test set comprised entirely new subjects not included in the training/validation phase. This approach aimed to prevent biases that might arise if the model learned to classify subjects based solely on subject identities rather than generalizable features distinguishing MDD from HC. Consequently, our refined approach involved training with 20 subjects and testing exclusively on 12 unseen subjects (6 MDD and 6 HC).

Results from this experiment showed a testing accuracy of 89.72%, revealing a significant drop from the initial validation accuracy of the proposed model. This decline suggests potential information leakage in the earlier validation phase and emphasizes the necessity for robust validation protocols to ensure model generalization. This revised methodology provides a more reliable assessment of the model's performance, affirming its effectiveness in predicting MDD across new subjects while maintaining superior accuracy and lower complexity compared to baseline models.

C. EXPERIMENT 3: SUBJECT-INDEPENDENT (TESTING ON ALL 32 UNSEEN SUBJECTS)

In Experiment 3, we adopted an iterative approach to thoroughly validate the model across all 32 subjects. Each iteration involved testing on a distinct subset of 8 subjects (4 HC and 4 MDD) to ensure comprehensive evaluation.

Looking at TABLE 2 Average Testing Accuracy on 32 Subjects), a consistent trend is noted, where integrating the encoder module enhances accuracy across all models compared to their baseline configurations. For instance, the CNN model with an encoder (Encoder-CNN) improved from 65.41% to 75.98%, demonstrating a substantial enhancement in predictive capability. Similarly, the LSTM model combined with an encoder (Encoder-LSTM) showed an increase in accuracy from 72.85% to 76.92%.

Fig. 8 shows the results of all models with sensitivity and specificity alongside average accuracy as evaluation metrics. The proposed Encoder-LSTM-ATTN architecture emerged as the top performer, achieving an average testing accuracy of 84.93%, sensitivity (Sen) of 88.89%, and specificity (Spec) of 80.98%, surpassing all other models while maintaining lower complexity.

Fig. 9 presents the classification results obtained from the proposed model's predictions on each tested subject, categorized by their respective diagnoses: MDD and HC. Analysis of individual subject accuracies revealed variability among both MDD and HC groups, ranging from 54.44% to 100% for MDD and 61.11% to 96.67%% for HC. The variability among subjects' accuracies may stem from factors such as individual variations in EEG patterns and overlapping features between HC and MDD.

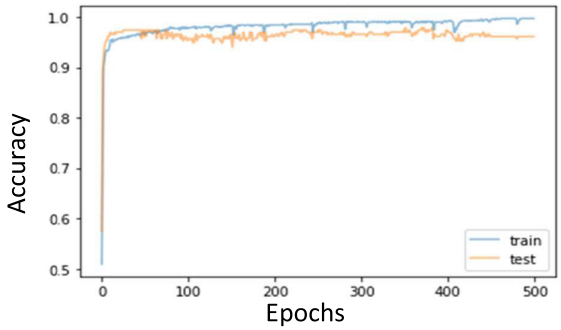

FIGURE 6. The accuracy curve for the encoder-LSTM model.

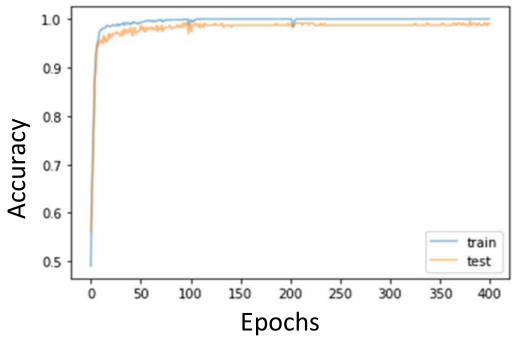

FIGURE 7. The accuracy curve for the proposed encoder-LSTM-ATTN model.

In summary, while the initial high validation accuracy may have been misleading due to data leakage, the subsequent experiment's results provide a more reliable assessment of the model's performance, showcasing its effectiveness in generalizing to new subjects for MDD detection. Despite this reduction in accuracy, it is important to note that the proposed model maintained its superiority, exhibiting the best accuracy while simultaneously showing low complexity compared to baseline models.

