IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 September 2022, accepted 14 October 2022, date of publication 19 October 2022, date of current version 25 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3215632

RESEARCH ARTICLE

Explainable Computer-Aided Detection of Obstructive Sleep Apnea and Depression

MOSTAFA M. MOUSSA, YAHYA ALZAABI, AND AHSAN H. KHANDOKER, (Senior Member, IEEE)
Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
Corresponding author: Mostafa M. Moussa (mostafa.moussa@ku.ac.ae)

This work was supported by the King Abdulaziz University-Khalifa University Joint Research Program (KAU-KU JRP), KU Project, under Grant 8474000376.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the ACPN on 10/02/2017 with IRB under Application No. 0019.

ABSTRACT Obstructive Sleep Apnea Syndrome (OSAS) and Major Depressive Disorder (MDD) are common conditions associated with poor quality of life. In this work, we aim to classify OSAS and depression in patients with OSAS using machine learning techniques. We have extracted features from electrocardiograms (ECG), electroencephalograms (EEG), and breathing signals from polysomnography (PSG) at specific 5-minute intervals, where the participants' statuses are known, meaning we do not need breathing signals. These statuses include sleep stage, whether or not they have depression, or an apneic event has occurred. The PSGs were recorded from a total of 118 subjects with a 75/25 split for training and testing and the resultant features were used in sleep staging and classifying OSAS and depression in OSAS patients. Sleep staging was best done with random forest without feature selection, yielding an accuracy of 70.52 % and F1-Score of 69.99 %. The best classification performance of OSAS happened during deep sleep without feature selection and SVM, which yielded an accuracy of 98.36 % and F1-Score of 98.82 %. All sleep stages with Chi^2 ANN yielded an accuracy of 72.95 % and F1-Score of 73.43 % for classification of depression in OSAS patients. Results show promise in detecting OSAS and depression in OSAS patients, and the Bland-Altman plot shows that posterior probability provides comparable means of detecting OSAS to the apnea-hypopnea index (AHI). Besides detection of OSAS in depressed patients, this work serves to classify depression and give insights into relevant sleep stages to both of those conditions, allowing better planning for polysomnography.

INDEX TERMS Electroencephalography (EEG), electrocardiography (ECG), obstructive sleep apnea syndrome (OSAS), depression, sleep staging, machine learning.

I. INTRODUCTION

Obstructive Sleep Apnea Syndrome (OSAS) or Obstructive Sleep Apnea (OSA) is a breathing disorder characterized by brief obstruction of upper airways during sleep, disrupting it. OSA differs to Central Sleep Apnea (CSA) in that the cessation of breathing is caused by blockage of airways due to physical causes such as obesity, not due to a dysfunction in the brain or spinal cord. Both, however, are associated with metabolic diseases, mood disorders, reduced cognitive performance, increased risk of accidents, depression, memory loss, cardiovascular complications, such as arrhythmias, coronary heart disease, heart failure, and strokes among other effects [1], [2], [3]. OSAS is a relatively common condition of sleep-disordered breathing, affecting 3-7 % of men and 2-5% of women in the general population [4], increasing to about 24 % and 9 % respectively when polysomnographic criteria are solely considered [5], [6].

Depression or Major Depressive Disorder (MDD) is a common mental disorder characterized by reduced production of

The associate editor coordinating the review of this manuscript and approving it for publication was Humaira Nisar.

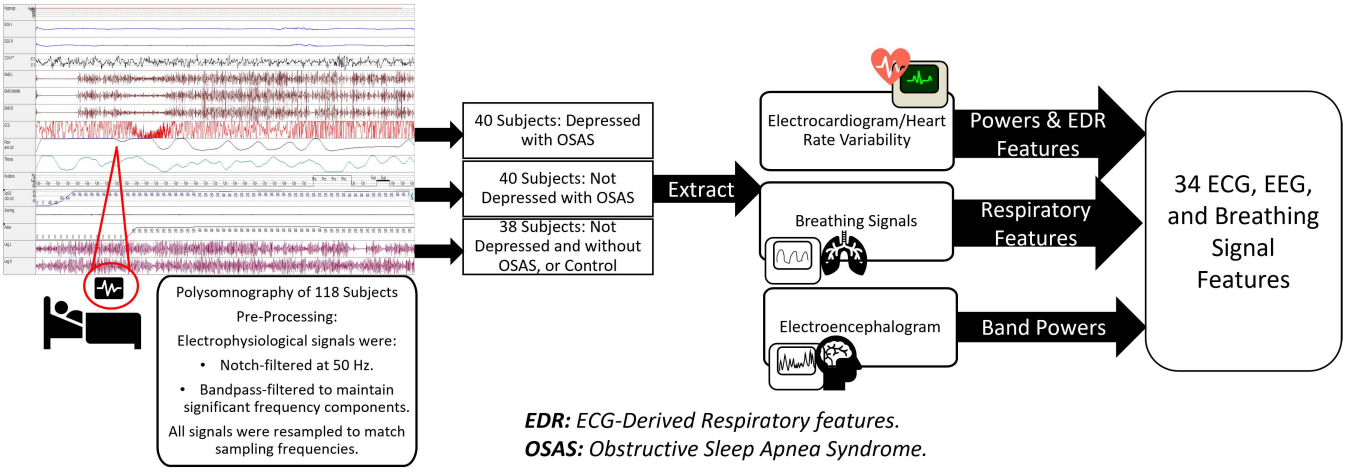

FIGURE 1. Data acquisition and feature extraction. Powers refer to various frequency bands in both ECG and EEG, but also with various channels for EEG.

certain neurotransmitters in the brain that affects 10 % of the population [7]. Patterns described by Murray et al. in [8] 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 hypersomnia) 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 [9]. Furthermore, factors such as traumatic events, chronic conditions, and abuse of alcohol and drugs contribute to the risk of depression [9].

The use of electrophysiological signals is extremely common in the medical field. ECG signals, are the first signals measured to rule out causes of symptoms that could pertain to heart disease, EEG signals in physical conditions that have to do with the brain or psychological ones, breathing signals for respiratory conditions, etc. For instance, electrophysiological signals can be used to detect and manage strokes, a condition comorbid with obstructive sleep apnea, as explained in Hussain et al.'s works [10], [11].

Studies by Yue et al., Björnsdóttir et al., and Ejaz et al. [12], [13], [14] show that sleep apnea-hypopnea is correlated with decline in general psychological well-being and that depression is indeed prevalent in untreated OSAS patients. Yue et al. found that the 30 patients suffering from sleep apnea-hypopnea syndrome (SAHS) on average have a higher general severity index (GSI) of the Symptom Checklist-90 (SCL-90) as well as higher scores for depression with a t-value of 2.62 ($P < 0.05$), among other symptoms like somatization, obsession-compulsion, anxiety, and hostility than the 30 control subjects [12]. The prevalence of depression in OSAS patients warrants an investigation into means to detect depression in OSAS patients, or in the case of untreated or undiagnosed OSAS, potentially use depression detection to diagnose OSAS. In this work, we are concerned with OSAS and depression in patients with OSAS and have identified the need for robust systems that could detect these two conditions. To that end, we propose the use of machine learning algorithms trained with a suitable dataset to detect OSAS and depression in OSAS patients, as shown in Figure 1 and Figure 2. We consider a multitude of machine learning techniques to detect OSAS and depression and give other insights into the better sleep stages to record to detect OSAS and depression, what features are more important for extraction, and what metrics to use or not use when diagnosing either one. Using these features instead of the raw signal may compromise performance to a certain degree, albeit likely to a small degree at least in the case of OSAS, but results in significantly faster and less hardware-taxing training. In other words, it saves clinicians time and money.

Machine learning methods are not unheard of in diagnosis or detection of both OSAS and MDD; plenty of works use machine learning different datasets involving detection of OSAS or depression. Khandoker et al. [3] use SVMs in OSAS detection with heart rate variability (HRV) and ECG-derived respiration (EDR) features decomposed into 14 levels via 10th order Daubechies wavelets extracted from 8-hour ECG recordings of 125 subjects with 83 subjects used for training with leave-one-out (LOO) cross-validation and 42 for testing. Their features were 14 HRV and 14 EDR variances of coefficients representing variability powers. SVMs with polynomial (d = 2, 3, and 4), Radial Basis Function (RBF) ($\sigma = 0.1, 0.5$, and 1), and linear kernels are used with different regularization parameters, C (0.1, 0.8, and 10) and different features (the 6 HRV, 6 EDR, and 4 HRV+EDR). For estimating the relative severity of OSAS using the classifier, the posterior probabilities of SVM outputs were calculated and then compared with the AHI. Maximum validation accuracy reached 100 % (so did specificity, sensitivity reach 100 %, and Cohen's κ and AUC reach 1.0) with 3rd-degree Polynomial kernel SVM and a C of 1 (0.8), which in turn yielded a testing classification accuracy of 92.85 % and a Cohen's κ value of 0.85. The posterior probability results suggest

