Depression in Obstructive Sleep Apnea Patients: Is Using Complex Deep Learning Structures Worth It?

Mostafa Moussa, Yahya Alzaabi and Ahsan Khandoker

Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi, 127788, U.A.E.

Keywords: Depression, Electroencephalography, Electrocardiography, Breathing Signals, Gated Recurrent Unit Long Short-Term Memory Networks.

Abstract: The prevalence and severity of depression make it imperative to develop a means to automatically detect it, so as to alleviate the associated mental effort and cost of seeing a dedicated professional. Depression can also co-exist with other conditions, such as Obstructive Sleep Apnea Syndrome (OSAS). In this paper, we build upon our previous work involving sleep staging, detection of OSAS, and detection of depression in OSAS patients, but focus solely on the latter of the three. We use features extracted from EEG, ECG, and breathing signals of 80 subjects suffering from OSAS and half of which also with depression, using 75 % of this 80-subject dataset for training and 10-fold cross-validation and the remainder for testing. We train three models to classify depression: a random forest (RF), a three-layer artificial neural network (3-ANN), and a gated-recurrent unit long short-term memory (GRU-LSTM) recurrent neural network. Our analysis shows that, like our previous work, the 3-ANN is still the best performing model, with the GRU-LSTM following closely behind at an accuracy of 79.0 % and 78.6 %, respectively, but with a smaller F1-score at 80.0 % and 81.6 %. However, we believe that the large increase in computation time and number of learnable parameters does not justify the use of GRU-LSTM over a simple ANN.

1 INTRODUCTION

Major Depressive Disorder (MDD) is a common mental disorder characterized by reduced production of certain neurotransmitters in the brain that affects 10 % of the population (Gao et al., 2018). Patterns described by Murray et al. in (Murray et al., 2012) show that depression is consistently on the rise as a prevalent cause of morbidity or disability and its effects include but are not limited to, memory loss, irritability, loss of interest, disordered sleep (insomnia or hypomnia) and eating (weight loss or gain), tiredness and lethargy, anxiety, reduced cognitive and/or motor performance, feelings of inadequacy, inability to concentrate, self-harm or suicidal ideation or attempt, and unexplained physical pain (Strock, 2002).

Obstructive Sleep Apnea Syndrome (OSAS) is a condition characterized by cessation of breathing during sleep specifically due to airway blockages primarily caused by muscles, mainly the genioglossus. Though OSAS is not as prevalent as depression and has vastly differing causes, it can still occur in patients with depression, or vice versa. It is thus not unlikely that a selected dataset of OSAS patients would include those with depression as well, as is the case in our previous works and this current one (Moussa et al., 2022). Though depression was an important part of these previous works, OSAS was the main focus and depression was classified as a comorbidity.

From the literature, we know sleep apnea and hypopnea are correlated with lower quality of life in general including in large part psychological health. That is to say depression is relatively prevalent in people who suffer from OSAS (Yue et al., 2003; Björnsson et al., 2016; Ejaz et al., 2011). In one of the aforementioned works, Yue et al. found that the 30 patients suffering from sleep apnea and hypopnea have higher scores for depression with a t-value of 2.62 (P < 0.05) (Yue et al., 2003).

In the literature, we have seen plenty of works wherein the authors use electrophysiological signals, such as ECG (Zang et al., 2022) or EEG (Mumtaz et al., 2018; Hosseinifard et al., 2013) in classifying major depressive disorder or depression. The use of varying machine learning algorithms played a critical role in classification in these works, which...
Figure 1 gives an abstract idea of our methodology, as well as the architectures/algorithms we used in our work. The contribution in our work lies mainly in classification of depression in OSAS with the novel dataset via machine learning and a simple deep learning architecture, and gauging what would make switching to deep learning worth the increase in computational cost and subsequently, physical cost.

2 METHODOLOGY

2.1 Dataset and Processing

Seeing the extensive use of electrophysiological signals for classification of depression, we elected to use electroencephalography (EEG), electrocardiography (ECG), and breathing signals for that purpose. We focus particularly on depression in subjects that are suffering from OSAS, so while our results may not necessarily be applicable to the general population, they can provide a suitable baseline for OSAS patients. For the purpose of detecting depression alone in OSAS patients, we use a subset of the dataset described in our previous work (Moussa et al., 2022); instead of using the electrophysiological signals of 118 subjects, we use that of 80 subjects. These 80 subjects consist of 40 with depression and OSAS and 40 with OSAS alone, collected from the American Center of Psychiatry and Neurology (ACPN) in Abu Dhabi, UAE, meaning we omit the 6 healthy subjects from this set and the 32 supplementary healthy subjects from the STAGES dataset (Zhang et al., 2018). The 80 subjects selected only from the original study, excluding the STAGES healthy subjects, consist of 48 male subjects and 32 female, all UAE Nationals between the ages of 20 and 66 with a mean age of 44.2 ± 10.9
years-old at the time of the study. This study was approved by the Institutional Review Board (IRB) of the ACPN on the 2nd of October, 2017 with IRB reference number 0019.