VI. DISCUSSION

The outcomes of the proposed model were reaped from the combination of Encoder, LSTM, and Attention modules. The Encoder assisted in minimizing the input size along with channels selection and provided a better representation of the input data and reduced the overall TNP or complexity.

The LSTM provided a great powerfulness in the extraction of temporal properties of brain's activity variations among the sequential windows. The Attention mechanism extends the model capabilities of achieving a high performance by flexibly focusing on important information in the long sequence and clarifying unclear data resulted due to compression by the Encoder and the LSTM small hidden units.

TABLE 2. Experimental results of state-of-the-art efficient CNN, LSTM, and CNN-LSTM and the proposed encoder-LSTM-ATTN Model.

Model (Input Size)	Filter	Kernel Size	Pool Size	Strides	LSTM Units	Validation Accuracy (80:20 Subject-Dependent)	Testing Accuracy on 12 Subjects	Average Testing Accuracy on 32 Subjects	TNP
CNN (9,10)	64	5	5	2	N/A	95.89%	62.44%	65.41%	3,394
LSTM (30,90)	N/A	N/A	N/A	N/A	32	96.92%	73.93%	72.85%	15,810
CNN-LSTM (30,9,10)	16	2	2	1	16	95.89%	78.18%	76.80%	5,554
Encoder-CNN (30,20)	32	2	-	2	N/A	97.77%	73.93%	75.98%	5,666
Encoder-LSTM (30,20)	N/A	N/A	N/A	N/A	4	95.00%	85.45%	76.92%	3,802
Encoder-CNN-LSTM (30,20)	16	2	-	2	16	97.77%	82.12%	81.07%	6,194
Encoder-CNN-LSTM-ATTN (30,20)	16	2	-	2	16	98.88%	83.03%	81.50%	7,731
Proposed: Encoder-LSTM-ATTN (30,20)	N/A	N/A	N/A	N/A	4	99.57%	89.72%	84.93%	4,355

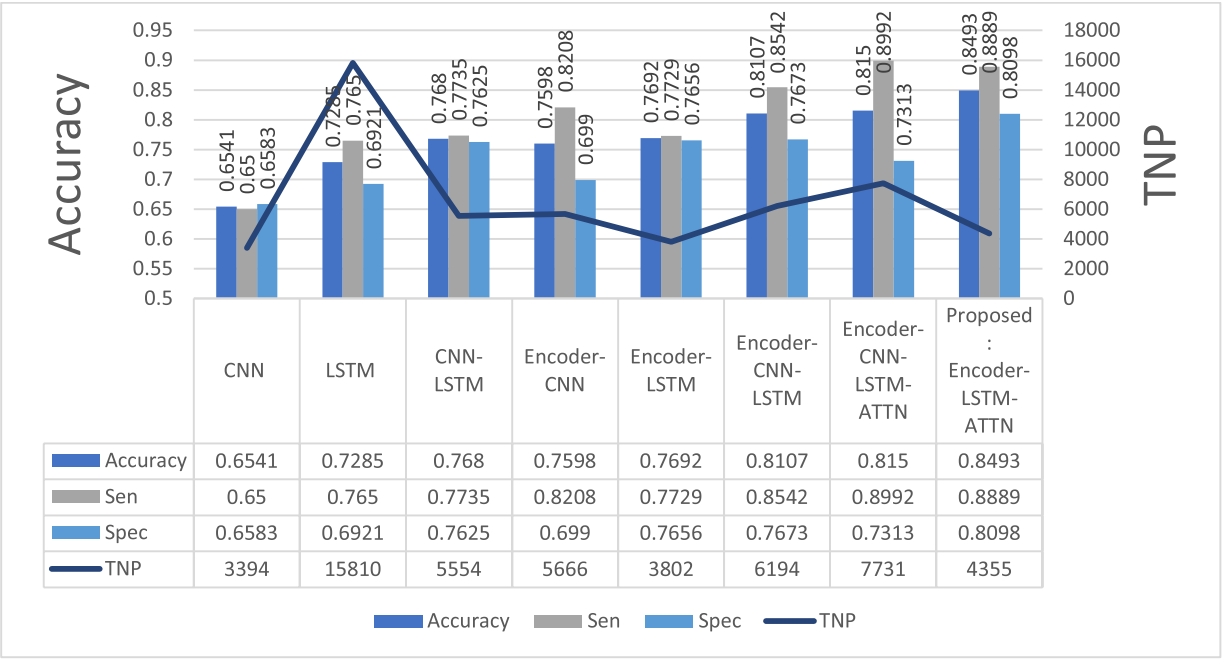

FIGURE 8. Experiment 3 (testing on 32 subjects): models results.