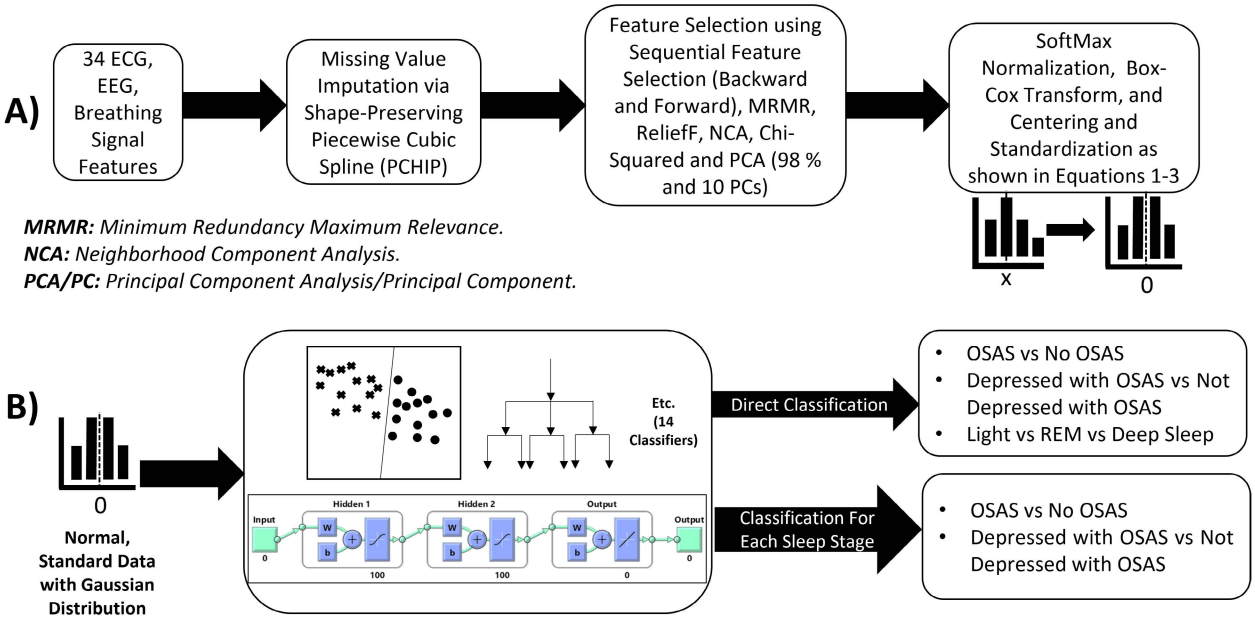

FIGURE 2. A) Data processing and feature selection steps. B) Classification steps.

superior performance of SVMs in OSAS recognition supported by wavelet-based features of ECG. The results demonstrate considerable potential in applying SVMs in an ECG-based screening device that can aid sleep specialists in the initial assessment of patients with suspected OSAS [3].

Bozkurt et al., Sheta et al., Dey et al., Erdenebayar et al., Padovano et al., Nazli and Altural, and Srinivasulu et al. [15], [16], [17], [18], [19], [20], [21] also use ECG signals to detect apnea events. These works use DTs [15], [16], [20], [21], random forest (RF) [20], LDA [16], logistic regression [16], Naive Bayes (NB) [16], SVM and KNN [15], [16], [19], [20], [21], ensemble classifiers [15], [16], [21], artificial neural networks [18], [20], convolutional neural networks (CNNs) [16], [17], [18], recurrent neural networks (RNNs) [16], [18], long short-term memory recurrent networks [16], [18] combinations of CNN and LSTM or CNN-LSTM [16], and gated recurrent unit (GRU) RNNs for classification [18] to varying degrees of success of detecting sleep apnea events.

Bozkurt et al. applied their methods on IIR-filtered (0.25-100 Hz, notch at 50 Hz) QRS intervals and HRV derived from ECG of 10 subjects and obtained a maximum testing accuracy of 85.12 %, sensitivity of 85 % and specificity of 86 % with an ensemble of DT, KNN, and SVM and 50 % of the features selected via Fisher algorithm [15]. Sheta et al. extract 9 general features from 70 ECG recordings in the CinC database and rank them using the relief algorithm [22], [23], IIR-filter out the power-line interference (60 Hz), and pass to their classifier, getting a maximum testing accuracy with 5-fold cross-validated CNN-LSTM at 86.25 %, F1-score at 87.68 % and AUC of 0.95 [16]. Dey et al. obtain a testing accuracy of 98.91 %, sensitivity of 97.82 %, and specificity of 99.20 % with their convolutional neural network applied to 7-10-hour, single-channel ECG recordings from 35 subjects from the Apnea-ECG dataset [17], [22], [24]. Erdenebayar et al. use ECG recordings of 86 subjects, 69 for training and 17 for testing as input to their convolutional and recurrent neural networks and obtain the best performance with GRU-RNN with an accuracy and sensitivity of 99.00 % [18]. Just like Dey et al. [17], Padovano et al. used the Apnea-ECG dataset, but with SVM and KNN instead of CNNs and more extensive pre-processing. Namely, selecting the middle five hours of recording, baseline drift removal, IIR filtering, and R-peak detection using Pan-Tompkins algorithm, followed by feature extraction from HRV in addition to common complexity measures and Renyi and Tsallis entropies, followed by sequential backward and forward feature selection (SBFS and SFFS) to reduce features. This resulted in a maximum accuracy of 81.4 % with an AUC of 0.83 for SVM with SBFS used for feature selection [19]. Nazli and Altural extract 25 HRV features and obtain a maximum testing accuracy of 85.26 % with random forest, which yielded a sensitivity of 75.44 % and a specificity of 91.40 %, though the dataset used and validation methods are unclear [20]. Finally, Srinivasulu et al. preprocess their ECG signals by first filtering out powerline interference with a notch filter, wavelet denoising using sym8 wavelet with heuristic threshold selection, and segmenting them into 1-minute intervals. In the end, they extract 5 features including wavelet entropy of the ECG signal from the 16 polysomnographs available in the MIT-BIH database [22], [25] with a 90/10 training/testing split. This results in a maximum accuracy of 89.60 %, sensitivity of 95.40%, and specificity of 66.10 % with bagged trees [21].

Uçar et al. use IIR-filtered (type II Chebyshev 0.1-20 Hz) abdominal PPG signals from 5 subjects and KNN,

RBF-neural network, probabilistic neural network, and multilayer feed-forward neural networks (MLFFNN) to classify data into sleep apnea or normal classes with a 75/25 training/testing split and 5, 10, and 20-fold cross-validation on the training set. 1,234 arrests were recorded, 722 of which belonged to the apnea class, and 34 features are extracted and reduced to 32 via Mann-Whitney U test. The MLFFNN yielded the best testing performance with an accuracy of 97.07 % and a κ of 0.93 [26]. Hafezi et al. aim to classify sleep apnea events using in-laboratory PSG recordings and an accelerometer attached over subjects' suprasternal notches of 69 subjects in order to develop a model that works well with wearable devices. Baseline drift is removed by band-pass filtering (0.1-25 Hz), 21 features are extracted and prepared as input to the CNN-LSTM classifier with 5-fold cross-validation. The best performance obtained included accuracy of 84.00 %, a sensitivity of 81.00 %, and a specificity of 87.00 % with an AHI-cutoff of 15 events per hour [27]. Hajipour et al. compare logistic regression with RF with tracheal breathing sounds for daytime diagnosis of OSAS. They obtained and processed suprasternal notch microphone recordings from 199 subjects and used the Bonferroni approach to remove redundant features resulting in 85 features used. RF proved better than the least absolute shrinkage and selection operator or LASSO-regularized logistic regression in terms of accuracy, sensitivity, and specificity at an average of 82.10 %, 84.20 %, and 79.50 % compared to 79.30 %, 82.2 %, and 75.8 % [28]. The work done by Schätz et al. [29] similarly focuses on sleep apnea detection using breathing signals acquired by depth sensors with a simple neural network and K-means clustering in contrast with PSG data. Their work shows great promise in use of breathing signals alone acquired via off-the-shelf sensors as input machine learning for apnea detection, with an accuracy of 96.8 % and an F1-score of 92.2 % with their neural network. They use breathing signals and PSG data collected from 57 patients, of which 25 had sleep apnea, and use 70 % of the data segments for training with 5-fold cross-validation [29]. It is important to note that most apneas recorded in this work were also obstructive, showing great applicability to our future work. Keshavarz et al. use the best-ranked features of 231 subjects' PSGs obtained via rank widget with information gain method with 10-fold cross-validation as input to an ANN, NB, LR, KNN, SVM, and RF, of which ANN yields the best accuracy at 74.91 % [30]. Rodrigues Jr et al. also aim to classify OSAS into one of 4 classes based on AHI, and in addition, aim to predict AHI with 60 classification and regression algorithms for data of the 1,042 subjects in the MARS dataset [31]. ExtraTrees yields the best results both as a regressor and classifier, as it yielded an R^2 value of 0.5, a root MSE (RMSE) of 16.25, a mean absolute error (MAE) of 9.6 for regression, and an accuracy of 75 %, an F1-score of 66 %, and an AUC of 85 % for classification [31]. Kim et al. use LR, SVM, RF, and extreme gradient boosting (XGBoost) with a total of 7 features extracted from PSGs and anthropometric data- reduced from 92 via a permutation feature importance algorithm- to predict OSAS in 279 subjects with a 70/30 training/testing split. SVM showed the best OSA prediction results with an accuracy, sensitivity, specificity, F1-score, and AUC of 83.33 %, 80.33 %, 86.96 %, 84.23 % and 0.87, respectively [32]. Works like Almuhammadi's et al. and Vimala's et al. extract and use EEG signals alone from the PSGs in the MIT-BIH database [22], [25], similarly to how Srinivasulu et al. use ECG signals alone [21], [33], [34]. Both works extract the five frequency bands corresponding to the five different brain waves, but Almuhammadi et al. use SVM, ANN, LDA, and NB with 2 features extracted from 30-second EEG epochs for classification with 90 % of the data used for training, whereas Vimala et al. SVM, ANN, and KNN with 14 EEG features with two-thirds of the data used for training and validation. SVM yielded the best result for both setups; Almuhammadi et al. obtained an accuracy, sensitivity, and specificity of 97.14 %, 97.01 %, and 97.26 %, and Vimala et al. obtained 99 %, 100 %, 98 %, respectively [33], [34].