Among the 80 subjects, 2 had an Apneahypopnea Index (AHI) less than 5, 27 had an AHI between 5 and 15, 27 had an AHI between 15 and 30, and 24 had an AHI above 30. Since we know the status of both depression and apnea, we can train supervised machine learning models to classify our subjects into one of two classes: depressed or not depressed, both with OSAS. We can also better partition them according to AHI, sleep stage, and depression status to further investigate the effects of certain conditions on classification performance in other works, as we did for sleep stages in (Moussa et al., 2022).

As previously stated, we primarily use EEG, ECG, and breathing signals, namely airflow, oxygen saturation, and thoracic effort, in addition to other information “signals”, such as the hypnogram detailing sleep stages. These are not the only recorded signals, however. The subjects undergo overnight polysomnography, which conventionally include the aforementioned signals in addition to chin and leg electromyography (EMG), electrooculography (EOG) for both eyes, and abdominal effort. Chin EMG (Al-Angari, 2008; Moradhasel et al., 2021) could pave the way for better detection of OSAS due to the more direct causal effect between the condition and dilator muscles, and could even facilitate the use of sensors directly with the genioglossus muscle instead of chin placement.

The main five signals are recorded by means of an 8-channel EEG cap for brain signals, an ECG for heart signals, a spiro meter for airflow, a pulse oximeter for oxygen saturation, and a piezoelectric belt for thoracic movement. The EEG channels used are O1, O2, C3, C4, F3, and F4 with A1 and A2 according to the 10-20 convention, as shown in Figure 2, and the other signals are recorded via standard leads/sensors and standard lead/sensor placement.

After obtaining the signals, some processing would be required to ensure the data is clean and ready for feature extraction, selection, and eventually, classification. Since the EEG, ECG, and breathing signals are sampled at 200 Hz, 100 Hz, and 10 Hz, respectively. The EEG and ECG are also put through a 50 Hz Notch filter to remove the power-line interference and all three signals are put through bandpass filters in previous work to be published by Yahya Alzaabi; the breathing signals and ECG at 0.1-0.4 Hz and the EEG at 0.5-30 Hz to keep beta, theta, alpha, and delta waves. Following filtering, the signals are split into 3-minute intervals selected manually by inspection mainly based on whether or not an apnea has occurred, so as to avoid artifacts. This results in a total of 1,424 intervals or observations from the 118 subjects, of which 1,005 observations are from our 80 subjects. For each of these observations, we compute a set of 34 features, 24 from EEG signals, 6 from the ECG signals/heart rate variability (HRV), 1 directly from airflow, and 3 from the interaction between airflow and ECG/HRV, or more specifically R-R interval (RRI) signals. The EEG features are simply average powers extracted for each brain wave from each electrode, with the exception of the reference electrodes, the ECG features include the average very low frequency, low frequency, and high frequency powers, a normalized set of the latter two, and the ratio/division between the latter two. The singular breathing signal/airflow feature is the respiratory frequency, and the remaining three features are the respiratory sinus arrhythmia (RSA), the normalized RSA, and the time-dependent phase coherence between RSA and airflow (phases extracted via Hilbert transform), also known as lambda (λ).

After taking care of noise with filtering and manual selection of intervals and extracting our feature set, we fill in missing values using shape-preserving piecewise cubic spline interpolation (Fritsch and Carlson, 1980; Kahaner et al., 1989), also known as Pchip, then follow that by Softmax normalization, Box-Cox transform (Box and Cox, 1964) to ensure normal probability distribution, and z-score normalization (Moussa et al., 2022). These processing steps are described in Equations 1-3, where Data1 is the Softmax normalized data, Data2 is Data1 with probability distribution made approximately normal, and DataFinal is centered and standardized Data2. Box-Cox transform is a non-linear power transform that makes the data probability distribution approximately normal by finding an optimal value of an exponent (λ).
that results in the best normal distribution approximation. Looking at Equation 2, we can conclude that Box-Cox transformation would require the input data to be positive, which we achieve via Softmax normalization before applying the power transform.