The results lend strong support to the argument that it is possible to create a high-performing efficient model considering the success pillars: a solid EEG signal processing, proper channels selection, extraction of impactful features, and employment of on-purpose networks with a flexible design.

One concern about the findings is the presence of noise and redundant information in the input, as revealed by the testing outcomes without the encoder. This implies that some EEG bands are not necessarily important. Thus, the results drawn must be replicated on only impactful EEG bands for this classification problem. Researchers are encouraged to find affirmative and consistent evidence about EEG biomarkers for depression. This may improve aspects of the model' generalizability. Besides, it will enhance efficiency by reducing the overall model complexity and real-time processing energy and memory.

TABLE 3 introduces a comparison between the proposed Encoder-LSTM-ATTN model and recent state-of-the-art results. Apparently, our proposed model showed superior performance, compared to models developed in [11], [12], [13], and [14], with a validation accuracy of 99.57%. In addition, it required the lowest TNP.

Our model was efficient to a greater extent by 99.65% decrease in TNP compared to the significantly complex model introduced by [13] which required 1,276,898. They used raw EEG signals of only (FP1-T3 channel and FP2-T4 channel) as input to their model.

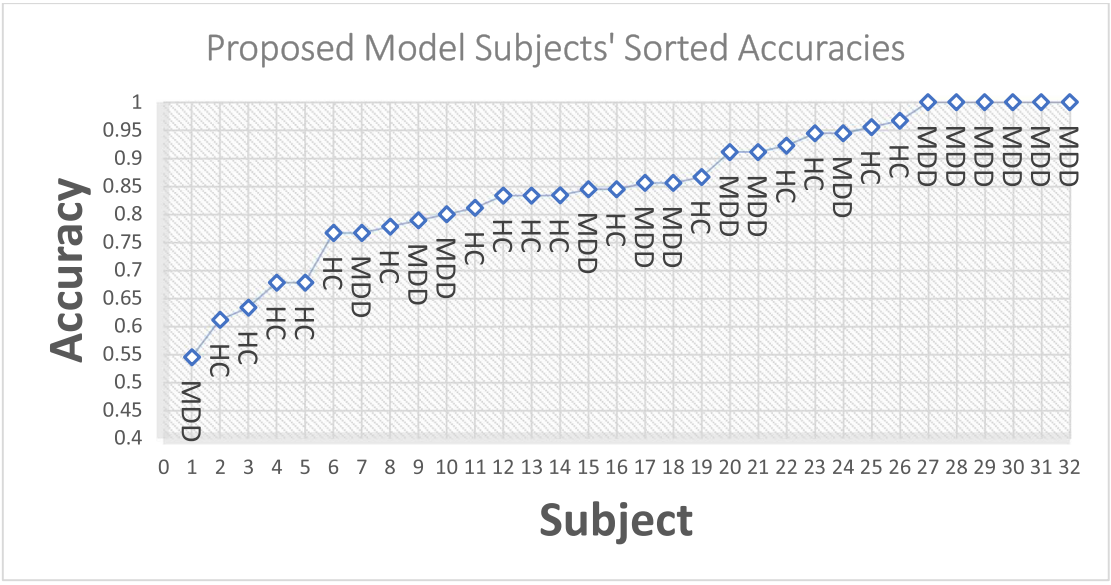

FIGURE 9. Proposed model classification accuracy per subject.

In [14], a powerful model is proposed with 2 CNNs, 1 GRU, and an Attention module. They used 16 channels (selected out of 128) and extracted PSD of the 5 bands. The TNP of this model was not mentioned. However, the model was constructed with 128 and 256 filters and 256 hidden units for GRU. Therefore, it is sufficient to say that their model requires much higher TNP in contrary to our proposed model.