Korkalainen et al.'s work also does not center on detecting OSAS, rather focusing on the effect of sleep apnea severity on sleep staging, considering awake, REM, the two stages of light sleep, and deep sleep. They use PSGs of 153 healthy individuals from the Sleep-EDF dataset [22], [35] and 891 with suspected OSAS from a clinical dataset and a CNN-Bidirectional LSTM model with 10-fold cross-validation, and obtained a maximum staging accuracy of 83.90 % and κ of 0.78 with the public dataset, and a staging accuracy of 83.80 % and κ of 0.78 with the clinical dataset with single-channel EEG and EOG [36].

From a broader perspective, Hussain et al. [37] use EEG signals collected from 157 subjects from the recently published Haaglanden Medisch Centrum sleep staging database [38] in sleep staging. Machine learning techniques utilized in their efforts to perform multi-class classification between wakefulness, non-REM stage, and REM include simple neural network (multi-layer perceptron), C5.0 decision tree algorithm, and Chi-square Automatic Interaction Detection (CHAID) algorithm. Although our work is similar in the methodology they employed in sleep staging, their classification problem involves 5 classes, whereas ours has 3 as we use deep sleep as a singular stage without subdivision into N2 and N3. Furthermore, our machine learning classifiers are simpler and thus may be detrimental to sleep staging performance. EEG signals are filtered at 60 Hz to remove power-line interference, independent component analysis (ICA) was used to remove eye-blink and general muscle artifacts, and the EEG signals were band-passed filtered between 0.5 and 44 Hz (delta to gamma ranges). Afterward, the authors extracted features from the time-domain and frequency-domain EEG signals, followed by feature selection via statistical significance obtained from one-way ANOVA F-test to feed into the machine learning classifiers. The C5.0 model yielded the best classification performance in terms of accuracy at 91.0 %, showing promise for the use of similar methodology in our work [37].

Works regarding detecting depression also vary in methodology; works commonly use EEG [39], [40], ECG [41], speech audio data [42], scoring systems, such as the Depression, Anxiety, and Stress Scale questionnaire (DASS 21) [43], or magnetic resonance imaging (MRI) data to train machine learning models in their efforts to diagnose depression [44]. Mumtaz et al. and Hosseinifard et al. use SVM, LR, and NB with 64 subjects' EEG synchronization likelihood (SL) features and KNN, LDA, and LR with band powers and detrended fluctuation analysis (DFA), Higuchi fractal dimension, correlation dimension, and Lyapunov exponent extracted from 90 subjects' EEG recordings, respectively to diagnose MDD. SVM with 10-fold cross-validation yielded the best results for Mumtaz's et al. study with an accuracy of 98.00 %, a sensitivity of 99.9. %, a specificity of 95.00 % and an F1-score of 97.00 % and LR with LOO cross-validation applied to two-thirds of the data (training set) yielded the best performance for Hosseinifard's et al. study with an accuracy of 90.00 % [39], [40].

Zang et al. use raw ECG recordings of 74 subjects as input to their CNN and obtained an accuracy of 93.96 %, a sensitivity of 89.43 %, a specificity of 98.49 %, and an F1-score of 93.67 % [41]. McGinnis et al. used a novel speech task designed to elicit an anxiety response in 71 children between 3 and 8 years old who were fluent in English and recorded audio data sampled at 48 kHz and processed with a voice activity detector. LR, SVM, and RF with LOO cross-validation were used with eight z-score normalized, Davies-Bouldin Index-selected features, of which LR yielded the best performance with an accuracy of 80.00 %, a sensitivity of 54.00 %, a specificity of 93 %, and an AUC of 0.75 in diagnosing internalizing disorders (anxiety and depression) [42]. Priya et al. used the Depression, Anxiety and Stress Scale questionnaire (DASS 21) on Google forms to collect responses from a total of 348 participants, which was split 70/30 into training and testing sets. This data is then applied to DT, RF, SVM, KNN, and NB, which resulted in the best performance with an accuracy, sensitivity, specificity, and F1-score of 85.50 %, 85.00 %, 91.70 %, and 83.60 %, respectively [43].

A. CONTRIBUTION

As seen previously, works regarding OSAS, depression, and sleep stage classification focus on anthropometric patient information, questionnaire results, and certain electrophysiological signals, such as EEG, ECG, EMG, EOG, etc., and less commonly, audio signals for depression at least. Furthermore, machine learning is a common research stratagem used in detecting/classifying these conditions, as it provides a robust, automated means to classify with lower costs and a lower chance of human bias compared to human classification if trained appropriately. Thus, the main contribution of our work lies in:

- Using a novel dataset that includes ECG/HRV-extracted features; namely lambda, respiratory sinus arrhythmia, very low, low, and high-frequency powers, as well as EEG-extracted brain wave powers for each patient at each instance.
- Classifying sleep stages to assess the possibility of skipping manual scoring, along with classifying depression and OSAS in parallel.
- Classifying OSAS with each of the three sleep stages.
- Classifying depression in OSAS patients with each of the three sleep stages.

In the subsequent sections, we describe the methodology undertaken to collect, process, and use data in classification (Section II), the results obtained from our experimentation (Section III), the discussion of our results, and comparison with similar works in the literature (Section IV), and finally, concluding remarks and future directions in which this work could be expanded (Section V).

II. METHODOLOGY

Before explaining the dataset, features and feature selection processes, machine learning techniques, and performance evaluation methods, the algorithm II below briefly describes the methodology.

1: Data Table, Size = Observations × Features Features = 34 All-Stages: Observations = 1,424 Light Sleep: Observations = 675 REM Sleep: Observations = 476 Deep Sleep: Observations = 273
2: Fill in Missing Values via PCHIP
3: Softmax Normalization (1), Box-Cox Transform (2), Centering and Standardization (3)
4: **for** i = 1 to 9, Where 9 is the Number of Feature Configurations **do**
5: Features in Line **1** of the Algorithm Changes Based on the Problem and Feature Configuration
6: Partition Training/Validation and Testing Data and Labels
7: **for** i = 1 to 14, Where 14 is the Number of Classifiers **do**
8: Train and 10-Fold Validate OSAS, Depression, and Staging Models
9: Test Models and Acquire Performance and Interpretability Metrics and Plots
10: **end for**
11: **end for**
12: **return** Accuracies, Sensitivities, Specificities, Precisions, F1-Scores, Kappa Values For All 126 Models-Feature Configurations
13: **return** Posterior Probability, Confusion Matrix, Testing ROC, and SHAP Plots of Best Model-Feature Configuration For Each Problem
14: **return** Bland-Altman Plot of Best OSAS Model vs Softmax-Normalized AHI

A. SUBJECTS AND DATA COLLECTION

A total of 118 patients (59 male, 57 female) of ages ranging from 1 to 81 years old were considered in this work, those with OSAS making up 80 of the 118 and 6 without OSAS

TABLE 1. Brief description of dataset used.

Data Source	Total Number of Subjects In Dataset	Total Number of Subjects Taken	Subjects with OSAS and Depression	Subjects with OSAS Only	Healthy/Control Subjects	Mean Age (Years)	Collected From
Our dataset	86	86	40	40	6	45.2 ± 13.6	ACPN
STAGES Dataset	1,500	32	0	0	32	"Adult/ Adolescent"	Stanford University, Bogan Sleep Consulting, Geisinger Health, Mayo Clinic, MedSleep, and St. Luke's Hospital
Total	1,586	118	40	40	38		

or depression, all 86 being UAE nationals over 18 years of age. Studies involving these subjects were conducted by the American Center for Psychiatry and Neurology. The remaining 32 of the control group were collected from the Stanford Technology Analytics and Genomics in Sleep (STAGES) dataset [45]. Both datasets are briefly described in Table 1. These patients are split into 3 groups, 40 in the first group, who were depressed with OSAS, 40 in the second, who were not depressed but had OSAS, and the remaining 38 in the control group, who were not depressed and did not suffer from OSAS. The EEG and ECG signals sampled at 200 Hz and 100 Hz, respectively were resampled to the same sampling frequency in order to synchronize the signals, along with breathing signal, sampled at 10 Hz. Furthermore, the former two signals were notch-filtered at 50 Hz to remove power-line interference, and band-pass filtering was applied to the ECG and breathing signals from 0.1-0.4 Hz, and the EEG signals from 0.5-30 Hz. The apneic segments were selected manually by inspection of artifacts pertaining to motion, eye-blinks, and heart beats. Features were extracted from these subjects at specific observation conditions, namely occurrence of apnea, occurrence of depression, and sleep stage the patient is in at that segment. The extracted data included 34 features shown in Table 2, detailing various EEG brain wave powers from channels O1, O2, C3, C4, F3, and F4 with A1 and A2 as reference channels, as shown in Figure 3, ECG/HRV features, such as very low-frequency, low-frequency, and high-frequency powers, the ratio of low-to-high-frequency powers, and features extracted from breathing signals (flow, thorax, oxygen saturation) and ECG/HRV such as respiratory frequency, lambda, and respiratory sinus arrhythmia (RSA). The mean power of the ECG signal is extracted at various frequency ranges, named very low, low, and high frequency, and the low and high frequency powers are normalized. The ratio between low frequency and high frequency powers is also extracted and normalized for a total of 7 features. Respiratory frequency is the sole feature extracted from breathing signals independently of other signals, and, as its name implies, is the breathing rate.

Lambda and respiratory sinus arrhythmia represent phase coupling between ECG (R-R Intervals) and thorax signal, and reduction in R-R interval duration during inspiration and extension during expiration seen in ECG and breathing signals, respectively. This gives us two features pertaining to both ECG and breathing signals. EEG features are entirely independent of ECG and breathing signals, however. The mean powers of frequency bands beta, theta, alpha, and delta are extracted for channels O1, O2, C3, C4, F3, and F4, so 6 channel × 4 bands for a total of 24 features.