\[
\text{Data1} = \frac{1}{1 + \exp\left(\frac{\text{mean(Data)} - \text{Data}}{\text{std(Data)}}\right)} \quad (1)
\]

\[
\text{Data2}(\lambda) = \begin{cases} \frac{\text{Data}^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(\text{Data})/\lambda, & \text{if } \lambda = 0 \end{cases} \quad (2)
\]

\[
\text{DataFinal} = \frac{\text{Data2} - \text{mean(Data2)}}{\text{std(Data2)}} \quad (3)
\]

Despite extracting 34 features for each observation, we do not use the full feature set in this work. As we have seen in (Moussa et al., 2022), using $\chi^2$ to select features whose importance score is greater than or equal to the average feature importance score, along with the bi-layer artificial neural network (ANN) yielded the best classification result for depression compared to other feature selection algorithms including sequential feature selection, neighborhood and principal component analysis, maximum relevance minimum redundancy and RelieFF algorithms, so we opt to directly apply $\chi^2$ in feature selection, ending up with six features out of the thirty four. The six selected features by this technique are all extracted from EEG signals, also surprisingly from only two channels. These features include the average powers of beta, theta, and alpha waves from channels F3 and F4. In the context of feature selection on MATLAB, the function examines whether each of our 34 features is independent of the depression status using individual $\chi^2$ tests. The score output from this function is the negative of the common logarithm of the p-value, and we know a small p-value indicates that the corresponding feature is dependent on the label is an important feature. This score would approach infinity as the p-value approaches zero. Our analysis concluded that the aforementioned six features have an infinite score, hence were selected as our features.

Now that signal processing has concluded, we have a clean dataset of 1,005 observations each with six features with an approximately normal probability distribution and no missing values. The 80 subjects are then split into two sets, one for training and 10-fold cross-validation and comprises the observations of 75% of the subjects, and the other set for testing and comprises the observations of the remaining 25% of the subjects. The labels are likewise partitioned in the same manner, culminating in a partitioned dataset ready to be input to machine learning algorithms.

### 2.2 Classifiers and Performance Evaluation

As we saw in Section 1, machine learning is commonly used in detecting depression in the literature, due to its automated nature, the simplicity of its metrics, and the insights it could help us derive regarding the nature of the condition, the widely established methods of diagnosing depression, or the features used in classification. In addition, it has social benefits as it reduces the need for human interaction in diagnosis.

Artificial neural networks (ANNs) and deep learning techniques use the back-propagation algorithm to minimize a loss function, and to automatically extract features with the major difference being an added function or layer. In convolutional neural networks, the added function would be convolutional layers, which, as their name suggests, convolve the input to reduce its size, producing a smaller feature map. In gated recurrent unit long short-term memory (GRU-LSTM or GRU) networks, the added function(s) are an update and reset gates that control the flow of information (Erdenebayar et al., 2019).

As we have previously tested out numerous classifiers in (Moussa et al., 2022), we opt to directly compare the best-performing model in that work (ANN), with a deep learning technique- a GRU-LSTM network, and getting the results with random forest as some form of baseline. This is because random forests are known for their generally robust performance and relative simplicity compared to deep learning techniques. The random forest (RF) used was the same as the previous work; bagged trees with surrogate decision split and 200 learning cycles. However, some changes were made to the ANN model to better optimize it for the problem. The model, named 3-ANN, now consists of three hidden layers instead of two with 100 units each and a regularization term (lambda) of 0.01 instead of 0 in between the input and output layers. The GRU-LSTM model is new, as it would have been difficult to employ prior to hardware upgrade from a machine with the Nvidia GTX 1050Ti GPU to one with RTX 3080, and consists of a total of 15 layers, as shown in Figures 1 and 3. These layers begin with a feature input layer of size $6 \times$ Number of training samples, followed by a gated recurrent unit of size 10 and a 40% dropout layer. Afterwards, we have three fully connected layers with 50 units with batch normalization and ReLu following each one. Then finally, we have our “output” layer, which consists of a fully connected layer with 2 units, a SoftMax layer, and the actual output layer, since our entire analysis and net-
Figure 3: Layer descriptions and number of learnable parameters of the GRU-LSTM model. Each layer has the number of units under its name and/or any additional options (i.e. normalization, dropout, number of channels).

work design are done on MATLAB. The weights are initialized via the Glorot initializer (Glorot and Bengio, 2010). The 3-ANN uses the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm in training while the GRU uses stochastic gradient descent with a momentum of 0.8. The additional training options for the GRU include a mini-batch size of 32, a fixed learn rate of 0.01, an L2-regularization term of 0.005, a validation frequency of 1 (each epoch), and a maximum number of epochs of 1000. Figure 3 shows a brief description of each layer and gives an idea about the number of associated computations.