Other studies developed CNN models without LSTM. For example, the model proposed by [11] includes 5 CNNs and 2 FC layers which take windowed EEG signals. This model required 363,882 TNP. It must be pointed out that complex designs were needful in cases where raw EEG in inputted to the models.

Other machine learning models demonstrated lower classification accuracy. For example, the work in [33] proposed support vector machine (SVM), and it showed 81.8% accuracy using EEG features and 92.7% using combined fNIRS and EEG features. Another SVM model was developed by [34], using only 2-channel frontal EEGs. The model yielded 90% accuracy.

From the short review above, key findings emerge: the usage of raw EEG signals and large input force a complex model design with greater number of filters, hidden units, and neurons. Also, despite the utilization of reduced number of channels in [13] and [14], the models are still complex due to the size of input and the selection of large design parameters. Less complex models such as SVM [33], [34] showed decreased performance compared to CNN, LSTM, and CNN-LSTM models. Based on the comparisons presented in TABLE 3, it turns out that our proposed model is optimal as it achieved a trade-off between efficiency and performance.

Addressing the identified leakage issue, we conducted another comparison with state-of-the-art models to ensure fairness, see TABLE 4. Our study employed an Encoder-LSTM-ATTN architecture and tested it on all 32 subjects in an iterative manner, achieving an average testing accuracy of 84.93%. This approach allowed us to evaluate the model's performance on completely unseen data, mitigating overfitting or data leakage concerns.

In comparison, previous studies by [35] and [36] utilized different methodologies. For example, [35] employed a multi-head self-attention mechanism and parallel two-branch CNN with leave-one-subject-out cross-validation, achieving a testing accuracy of 91.06%. Whereas [36] utilized the Inception Time model with 10-fold cross-validation at the subject level, achieving a testing accuracy of 91.67% (19 channels) and 87.50% (10 channels). Our model's testing accuracy demonstrates competitive performance compared to these state-of-the-art approaches, with the added advantage of reduced model complexity, indicated by fewer parameters (4,355).

The proposed model (saved in.h5 format) requires only 126 KB. The EEG processing Python script has a memory footprint of 14 KB and takes 8.20 seconds to execute on a 1.60 GHz CPU. The algorithm applies the pre and post processing operations, discussed earlier, on each subjec's EEG recording of 300 seconds (30 windows). Our cost-effective model and algorithm have fewer hardware necessities and high compatibility with the microcontroller, hence, can adaptively and easily fit on it compared to other computationally expensive models.

It might be argued that the processing operations done to extract the spectral features also contribute to the overall complexity of the system. However, in the case of multi-tasking

TABLE 3. Comparison of the proposed encoder-LSTM-ATTN with existing state-of-the-art models.

Study	Dataset	Hidden Layers	Hidden Units/Neurons	Filters	Kernel	Strides, Pool Size	Input	Accuracy	TNP
[11]	33 Sub.	5 CNN, 2 FC	FC (16, 8)	128, 64,64,32,32	5,5,5,3,2	Stride =1 Pool = 2	4-sec windows in 19 channels	99.37%	363,882
[12]	45 Sub.	1 CNN, 2 LSTM, 2 FC	LSTM (64, 32) and FC (16, 16)	64	5	Stride=1	64 channels, window size=20	99.10%	46,658
[13]	30 Sub.	4 CNN, 1 LSTM, 1 FC	LSTM (32) and FC (64)	64,128, 128, 32	5,3,13,7	Stride=1 Pool=2	Raw EEG 2000 samples in FP1-T3 cand FP2-T4	99.12%	1,276,898
[14]	53 Sub.	2 CNN,1 GRU, Attention	GRU (256)	128, 256	5	Pool = 2	PSD of 5 bands in 16 channels	99.33%	N/A
This Study	32 Sub.	Encoder, 1 LSTM, Attention	LSTM (4)	N/A	N/A	N/A	Compressed Selected 9-channels ABP and SD features of the 5 bands. Vector Size=20	99.57%	4,355

TABLE 4. Comparison with state-of-the-art (mitigating leakage in model evaluation).