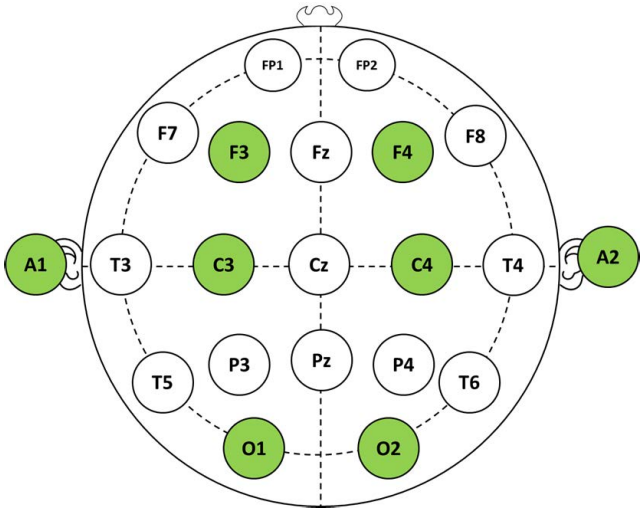

FIGURE 3. 10-20 8-channel electrode configuration. The highlighted electrodes are the ones used.

Data collection and use were approved by the Institutional Review Board (IRB) of the American Center for Psychiatry and Neurology in Abu Dhabi on the 2nd of October 2017, with protocol number or IRB reference number of 0019. Each subject had between 2 and 20 observations, depending on the group to which they belonged, the sleep stage they were in at the time of observation, and whether or not respiratory events occurred during that observation. Each observation represents the 34 features extracted from one 5-minute segment of the signals. A maximum of three segments was selected for each sleep stage, and segments were selected when apneas occurred and otherwise, as well as considering the depression status of the subject, resulting in a total of 1,424 observations.

B. PROBLEM DEFINITION AND PROCESSING

Thus, we can think of four classification problem sets or experiments with this data. For the first experiment, we consider sleep stage a classification target along with each of the

three groups. Thus we have three classification problems in this experiment. These problems are:

- Classifying sleep stage the patient is in for this observation.
- Classifying whether or not the patient has sleep apnea using the group data wherein we label the groups ''Depressed with OSAS'' and ''Not depressed with OSAS'' as ''OSAS+'' and ''Control'' as ''OSAS-''.
- Classifying whether or not the patients who have OSAS are depressed using the group data wherein we omit the ''Control'' group altogether, and keep the remaining two groups ''Depressed with OSAS'' and ''Not depressed with OSAS''.

For the second, third, and fourth experiments, sleep stage is not a classification target but a label used to split the data, so we only have the second and third classification problems mentioned above for light sleep, REM sleep, and deep sleep, respectively, omitting the ''Control'' group observations for classifying depression and keeping all observations for classifying OSAS. As one can note, we use three sleep stages, meaning we omit the awake stage, and combine N1 and N2 of light sleep together to form one class, unlike [36].

For some of these observations, some features have missing values. Since we perform classification on MATLAB, missing values do not necessarily hinder the training process in most classifiers, as some deal with them by omitting the observation altogether, and some train with them present as long as the whole observation does not comprise missing values for all features, though that introduces some bias and necessitates we forgo some important processing steps to be mentioned later. Therefore, it is best to remedy this issue by filling in the missing values before proceeding with classification. To do so, we consider several filling methods and select based on the patterns observed in data and based on average obtained cross-validation performance with all classifiers. Methods used to fill in missing values include filling them in with the nearest non-missing value, linear interpolation of nearest non-missing values, shape-preserving piecewise cubic spline interpolation (pchip), and using KNN to impute missing values. Pchip yielded the best cross-validation performance among these and was henceforth the method used for all subsequent experimentation.

After missing values are filled, the probability distribution of the data is checked, and it was discovered that all features had non-normal distributions, a few were skewed to one side, but most had several peaks. Power transformations were used to ensure our data is normal to improve performance. The non-linear Box-Cox transform was used to make features' probability distribution approximately normal and is shown in Equation 2. The algorithm searches for the value of λ that maximizes the Log-Likelihood Function (LLF) and uses this optimal value to compute the data transform or the result of the equation. The Box-Cox algorithm finds λ by computing the geometric mean of the data to be transformed. Thus the data must all be positive to avoid imaginary numbers. Though this can be remedied by adding a constant value to the data, we perform Softmax normalization, as shown in Equation 1 and feature selection (FS) prior to Box-Cox transform and as described in Subsection II-C. The output of Softmax normalization can be interpreted as a probability distribution, which allows us to gauge if power transformation is necessary, and if so, ensures that the input to the Box-Cox transform is always positive. Applying the Box-Cox transform is followed by centering and standardization to have a mean value of zero and a standard deviation of 1 as shown in Equation 3, also known as z-score normalization.

$$Data_{norm} = \frac{1}{1 + \exp \frac{mean(Data) - Data}{std(Data)}} \tag{1}$$

$$Data_{normpdf}(\lambda) = \begin{cases} \dfrac{Data_{norm}^{\lambda} - 1}{\lambda}, & if\, \lambda \neq 0 \\ log(Data_{norm}), & if\, \lambda = 0 \end{cases} \tag{2}$$

$$Data_{normpdf,cent,std} = \frac{Data_{normpdf} - mean(Data_{normpdf})}{std(Data_{normpdf})} \tag{3}$$

After Softmax normalization, feature selection, probability density function normalization using Box-Cox transform, and centering and standardization, the data is split into two sets. The first set is used for training and validation and comprises the observations of 89 of the 118 patients or 75% of the total number of patients. The second set is used for testing, as it is unseen by the trained classifiers, and comprises the remaining 29 patients, or 25 % of the patients. Data segment labels, being apnea status, depression status, and sleep stage are the output targets corresponding to each observation set. These labels or targets are selected simply based on whether or not an apnea has occurred in the selected segment, whether or not the subject suffers from depression, and the sleep stage the subject was in during that segment. These labels are also split accordingly to the data segments' training/testing split.

We perform experimentation with missing values regardless of omission or bias to witness the effect they may have on results, though without Softmax normalization and Box-Cox transformation. That is because Softmax would change any one column/feature with even a singular missing value into missing values entirely, which does not work with some classifiers, and certainly not with Box-Cox transform, as it requires strictly positive values. Before describing our classifiers, it is best to describe our feature selection methods and feature configurations used.

C. FEATURE SELECTION

Feature selection is an important part of processing extracted features to ensure the input to the classification algorithms would produce the best learning performance i.e. that the model produced would both train well, and generalize well to new data. In this work, we consider nine feature configurations:

1) No feature selection,
2) feature selection by sequential backward feature selection (SBFS) using KNN to select features,

TABLE 2. Selected features for all stages with pchip missing-value-filling and for each stage with pchip filling. *Note: Staging has the same selected features as OSAS but only appears in case A when all stages are kept. (A: All Sleep Stages, L: Light Sleep, R: REM Sleep, D: Deep Sleep).* ***Both OSAS and Depression*** ***OSAS Only*** ***Depression Only*** ***Neither OSAS nor Depression.***

Feature Number	Name	Sleep Stages in Which Features Appear for Each Feature Selection Technique																							
		SBFS				SFFS				MRMR				ReliefF				NCA				Chi2			
		A	L	R	D	A	L	R	D	A	L	R	D	A	L	R	D	A	L	R	D	A	L	R	D
1	Lambda1																								
2	RSA1																								
3	RespFreq1																								
4	normRSA																								
5	vLFpower																								
6	LFpower																								
7	normLFpower																								
8	HFpower																								
9	normHFpower																								
10	LF/HF																								
11	POWER_DELTA_F3																								
12	POWER_THETA_F3																								
13	POWER_ALPHA_F3																								
14	POWER_BETA_F3																								
15	POWER_DELTA_F4																								
16	POWER_THETA_F4																								
17	POWER_ALPHA_F4																								
18	POWER_BETA_F4																								
19	POWER_DELTA_C3																								
20	POWER_THETA_C3																								
21	POWER_ALPHA_C3																								
22	POWER_BETA_C3																								
23	POWER_DELTA_C4																								
24	POWER_THETA_C4																								
25	POWER_ALPHA_C4																								
26	POWER_BETA_C4																								
27	POWER_DELTA_O1																								
28	POWER_THETA_O1																								
29	POWER_ALPHA_O1																								
30	POWER_BETA_O1																								
31	POWER_DELTA_O2																								
32	POWER_THETA_O2																								
33	POWER_ALPHA_O2																								
34	POWER_BETA_O2																								

3) feature selection by sequential forward feature selection (SFFS) using KNN to select features,
4) feature selection using the Minimum Redundancy Maximum Relevance (MRMR) algorithm [46], taking features whose importance score is greater than the mean of dominant scores,
5) feature selection using the ReliefF algorithm [47], taking features whose importance score is greater than zero,
6) feature selection using neighborhood component analysis (NCA), taking features whose weight is greater than the mean of the weights,
7) feature selection using uni-variate feature ranking using chi-square tests, taking features whose importance score is greater than or equal to the mean of all scores (equal is included since the mean is usually infinity),
8) feature selection using principal component analysis (PCA), keeping the first 10 components,
9) feature selection using PCA, keeping components that explain 95 % of the variance, which results in one principal component remaining.

The first configuration involves taking all original 34 features as they are, the eighth and ninth involve getting representations of the features in the principal component space, whereas all the other configurations involve actually operating on and selecting from the initial 34 features. This is done for all sleep stages together and for each one separately, for the original data and for the data with the "Control" group removed. The feature selection criteria for MRMR, ReliefF, NCA, and Chi^2 were chosen based on inspection of plots of feature ranking based on their importance score, like the one shown in Figure 4. Figure 4 shows the feature names, as well as feature ranking using one of our 8 feature selection techniques- MRMR- to select features after pchip was used to fill in missing values and with all three sleep stage classes included.