We use 10-fold cross-validation, as stated earlier, to ensure our models do not overfit. With the deep learning model on MATLAB, this is implemented by training a model using 9 folds and the last fold for validation, and repeating while changing the validation fold until all folds have been used for validation, and then the model with the best validation performance is taken. With the other two models, the implementation on MATLAB is more automatic than having to select a model based on validation performance algorithmically.

After the models are trained and validated, we measure their performance with the testing set. Classification performance is measured by accuracy, sensitivity, specificity, precision, F1-score, the area under receiver operating characteristics (ROC) curve (AUC), Cohen’s χ coefficient (Cohen, 1960), and Matthews correlation coefficient (Matthews, 1975). Accuracy measures how many instances were correctly classified, sensitivity measures the number of instances correctly classified positive out of the actual positive instances, specificity measures the number of instances correctly classified negative out of the ac-
Figure 4: Testing confusion matrix, receiver operating characteristics (ROC) curve, and posterior probability plot of the 3-ANN model.

Figure 5: Testing confusion matrix, receiver operating characteristics (ROC) curve, and posterior probability plot of the GRU model.

Table 1: Testing performance of the three classifiers in classification of depression in OSAS patients.

<table>
<thead>
<tr>
<th>Model</th>
<th>RF</th>
<th>3-ANN</th>
<th>GRU-LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.71</td>
<td>0.84</td>
<td>0.76</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>67.6</td>
<td>79.0</td>
<td>78.6</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>56.8</td>
<td>79.7</td>
<td>89.9</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>79.7</td>
<td>78.2</td>
<td>66.2</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>75.7</td>
<td>80.3</td>
<td>74.7</td>
</tr>
<tr>
<td>F1-Score (%)</td>
<td>64.9</td>
<td>80.0</td>
<td>81.6</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.36</td>
<td>0.58</td>
<td>0.57</td>
</tr>
<tr>
<td>MCC</td>
<td>0.37</td>
<td>0.58</td>
<td>0.58</td>
</tr>
</tbody>
</table>

3 RESULTS AND DISCUSSION

As stated earlier, we compute the testing accuracy, sensitivity, specificity, precision, F1-score, AUC, Cohen’s $\kappa$ coefficient, and Matthews correlation coefficient (MCC) for the three classifiers. Although we compute all metrics, we mainly look at the accuracy, F1-score, $\kappa$, and MCC in comparison in order to make a conclusion regarding the best classifier for detecting depression in OSAS patients with our dataset and processing steps.

Table 1 shows comparable performance between the 3-ANN and the GRU-LSTM and shows both beating the random forest model in all metrics but specificity. The 3-ANN has a higher AUC, accuracy, specificity, precision, and $\kappa$ than the GRU, but the GRU has a higher sensitivity, F1-score and they both have almost the same value of the Matthews correlation coefficient.

The reasons the performance of the two neural network models is similar could include the relatively small size of the available dataset, the use of only 6 out of the 34 features, the simplicity of the selected features, or the simplicity of the supposedly more complex model (GRU). The first reason is simple enough; artificial neural networks generally re-