Study	Model	Testing Strategy	Testing Accuracy	TNP
[35]	Multi-head self-attention mechanism and parallel two-branch CNN	Leave-one-subject-out cross-validation.	91.06%	3,297,010
[36]	InceptionTime model (6 Inception modules)	10-fold cross-validation at the subject level.	91.67% (19 channels). 87.50% (10 channels).	N/A
Our Study	Encoder-LSTM-ATTN	Subject-Independent (Iterative Testing on all 32 Unseen Subjects).	84.93%	4,355

mental health monitoring using EEG wearables, the PSD of the five bands are common features and spectral analysis is inevitable for the detection of other disorders such as anxiety and stress. On the other hand, relying on pre-processed signals as input to each disorder-specific model will result in explosion of models. Thus, designing an efficient model from scratch based on important PSD features will reduce the overall complexity and provide better interpretability.

One limitation of this study is its small sample size, which restricts the generalizability of findings. Scaling up to a larger number of subjects would inevitably increase the complexity of the model to accommodate more diverse patterns.

Future studies aiming for low-complexity models should investigate the stability of the proposed model when trained on larger datasets. The current model, with only 4 LSTM units, may struggle to capture complex patterns adequately. While suitable for small datasets, it could potentially underfit more complex datasets. Larger datasets may necessitate increasing the number of LSTM units to effectively capture the variability and complex patterns inherent in the data.

Also, the feasibility of designing a single model for the detection of multiple disorders can be studied in the future to achieve better overall efficiency.

VII. CONCLUSION

The findings of this study provide conclusive support for the effectiveness of the proposed model design and the usefulness of the combined encoder, LSTM, and attention mechanism modules. These modules served as mitigating factors to the computational encumbrance and complexity of currently available models for depression detection. Our study underlines the importance of considering the expected obstacles of deploying computationally expensive EEG-based depression detection models in tiny devices. Also, we provided actionable insights on how to reduce mode's complexity and thereby speed up inference without a sacrifice in accuracy, to be applicable in embedded wearable tiny devices for future attempts.

ACKNOWLEDGMENT

Noor Faris Ali was with the Electrical and Communication Engineering Department, College of Engineering, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates.

REFERENCES

[1] S. Yasin, S. A. Hussain, S. Aslan, I. Raza, M. Muzammel, and A. Othmani, "EEG based major depressive disorder and bipolar disorder detection using neural networks: A review," *Comput. Methods Programs Biomed.*, vol. 202, Apr. 2021, Art. no. 106007.

[2] World Health Org. (Sep. 13, 2021). *Depression*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[3] A. Safayari and H. Bolhasani, "Depression diagnosis by deep learning using EEG signals: A systematic review," *Med. Novel Technol. Devices*, vol. 12, Dec. 2021, Art. no. 100102.

[4] J. S. Kumar and P. Bhuvaneswari, "Analysis of electroencephalography (EEG) signals and its categorization—A study," *Proc. Eng.*, vol. 38, pp. 2525–2536, Jan. 2012.

[5] F. S. de Aguiar Neto and J. L. G. Rosa, "Depression biomarkers using non-invasive EEG: A review," *Neurosci. Biobehavioral Rev.*, vol. 105, pp. 83–93, Oct. 2019.