The total numbers of selected features for all 8 feature selection methods are shown in Table 3 and the selected features are shown in Table 2. We note from the tables that reliefF and sequential backward feature selection tend to preserve more features than the other techniques, and that reliefF applied to deep sleep OSAS data without missing values preserves the largest number of features, as it preserves all features but one.

Feature selection algorithms notwithstanding, seemingly redundant features can be detected manually by inspection and removed prior to applying any of the feature selection methods, but that introduces human bias; features that are not considered redundant by the algorithms could be redundant to experimenters. After feature selection and processing steps are concluded, we are now ready to move forward with training the classifier models and testing them.

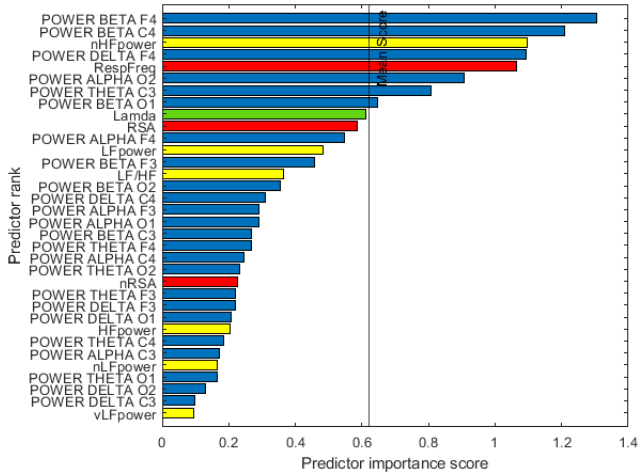

FIGURE 4. Feature ranking with pchip filling and all stage data via MRMR. The horizontal line represents the mean of the dominant features' scores and is the minimum selection threshold.

TABLE 3. Number of selected features for all stages with pchip missing-value-filling, and for each stage with pchip.

OSAS and Stage Classification						
Stage/FS Algorithm	SBFS	SFFS	MRMR	ReliefF	NCA	Chi^2
Pchip	24	17	8	33	9	9
Light Sleep	19	13	5	24	7	13
REM Sleep	24	16	4	28	19	12
Deep Sleep	17	8	4	33	7	13
Depression Classification						
Stage/FS Algorithm	SBFS	SFFS	MRMR	ReliefF	NCA	Chi^2
Pchip	22	14	4	25	27	6
Light Sleep	23	10	5	24	8	15
REM Sleep	19	16	4	26	4	17
Deep Sleep	16	8	5	28	5	18

D. MACHINE LEARNING ALGORITHMS, CLASSIFICATION, AND PERFORMANCE EVALUATION

As observed in Section I, machine learning is a commonly used technique in detecting OSAS, not only because it provides an automated method with which to classify apnea events, but also due to the diagnostic insights its results can give. These insights can allow clinical personnel to focus on recording signals or components of sleep staging with best classification performance or allowing them to forgo some steps, such as computing apnea-hypopnea indices (AHI).

Machine learning is also commonly used as an automated method to detect depression and can give similar insights into depression scoring systems, components of depression that can be focused on to provide a faster yet similarly accurate diagnosis, and perhaps even the causes and effects of the condition itself. Depression scoring demands patients to voluntarily divulge information, which is more difficult to some than simply having signals recorded as they sleep, so a less personal, more automated method of obtaining and analyzing data could be preferable. The main focus of our work is developing machine learning models that work towards detecting OSAS and detecting depression in OSAS patients, and finding the most optimal features and sleep stages to record in order to detect both of them.

As stated earlier, the data are split by subject into training-and-validation/testing sets 75/25, meaning the data of 89 subjects are used for training and validation and the data of the remaining 29 subjects are used for testing. We used 10-fold cross-validation with our training-and-validation set, meaning 9 folds are used for training and 1 for validation, and the model updates as that process iterates 10 times. Cross-validation ensures the model is less likely to over-fit because the entire set is used in tuning it. Hence, if one fold over-fits or ''memorizes'' the data instead of ''learning'', it can be offset by the other folds, generally resulting in a more realistic model.

After the data is split, classifiers have to be considered. Different classifiers have different principles of function; probabilistic models, which depend on computing the probability of a class occurring using the input random variables (observations and target classes), geometric models- which can further be split into linear models, like SVM, or distance-based models, like KNN- work by splitting the different classes with an optimal hyperplane, or placing neighboring observations into the same class using certain criteria (e.g. in KNN, the criterion is k), and logical models, like decision trees (DTs), which work by using logical expressions to divide observations. In contrast, classifiers like ANNs work by back-propagating and updating weights to obtain minimum loss, not reliant on separating classes per se, be it by an optimal hyperplane or distance. We used MATLAB's ClassificationLearner application to train around 30 supervised learning models:

- DTs (fine, medium, coarse).
- Discriminant Analysis (linear, quadratic).
- Logistic Regression.
- Naive Bayes (Gaussian distribution, kernel distribution).
- Linear, quadratic, cubic, coarse, and Gaussian/RBF SVM (fine, coarse, medium).
- KNN (fine, medium, coarse, cubic, cosine, weighted).
- Ensemble Classifiers: Boosted Trees (AdaBoost, LogitBoost, GentleBoost), Bagged Trees, Random Forest, Subspace Discriminant, Subspace KNN, RUSBoosted Trees.
- ANNs (Narrow, medium, wide, bi-layer, tri-layer).

These 30 models mostly comprise the same algorithms, but with different parameters which are usually followed by a model obtained with hyperparameter optimization using Bayesian Optimization. After preliminary optimization and trials with these classifier configurations in MATLAB's ClassificationLearner, we select the ones with the best validation performance, and narrow down our list of used classifiers to the following fourteen with the following options:

1) Gaussian distribution NB.
2) LDA with a gamma (regularization term) of 0.
3) DT with 100 maximum number of branch nodes split by Gini's diversity index.
4) KNN with k = 1 using Euclidean distance.
5) Gaussian/RBF SVM with automatically computed kernel scale and a box constraint of 1.

6) Bagged Trees with 1290 maximum number of branch nodes and 30 learning cycles.
7) Random Forest (bagged trees with surrogate decision split).
8) AdaBoost with 30 maximum number of branch nodes Decision Trees.
9) RUSBoost with 30 maximum number of branch nodes Decision Trees.
10) LogitBoost with 30 maximum number of branch nodes Decision Trees.
11) GentleBoost with 30 maximum number of branch nodes Decision Trees.
12) Subspace KNN with 30 learning cycles.
13) Subspace Discriminant with 30 learning cycles.
14) Bi-layer ANN with a lambda (regularization term) of 0, and 100 units in each of the two hidden layers.

Naive Bayes is an algorithm that applies density estimation to the data using Bayes theorem under the assumption that the features are conditionally independent, and uses Bayes theorem with additive Gaussian smoothing to estimate the posterior probability in the case of Gaussian naive Bayes. The classifier then assigns the observation to the class with the highest posterior probability.

LDA is an algorithm that assigns observations to the class with the highest posterior probability, and works by reducing the expected classification cost. LDA assumes the data has a Gaussian mixture distribution with the same covariance matrix for all classes, yet still generally performs well even if these assumptions are violated.

DTs compute the weighted impurity of a node, and estimate the probability of an observation belonging to that node. The tree then splits the observations in a node and selects the class that yields the largest impurity gain.

KNN finds the k observations and response class in the training set that are nearest to the observation to be predicted, and assigns the label with the highest posterior probability in the response class to this observation.

SVM models involve mapping features of a binary classification problem into a kernel space and finding an optimal hyperplane in that space that separates the two classes. For multi-class problems with n classes, n binary SVMs are trained and then combined either with each class against all the other classes (one-vs-all), or each two classes against each other (one-vs-one). This also goes for classifiers that are binary by default, such as binomial logistic regression. In our work, as stated in the numbered list above, we use RBF kernel SVM, which applies the RBF to our input. Equation 4 describes the radial basis function, where G is element (i,j) of the Gram matrix, x_i and x_j are vectors that represent observations i and j in.

$$G(x_i, x_j) = \exp(-||x_i - x_j||^2 \tag{4}$$

Ensembling classifiers is a method that improves classification performance at the expense of computational cost, as it involves combining the ''weak'' learners in one of two ways in order to make a ''strong'' learner. The weak learner can be trained independently in parallel and then combined in the process known as bootstrap aggregation or bagging, or they can be trained sequentially in the process known as boosting. Bagging involves the use of random subspace, uses the weighted average results of the weak learners, and aims to reduce variance, whereas boosting uses gradient descent, majority vote of the weak learners, and aims to reduce classification error. Our bagging algorithms include random forest, bagged trees, subspace KNN, and subspace discriminant, and our boosting algorithms include AdaBoost, RUSBoost, LogitBoost (for the binary problems), and GentleBoost (for the binary problems), all with decision trees as the weak learners.

Artificial neural networks (ANNs) generally make use of the back-propagation algorithm to minimize a loss function, automatically extracting features. In simpler terms, the first input is used to update the initial weights to yield a predicted output, which is then compared with the actual output that corresponds to this input and the error is computed. The network then updates the weights and yields a different output in an attempt to reduce this error. This repeats until an error threshold is reached or until a maximum number of iterations is reached. Despite their advantages compared to other machine learning techniques, neural networks require large amounts of data in order to see an improvement in performance. After we use these models to train and validate all data with all feature selection configurations, we test them on the holdout/testing set with 29 subjects' data. Classification performance is measured by accuracy, sensitivity, specificity, precision, F1-score, AUC, and Cohen's κ coefficient. Accuracy gives a measure of how many observations the model classified correctly out of the total number of observations, sensitivity measures the correctly classified observations of the positive class out of the total of actually positive predictions, and specificity measures the correctly classified observations of the negative class out of the total of negative predictions, precision measures the positive predictive value or the number of true positive observations divided by the total of predicted positive observations, F1-score is a harmonic mean that represents sensitivity and precision, AUC is a summary of the ROC curve that represents sensitivity and the inverse of specificity and measures the ability of a classifier to distinguish between classes, and Cohen's κ tells us how much better this classifier is than chance, considering class distribution as well as the number of classes.