<table>
<thead>
<tr>
<th>Work</th>
<th>Main Objective</th>
<th>Dataset</th>
<th>Machine Learning Algorithms</th>
<th>Significance</th>
<th>Limitations</th>
<th>Best Model</th>
<th>3-ANN</th>
<th>Logistic Regression</th>
<th>Support Vector Machines</th>
<th>Convolutional Neural Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Zang et al., 2022)</td>
<td>Classify Depression</td>
<td>74 subjects'raw ECGs</td>
<td>CNN, LR, SVM, and NB</td>
<td>Simplicity of methodology: The authors use raw ECG signals with CNNs in their analysis</td>
<td>Using CNNs with raw signals is inconvenient in resource-restricted environments</td>
<td>Accuracy (%)</td>
<td>93.96</td>
<td>SVM 98.00</td>
<td>LR 90.00</td>
<td>ANN 97.80</td>
</tr>
<tr>
<td>(Mumtaz et al., 2018)</td>
<td>Classify Depression</td>
<td>64 subjects'EEGs</td>
<td>KNN, LDA, and LR</td>
<td>Thorough analysis for some classic machine learning algorithms and features used are promising</td>
<td>No significant limitations found, though we would be interested to see how this setup performs with other datasets</td>
<td>Specificity (%)</td>
<td>99.90</td>
<td>N/A</td>
<td>N/A</td>
<td>79.70</td>
</tr>
<tr>
<td>(Hosseinifard et al., 2013)</td>
<td>Classify MDD</td>
<td>90 subjects'EEGs + 4 non-linear features</td>
<td>The authors present a thorough description of a robust methodology to classify depression in general, describing in detail their features, machine learning models and cross-validation schemes, as well as their novel dataset</td>
<td>Only the accuracy is reported</td>
<td>Deep training not thoroughly explored, and no automatic hyperparameter optimization via grid-search or Bayesian optimization</td>
<td>F1-Score (%)</td>
<td>97.00</td>
<td>95.00</td>
<td>N/A</td>
<td>80.00</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>Classify Depression in OSAS patients</td>
<td>1,005 observations extracted from EEG, ECG, and breathing signals of 80 subjects</td>
<td>Random Forest, 3-ANN, GRU-LSTM</td>
<td>Compares best depression in OSAS classification method in (Moussa et al., 2022) with deep learning</td>
<td>No significant limitations found, though we would be interested to see how this setup performs with other datasets</td>
<td>Accuracy (%)</td>
<td>93.67</td>
<td>97.00</td>
<td>80.00</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Due to the similarity of the MCC in particular, and the closeness of the accuracy, F1-score, and κ values, we cannot accurately say that one model outclasses the other for classification of depression in OSAS patients with this dataset. Instead, we can compare the resources required for each model and select the optimal one based on the less computationally expensive and less time-consuming one.

Despite similar performance, we see from Figure 3 that the number of learnable parameters is 6,500 for the GRU, comparatively smaller than that of 3-ANN at 21,102 (Weights + Biases: [(6×100) + (100×100) + (100×1) + (100×2)] + [(100×1) + (100×1) + (100×1) + (2×1)]). Despite that, it takes only 77.8 seconds to compute with a NVIDIA 1050 Ti GPU and significantly less with the NVIDIA 3080 GPU, whereas the GRU takes upwards of an hour to train with the latter. This could be attributed to the small size of mini-batches coupled with the large iteration/epoch limit, and the GRU layer itself. This makes the 3-ANN more suited for this problem, as it takes less time to train and is less demanding in terms of resources. Table 2 compares our work with similar works in the literature.

4 CONCLUSION

To sum up, the main goal of this work was to classify depression in OSA patients and investigate whether using deep learning over classic machine learning techniques is a worthy endeavor. The dataset included overnight EEG, ECG, and breathing signal recordings from 80 subjects, 40 of which were depressed with OSAS and 40 were not depressed but had OSAS. Afterwards, we extract 1,005 intervals from the sig-
nals depending on the status of obstructive apnea occurrence, in addition to depression status and sleep stage. We then process the data to ensure it is clean, has an approximately normal distribution, and is z-score normalized before we partition and input it into our three classifiers. We train three classifiers using the intervals or observations of 75 % of the subjects and perform 10-fold cross-validation on the same set, then test classifier performance with the data of the remaining 25 % of subjects. Using the Chi^2 algorithm to select the six most important features and ANN for classification yielded the best performance with an accuracy of 79.00 %, F1-score of 80.00 %, a κ of 0.58, a Matthews correlation coefficient of 0.58 and an AUC of 0.84, while also considering the low computational cost compared to the GRU-LSTM. The performance is promising, and we believe further preprocessing of the data, as well as further optimizing network architectures and hyperparameters and using more novel approaches like transformers could improve classification performance. In addition, implementing explainability metrics, like SHAP and descriptions would certainly make our work more accessible to clinical personnel, or even laypersons.

ACKNOWLEDGEMENTS
The authors would like to thank the American Center for Psychiatry and Neurology (ACPN) in Abu Dhabi for their invaluable contribution in sharing the polysomnography data and acknowledge the support of the biomedical engineering department and the Healthcare Engineering Innovation Center (HEIC) at Khalifa University of Science and Technology. The authors would also like to highlight the importance of the KAU-KU Joint Research Program, in particular, project DENTAPNEA between Khalifa University and King Abdulaziz University, in particular the advice of Dr. Angari, Dr. Balamesh, Dr. Khraibi, and Dr. Marghalani.

REFERENCES
Murray, C. J., Tos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K.,