[6] J. Shen, X. Zhang, X. Huang, M. Wu, J. Gao, D. Lu, Z. Ding, and B. Hu, "An optimal channel selection for EEG-based depression detection via kernel-target alignment," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2545–2556, Jul. 2021.
[7] J. J. Newson and T. C. Thiagarajan, "EEG frequency bands in psychiatric disorders: A review of resting state studies," *Frontiers Human Neurosci.*, vol. 12, p. 521, Jan. 2019.
[8] B. Hosseinifard, M. H. Moradi, and R. Rostami, "Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal," *Comput. Methods Programs Biomed.*, vol. 109, no. 3, pp. 339–345, Mar. 2013.
[9] V. A. Grin-Yatsenko, I. Baas, V. A. Ponomarev, and J. D. Kropotov, "Independent component approach to the analysis of EEG recordings at early stages of depressive disorders," *Clin. Neurophysiol.*, vol. 121, no. 3, pp. 281–289, Mar. 2010.
[10] A. Tendler and S. Wagner, "Different types of theta rhythmicity are induced by social and fearful stimuli in a network associated with social memory," *eLife*, vol. 4, Feb. 2015, Art. no. e03614.
[11] A. Seal, R. Bajpai, J. Agnihotri, A. Yazidi, E. Herrera-Viedma, and O. Krejcar, "DeprNet: A deep convolution neural network framework for detecting depression using EEG," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[12] G. Sharma, A. Parashar, and A. M. Joshi, "DepHNN: A novel hybrid neural network for electroencephalogram (EEG)-based screening of depression," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102393.
[13] B. Ay, O. Yildirim, M. Talo, U. B. Baloglu, G. Aydin, S. D. Puthankattil, and U. R. Acharya, "Automated depression detection using deep representation and sequence learning with EEG signals," *J. Med. Syst.*, vol. 43, no. 7, pp. 205:1–205:12, May 2019.
[14] Z. Wang, Z. Ma, W. Liu, Z. An, and F. Huang, "A depression diagnosis method based on the hybrid neural network and attention mechanism," *Brain Sci.*, vol. 12, no. 7, p. 834, Jun. 2022.
[15] B. A. Hickey, T. Chalmers, P. Newton, C.-T. Lin, D. Sibbritt, C. S. McLachlan, R. Clifton-Bligh, J. Morley, and S. Lal, "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review," *Sensors*, vol. 21, no. 10, p. 3461, May 2021.
[16] A. Abd-Alrazaq, R. AlSaad, F. Shuweihdi, A. Ahmed, S. Aziz, and J. Sheikh, "Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression," *NPJ Digit. Med.*, vol. 6, no. 1, pp. 1–16, May 2023.
[17] A. Ahmed, S. Aziz, M. Alzubaidi, J. Schneider, S. Irshaidat, H. A. Serhan, A. A. Abd-Alrazaq, B. Solaiman, and M. Househ, "Wearable devices for anxiety & depression: A scoping review," *Comput. Methods Programs Biomed. Update*, vol. 3, Jan. 2023, Art. no. 100095.
[18] S. Alhassan, A. Soudani, and M. Almusallam, "Energy-efficient EEG-based scheme for autism spectrum disorder detection using wearable sensors," *Sensors*, vol. 23, no. 4, p. 2228, Feb. 2023.
[19] Y. Zhang, Y. Savaria, S. Zhao, G. Mordido, M. Sawan, and F. Leduc-Primeau, "Tiny CNN for seizure prediction in wearable biomedical devices," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 1306–1309.
[20] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–37, Dec. 2023.
[21] S. Jang, W. Liu, and Y. Cho, "Convolutional neural network model compression method for software–hardware co-design," *Information*, vol. 13, no. 10, p. 451, Sep. 2022.