Interpretability is becoming an increasingly important aspect of machine learning in order to bridge the divide between machine learning specialists and clinical personnel. It is imperative to translate technical results into clinical terms to avoid confusion. Methods such as Shapley values, SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and partial dependence plots can all be used to provide explanations regarding feature importance and classification results. We will be using SHAP plots to explain the best-performing models for each of our classification problems.

E. LIMITATIONS

While the results obtained were satisfactory, as shown in Section III, the methodology has its limitations. One of which is the manual optimization of hyperparameters or using default settings in the software instead of using an algorithm like grid search or Bayesian optimization. Though that is associated with part of our contribution – developing a methodology that is relatively computationally inexpensive – the machine learning models may yield better results if either one of those techniques were used for hyperparameter optimization. Another limitation that also pertains to computational cost is associated with the use of classical machine learning algorithms and largely forgoing deep learning techniques. Deep learning techniques, particularly long short-term memory (LSTM) networks proved extremely time-consuming and taxing to the available hardware at the time compared to the stated machine learning techniques in preliminary testing. The use of deep learning techniques with better hardware could yield even better results with likely no increase in training time.

We can generally get an idea about associated computational cost by looking at the number of associated parameters and calculation, and subsequently how long it would take to train a model. However, in our analysis, computation cost was quantified in a simpler, binary fashion. More computationally expensive i.e., more parameters, more calculations, models would simply throw a memory error on MATLAB and fail to train. These models include convolutional neural networks, long short-term memory networks, and simple neural networks with larger hidden layer sizes, more hidden layers, or an extremely low initial learn rate. The device used to obtain these results is relatively low-cost and is standard-issue at Khalifa University. With no GPU and a weak Intel CPU, a new device with a powerful GPU (Nvidia GeForce RTX 3080) would not only allow us to use more complex machine learning techniques but would also open the possibility of using full PSG signals in training.

III. RESULTS

As stated earlier, we compute the testing accuracy, sensitivity, specificity, precision, F1-score, AUC, and Cohen's κ coefficient for all feature configurations with all classifiers. Since we have 14 classifiers, 9 feature configurations, and 2 or 3 classification problems per experiment, meaning we have large tables of results, we have elected to describe the results and showcase the performance of the best classifier and feature configuration for each problem in each experiment, as shown in Tables 4-7. Overall performance is gauged mainly by F1-score, accuracy, and κ considered together.

A. EXPERIMENT 1: CLASSIFICATION OF OSAS, DEPRESSION IN OSAS PATIENTS, AND SLEEP STAGE

The first experiment or problem set involves detecting OSAS, Depression in OSAS patients, and classifying sleep stages for data with missing values and after filling in missing values in order to observe the effect that filling them in has. Table 4 shows the best results for these problems. We can see a slight decrement in results when missing data was filled in, due to the inherent bias present in using missing values for some classifiers, or the complete omission of rows with missing values for other classifiers. We can also see that RF with features selected via SBFS yielded the best results for OSAS detection, ANN with features extracted via Chi^2 yielded the best results for detecting depression in OSAS patients, and RF without feature selection yielded the best performance for sleep stage classification.

TABLE 4. Summary of the best algorithm and feature selection configuration for OSAS, Depression with OSAS, and Sleep stage classification with missing values filled in via pchip.

	OSAS	Depression with OSAS	Sleep Stage
FS Technique	SBFS	Chi^2	No FS
Model	RF	ANN	RF
AUC	0.85	0.84	0.79
Accuracy (%)	91.18	72.95	70.52
Sensitivity (%)	96.44	70.95	69.35
Specificity (%)	73.17	75.19	83.69
Precision (%)	92.49	76.09	70.78
F1-Score (%)	94.43	73.43	69.99
$\varkappa$	0.73	0.46	0.34

TABLE 5. Summary of the best algorithm and feature selection configuration for OSAS, and Depression with OSAS for light sleep data.

	OSAS	Depression with OSAS
FS Technique	No FS	ReliefF
Model	Bagged Trees	GentleBoost
AUC	0.84	0.71
Accuracy (%)	90.85	70.83
Sensitivity (%)	98.61	74.07
Specificity (%)	35.00	68.89
Precision (%)	91.61	58.82
F1-Score (%)	94.98	65.57
$\varkappa$	0.44	0.41

B. EXPERIMENT 2: CLASSIFICATION OF OSAS AND DEPRESSION IN OSAS PATIENTS WITH LIGHT SLEEP

The second experiment involves detecting OSAS and Depression in OSAS patients using only the patient data whose sleep stage was labeled ''light sleep''. Table 5 shows the best results for these problems. We can see that bagged trees without feature selection yielded the best results for OSAS detection and boosted trees with GentleBoost with ReliefF yielded the best results for detecting depression in OSAS patients. However, OSAS classification shows low specificity and a relatively low κ value compared to pchip and the following sleep stages.

C. EXPERIMENT 3: CLASSIFICATION OF OSAS AND DEPRESSION IN OSAS PATIENTS WITH REM SLEEP

The third experiment involves detecting OSAS and Depression in OSAS patients using only the patient data whose sleep stage was labeled ''REM sleep''. Table 6 shows the best

TABLE 6. Summary of the best algorithm and feature selection configuration for OSAS, and Depression with OSAS for REM sleep data.

	OSAS	Depression with OSAS
FS Technique	SBFS	No FS/ReliefF
Model	RF	KNN/Subspace KNN
AUC	0.96	0.75
Accuracy (%)	98.13	67.03
Sensitivity (%)	100.00	60.00
Specificity (%)	87.50	90.48
Precision (%)	97.85	95.45
F1-Score (%)	98.91	73.68
$\varkappa$	0.92	0.35

TABLE 7. Summary of the best algorithm and feature selection configuration for OSAS, and Depression with OSAS for deep sleep data.

	OSAS	Depression with OSAS
FS Technique	No FS/ReliefF	PCA10
Model	SVM/Subspace Discriminant	DT
AUC	1.00	0.75
Accuracy (%)	98.36	73.81
Sensitivity (%)	100.00	90.00
Specificity (%)	94.74	68.75
Precision (%)	97.67	47.37
F1-Score (%)	98.82	62.07
$\varkappa$	0.96	0.45

results for these problems. We can see that RF with SBFS yielded the best results for OSAS detection and both Subspace KNN and KNN with ReliefF and no feature selection respectively, yielded the best results for detecting depression in OSAS patients.

D. EXPERIMENT 4: CLASSIFICATION OF OSAS AND DEPRESSION IN OSAS PATIENTS WITH DEEP SLEEP

The final experiment involves detecting OSAS and Depression in OSAS patients using only the patient data whose sleep stage was labeled "deep sleep". Table 7 shows the best results for these problems. SVM (in addition to LDA and subspace discriminant) without feature selection and subspace discriminant with reliefF yielded the best results for OSAS detection, whereas DT with 10 principal components following PCA yielded the best results for detecting depression in OSAS patients.

We note that deep sleep data, followed closely by REM sleep data, yield the best overall performance for detecting OSAS, and all sleep stage data and REM sleep data yield the best overall performance for detecting depression in OSAS patients. Sleep stage classification results appear only when sleep stages are not a classification target, so the best-observed performance is that described in subsection III-A. These best overall results are summarized in Table 8. The confusion matrices of OSAS, depression and sleep stage classification can be found in Figures 5-7. Figures 14-20 show the posterior probability plots obtained for all patient data relevant to the respective sleep stage(s) with the best set up for each problem (sleep stage(s), feature configuration, classifier, etc.). The Bland-Altman plot in Figure 15 shows the agreement between the AHI and our best classifier's results for diagnosing OSAS. Additionally, the ROC plots

TABLE 8. Summary of best overall setups for each problem.

	OSAS	Depression with OSAS	Sleep Stage
FS Technique	No FS/ReliefF	Chi^2	No FS
Model	SVM/Subspace Discriminant	ANN	RF
Sleep Stage	Deep Sleep	All stages	N/A
AUC	1.00	0.84	0.79
Accuracy (%)	98.36	72.95	70.52
Sensitivity (%)	100.00	70.95	69.35
Specificity (%)	94.74	75.19	83.69
Precision (%)	97.67	76.09	70.78
F1-Score (%)	98.82	73.43	69.99
$\varkappa$	0.96	0.46	0.34

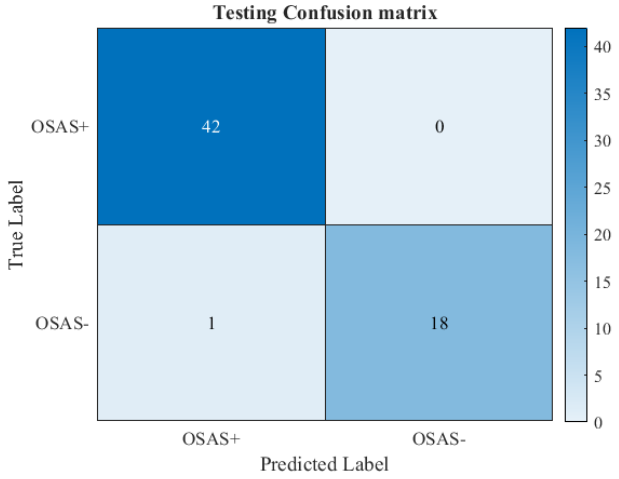

FIGURE 5. Testing confusion matrix of best OSAS detection setup.

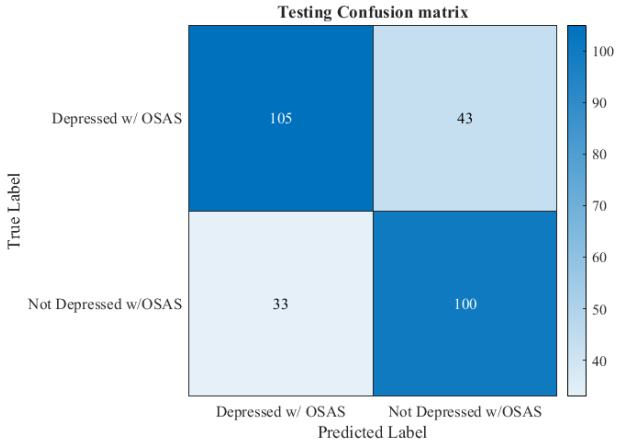

FIGURE 6. Testing confusion matrix of best detection of depression in OSAS setup.

are shown in Figures 8- 13 and the SHAP plots shown in Figures 16- 18 show the top 10 features (except in depression, where we have only 6 features) in prediction, using the first observation as our query point.

IV. DISCUSSION OF RESULTS

In this section, we explain the possible rationale behind the results observed in Section III, and make clinical recommendations based on them. Sleep staging plays a critical role in our work, as we can use it to observe how dependent classifying OSAS and depression in OSAS patients with this dataset and methodology are on sleep stages, thus we begin with it. Seeing no significant difference in classification results of OSAS (ANOVA p-value $= 0.14 > 0.05$) and depression

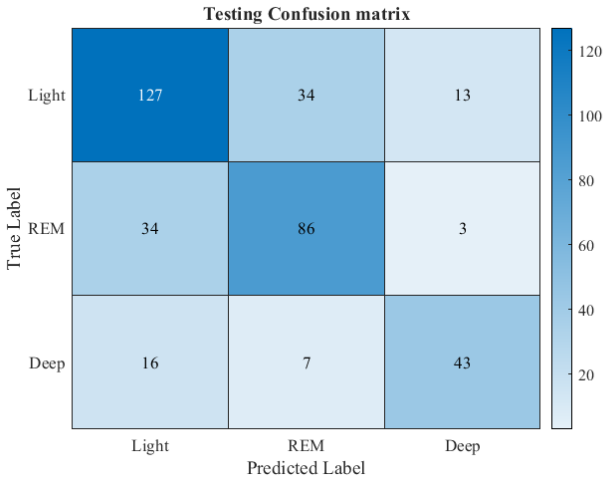

FIGURE 7. Testing confusion matrix of best sleep stage detection setup.

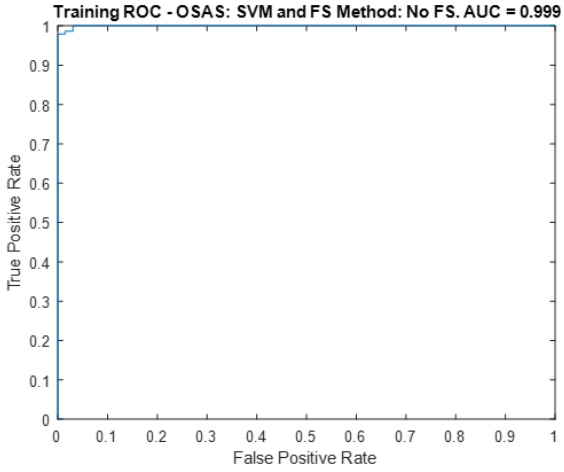

FIGURE 8. ROC plot of the best OSAS detection model in training/cross-validation.

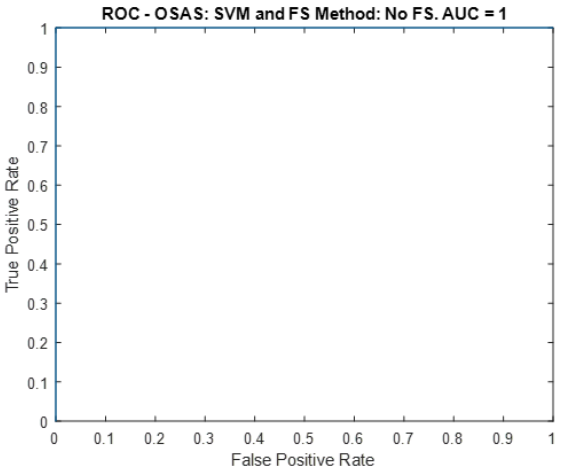

FIGURE 9. ROC plot of the best OSAS detection model in testing.

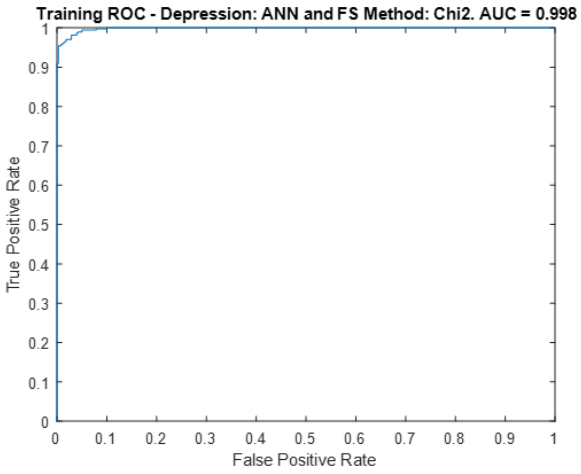

FIGURE 10. ROC plot of the best depression detection model in training/cross-validation.

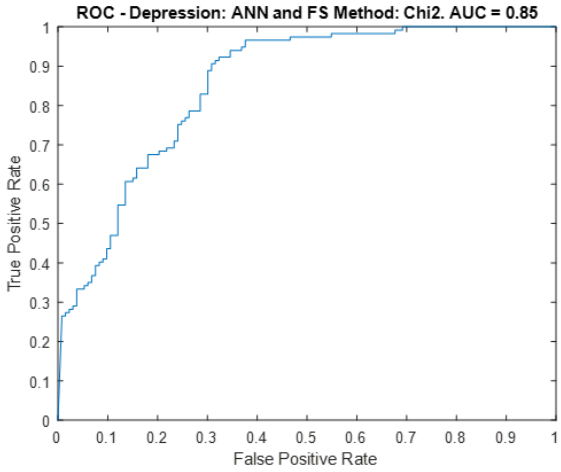

FIGURE 11. ROC plot of the best depression detection model in testing.

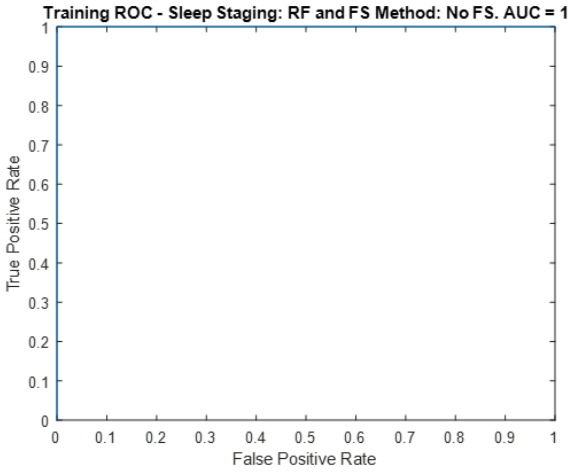

FIGURE 12. ROC plot of the best staging model in training/cross-validation.

in OSAS (ANOVA p-value $= 0.58 > 0.05$) in Tables 4 through 7, we conclude that sleep stages have no significant effect on OSAS and depression in OSAS classification performance with our best models.

That said, observations of patients that belonged to the deep sleep stage yielded better OSAS detection results than the other setups. Not to mention, the classes are clearly separated by looking at the posterior probability plot, which

TABLE 9. Comparison between the proposed methodology and similar works. *OSAS: Obstructive Sleep Apnea Syndrome, PSG: Polysomnography, EEG: Electroencephalography, EOG: Electrooculography, ECG: Electrocardiography, ANN: Artificial Neural Networks, NB: Naive Bayes, LR: Logistic Regression, KNN: K-th Nearest Neighbor, SVM: Support Vector Machine, RF: Random Forest, XGBoost: Extreme Gradient Boosting, CNN: Convolutional Neural Network, BiLSTM: Bidirectional Long Short-Term Memory Network, LDA: Linear Discriminant Analysis, DT: Decision Tree, FS: Feature Selection.*

Work	Keshavarz *et al.* [30]	Kim *et al.* [32]	Korkalainen *et al.* [36]	Hosseinifard *et al.* [40]	Proposed Method
Main Objective	Predict OSAS	Predict OSAS	Sleep Staging/ Classify 5 Sleep Stages, then measure effect of OSA severity on sleep staging	Classify Depression	Classify 3 Sleep Stages then Classify OSAS and Depression in patients with OSAS
Dataset	231 subjects' PSGs	279 subjects' PSGs and anthropometric data	1,044 subjects' PSGs, EEG and/or EOG data used (153 Sleep-EDF, 891 Clinical Dataset)	90 subjects' EEGs + 4 non-linear features	1,424 observations extracted from EEG, ECG, and breathing signals of 118 subjects
Machine Learning Algorithms	ANN, NB, LR, KNN, SVM, and RF	LR, SVM, RF, and XGBoost	CNN-BiLSTM	KNN, LDA, and LR	NB, LDA, DT, KNN, SVM, Bagged Trees, RF, Boosted Trees, Subspace KNN, Subspace Discriminant, ANN
Significance	The methodology is standard (CRISP-DM), they use a novel dataset with a decent number of samples, the machine learning models are simple, and cross-validation is used	Demographics, models, and results are well-described, the authors developed a web page-based application for prediction, and they used polyomnography signals as well as questionnaires	The machine learning model used is robust, the authors apply it on their own clinical dataset in addition to a known dataset (Sleep-EDF), the model based application for prediction, and they used polyomnography signals shows good results and cross-validation is applied	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	Computationally inexpensive method for detection of OSAS and depression in OSAS is developed using a novel dataset. Results support the reduction of extracted features and recorded sleep stages
Limitations	Classification performance could be better, more metrics can be used to gauge performance, and description of features would be beneficial	Better classification performance was witnessed in the literature, and validation/testing paritioning is not well-explained	One algorithm used, using others for comparison may be beneficial	No significant limitations found, perhaps the methodology may be complex	Deep learning not thoroughly explored, and no automatic hyperparameter optimization via grid-search or Bayesian optimization
OSAS - Accuracy (%)	ANN: 74.91	SVM: 83.33	N/A	N/A	No FS, SVM, AND ReliefF, Subspace Discriminant - Deep Sleep: 98.36
Depression in OSAS - Accuracy (%)	N/A	N/A	N/A	LR: 90.00	Chi^2, ANN - All Stages: 72.95
Sleep Stage - Accuracy (%)	N/A	N/A	Sleep-EDF: 83.90 Clinical Dataset: 83.80 OSAS Severity: 84.5% for individuals w/ OSAS diagnosis 76.5% for patients w/ severe OSAS.	N/A	No FS, RF: 70.52