[22] docs.arduino.cc. *Nano 33 BLE Sense | Arduino Documentation*. Accessed: Mar. 24, 2024. [Online]. Available: https://docs.arduino.cc/hardware/nano-33-ble-sense
[23] F. Laakom, J. Raitoharju, A. Iosifidis, and M. Gabbouj, "Reducing redundancy in the bottleneck representation of autoencoders," *Pattern Recognit. Lett.*, vol. 178, pp. 202–208, Feb. 2024.
[24] D. Walther, J. Viehweg, J. Haueisen, and P. Mäder, "A systematic comparison of deep learning methods for EEG time series analysis," *Frontiers Neuroinform.*, vol. 17, Feb. 2023, doi: 10.3389/fninf.2023.1067095.
[25] W. Mumtaz, L. Xia, M. A. M. Yasin, S. S. A. Ali, and A. S. Malik, "A wavelet-based technique to predict treatment outcome for major depressive disorder," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171409.
[26] L. Sun, Y. Liu, and P. J. Beadle, "Independent component analysis of EEG signals," in *Proc. IEEE Int. Workshop VLSI Design Video Technol.*, May 2005, pp. 219–222.
[27] H. Chang, Y. Zong, W. Zheng, C. Tang, J. Zhu, and X. Li, "Depression assessment method: An EEG emotion recognition framework based on spatiotemporal neural network," *Frontiers Psychiatry*, vol. 12, Mar. 2022, Art. no. 837149.
[28] S. D. Kumar and D. P. Subha, "Prediction of depression from EEG signal using long short term memory (LSTM)," in *Proc. 3rd Int. Conf. Trends Electron. Inform. (ICOEI)*, Apr. 2019, pp. 1248–1253.
[29] S. F. Naqvi, S. S. A. Ali, N. Yahya, M. A. Yasin, Y. Hafeez, A. R. Subhani, S. H. Adil, U. M. Al Saggaf, and M. Moinuddin, "Real-time stress assessment using sliding window based convolutional neural network," *Sensors*, vol. 20, no. 16, p. 4400, Aug. 2020.
[30] K. E. Atkinson, "An introduction to numerical analysis," *Math. Comput.*, vol. 54, no. 190, p. 903, Apr. 1990.
[31] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, "Robust information bottleneck for task-oriented communication with digital modulation," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2577–2591, Aug. 2023.
[32] G. Ding, A. Plummer, and I. Georgilas, "Deep learning with an attention mechanism for continuous biomechanical motion estimation across varied activities," *Frontiers Bioeng. Biotechnol.*, vol. 10, Oct. 2022, doi: 10.3389/fbioe.2022.1021505.
[33] L. Yi, G. Xie, Z. Li, X. Li, Y. Zhang, K. Wu, G. Shao, B. Lv, H. Jing, C. Zhang, W. Liang, J. Sun, Z. Hao, and J. Liang, "Automatic depression diagnosis through hybrid EEG and near-infrared spectroscopy features using support vector machine," *Frontiers Neurosci.*, vol. 17, Aug. 2023, doi: 10.3389/fnins.2023.1205931.
[34] A. Aderinwale, G. B. Tolossa, A. Y. Kim, E. H. Jang, Y.-I. Lee, H. J. Jeon, H. Kim, H. Y. Yu, and J. Jeong, "Two-channel EEG based diagnosis of panic disorder and major depressive disorder using machine learning and non-linear dynamical methods," *Psychiatry Res., Neuroimaging*, vol. 332, Jul. 2023, Art. no. 111641.
[35] M. Xia, Y. Zhang, Y. Wu, and X. Wang, "An end-to-end deep learning model for EEG-based major depressive disorder classification," *IEEE Access*, vol. 11, pp. 41337–41347, 2023.
[36] A. Rafiei, R. Zahedifar, C. Sitaula, and F. Marzbanrad, "Automated detection of major depressive disorder with EEG signals: A time series classification using deep learning," *IEEE Access*, vol. 10, pp. 73804–73817, 2022.

NOOR FARIS ALI received the B.Sc. degree (Hons.) in biomedical engineering from Ajman University, Ajman, United Arab Emirates, in 2021. She is currently pursuing the M.Sc. degree in biomedical engineering with the American University of Sharjah. She was a Researcher with the Electrical Engineering Department, United Arab Emirates University. She is working as a Graduate Teaching and Research Assistant. Her research interests include biomedical signal processing and analysis, embedded systems, edge machine learning, and deep learning for biomedical applications.

NABIL ALBASTAKI (Member, IEEE) received the B.Sc. degree in computer engineering from The University of Arizona, in May 1989, the M.Eng. degree in electrical engineering from Cornell University, in May 1993, and the Ph.D. degree in computer engineering from the University of Southern California, in May 2001. He joined United Arab Emirates University (UAEU), as a Teaching Assistant, in 1990. He is currently an Assistant Professor with the Department of Electrical Engineering, UAEU. He is also the Assistant Dean of research and graduate studies with the College of Engineering (COE). He has held several positions while at UAEU, the Director of the Continuing Education Center, the Assistant Dean of the Student Affairs, College of Engineering, and the Head of the Industrial Training and Graduation Projects Unit. His research interests include embedded systems, robotics, and digital systems design.

ABDELKADER NASREDDINE BELKACEM (Senior Member, IEEE) received the Ph.D. degree in information processing from Tokyo Institute of Technology, Japan, in 2015. He is currently an Assistant Professor with United Arab Emirates University (UAEU). Prior to joining UAEU, he was a Specially Appointed Researcher with the Department of Neurosurgery, Osaka University Medical School, and an Assistant Professor with the Endowed Research Department of Clinical Neuroengineering, Global Center for Medical Engineering and Informatics, Osaka University, Japan, from 2015 to 2019. His research interests include brain–computer/machine interface using human magneto- and electro-encephalography (MEG/EEG), human–machine interaction, artificial intelligence, robotics, and neuroscience. He was one of the recipient of the Prestigious International Award "MIT Technology Review Innovators Under 35 MENA" and brings over ten years of experience in brain–machine interface (BMI). One of his work was adopted on a cover page of IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, Volume 26-6, 2018. He has been the Vice Chair of the IEEE P2725.1 Working Group, since 2021. He has contributed in many international technical program committees, such as the IEEE International Conference on Systems, Man, and Cybernetics (SMC2018/20/21), as the Co-Chair of BMI workshops. He is the Guest Associate Editor of BMI on *Frontiers in Human Neuroscience*.

IBRAHIM (ABE) M. ELFADEL (Life Senior Member, IEEE) received the Ph.D. degree from Massachusetts Institute of Technology (MIT), in 1993. He is currently a Professor of computer and communication engineering with Khalifa University, Abu Dhabi, United Arab Emirates. Before his current academic position, he was with the corporate CAD organizations at IBM Research and the IBM Systems and Technology Group, Yorktown Heights, NY, USA, where he was involved in the research, development, and deployment of CAD tools and methodologies for IBM's high-end microprocessors. From 2012 to 2019, he led three Abu Dhabi-based, industrially funded research centers dedicated to the IoT, 3D Integration, and MEMS. He was a recipient of six invention achievement awards, one Outstanding Technical Achievement Award, and one Research Division Award, all from IBM. His other awards include the D. O. Pederson Best Paper Award from IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, the SRC Board of Directors Special Award for pioneering semiconductor research in Abu Dhabi, the Best Paper Award from the IEEE Conference on Cognitive Computing, Milan, Italy, in July 2019, and the 2022 Service Award from the International Federation of Information Processing (IFIP). He served on the Technical Program Committees for several leading conferences, including DAC, ICCAD, ASPDAC, DATE, ISCAS, AICAS, BioCAS, VLSI-SoC, ICCD, ICECS, and MWSCAS. He was the General Co-Chair of VLSI-SoC 2017, Abu Dhabi, United Arab Emirates, and the Technical Program Co-Chair of VLSI-SoC 2023, Sharjah, United Arab Emirates, and AICAS 2023, Hangzhou, China. He is the Technical Program Chair of CloudCom2024, Abu Dhabi, and the Technical Program Co-Chair of BioCAS 2024, Xi'an, China. He is an Associate Editor of IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR ARTIFICIAL INTELLIGENCE.

MOHAMED ATEF (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering, electronics and communications from Assiut University, Egypt, in 2000 and 2005, respectively, and the Ph.D. degree from the Institute of Electrodynamics, Microwave and Circuit Engineering, Vienna University of Technology, in 2010. From 2006 to 2007, he got a research scholarship from the Department of Microelectronics, Czech Technical University in Prague, he involved in the improvement of quantum dot optical properties. He was a Postdoctoral Researcher until the end of 2012. He visited the School of Microelectronics, Shanghai Jiao Tong University, China, from 2015 to 2017. He has been an Associate Professor with Assiut University, since 2016. In 2020, he joined the Electrical Engineering Department, United Arab Emirates University, United Arab Emirates, where he is currently an Associate Professor. He is the author of two books, such as "*Optical Communication over Plastic Optical Fibers: Integrated Optical Receiver Technology*" and "*Optoelectronic Circuits in Nanometer CMOS Technology*," and the author and co-author of more than 90 scientific publications. His research interests include optoelectronic integrated circuits, optical sensing, and biomedical circuits and systems. He served as a TPC member for many IEEE conference. He has been a member of the Biomedical and Life Science Circuits and Systems Technical Committee, since 2018. He was awarded the State Encouragement Award in Advanced Technological Sciences Serving the Engineering Sciences for 2018 from Egyptian Academy of Scientific Research and Technology (ASRT). He served as the Lead Editor for *Sensors* Special Issue (2019–2020), the Guest Editor for several IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS (IEEE TBioCAS) special issues, an Associate Editor for IEEE TBioCAS (2020–2023), and an Associate Editor for IEEE SENSORS JOURNAL since 2024.