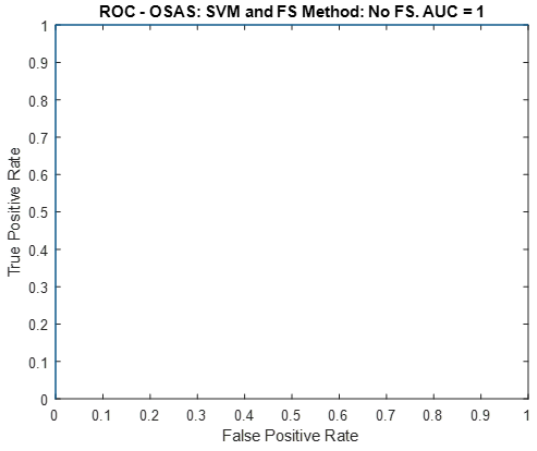

FIGURE 13. ROC plot of the best staging model in testing.

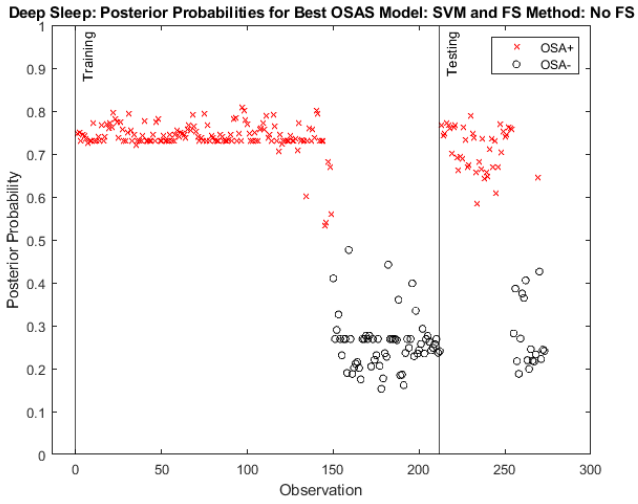

FIGURE 14. Posterior probabilities of the best OSAS detection setup (deep sleep, features selected with reliefF algorithm and classified with bagged trees). Posterior probabilities of the training and testing sets are separated as seen in the figure.

means our classifier leaves little room for misclassification. The best OSAS classifier incorrectly classifies only one out of 61 testing instances; an observation that belonged to the OSAS- group was misclassified as OSAS+. These results tell us that deep sleep is the stage most relevant in detecting OSAS, which would save clinicians time, computational cost, and effort when recording data for the purpose of discerning whether or not a patient has OSAS. The data used for OSAS+ patients included two distinct classes as well; depressed and not depressed, yet the model managed to accurately classify OSAS. This implies that the distinction between OSAS+ and OSAS- is clear, and/or that it may be difficult to distinguish depressed and non-depressed observations, at least when coupled with OSAS. The results obtained for detecting

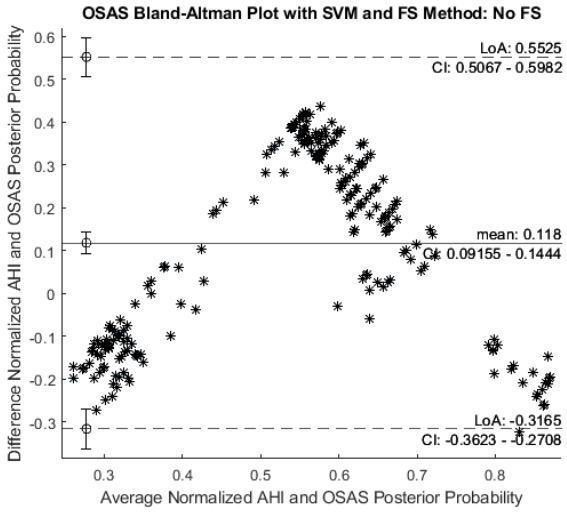

FIGURE 15. Bland-Altman plot of the relationship between the difference between normalized apnea-hypopnea index (AHI) and the posterior probability of the best OSAS detection model against their mean. The confidence interval (CI) and limits of agreement (LoA) are denoted in the figure, LoAs are the Mean±2*Standard Deviation.

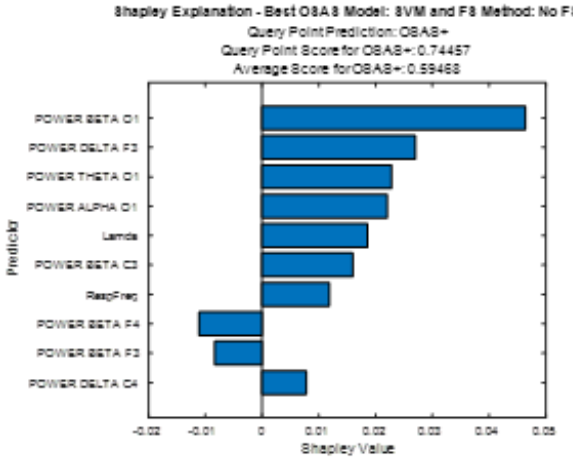

FIGURE 16. SHAP plot showing the top 10 features of the best OSAS detection model using the first observation as our query point.

depression were not as good as detecting OSAS, so that could support the latter hypothesis. We notice more confusion and therefore some difficulty in classifying depressed patients than non-depressed ones. However, the model provides an acceptable F1-score of 73.43 % with an acceptable κ value of 0.46, so both hypotheses hold up to an extent. By the same token, best depression results occurring when all sleep stage data are used could imply a lack of distinction between sleep stages, and/or that depression is independent of sleep stages. Seeing how sleep stage classification goes from the testing results and confusion matrix with an F1-score of 69.99 and a κ of 0.34, we can also say that both hypotheses hold up to an extent. The κ value is lower than 0.4, which tells us that while this classifier still performs better than chance, it is not significantly better. From the confusion matrix, we see that confusion in deep sleep is slightly higher than that in REM sleep, which is in turn slightly higher than that in light sleep.

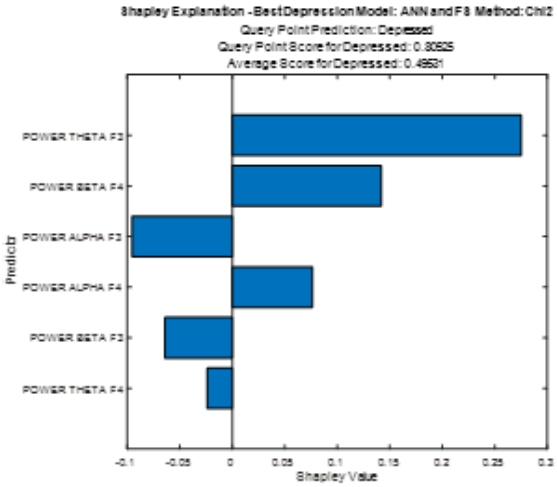

FIGURE 17. SHAP plot showing the top 10 features of the best depression detection model using the first observation as our query point.

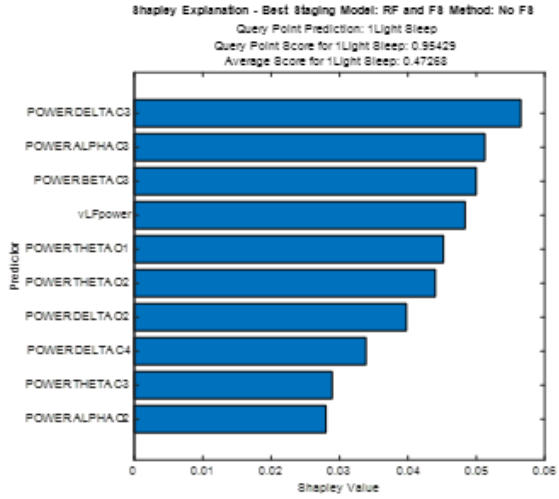

FIGURE 18. SHAP plot showing the top 10 features of the best staging using the first observation as our query point.

Due to the somewhat sub-par performance of sleep stage classification and the lack of statistically significant differences in the results obtained, we recommend using deep or REM sleep data if detecting OSAS and depression in patients with OSAS and avoiding all sleep stage data.

Our results further differ to the compared works because the dataset used includes subject with both depression and OSAS, the coincidence of which could influence the classification of either condition. As we can see, this may already be the cause of the decreased performance of depression classification; perhaps the focus on sleep apnea in the protocol design phase of the sleep study has biased the data to better work with OSAS rather than depression. Another possible conclusion could be the inverse; the occurrence of depression in patients suffering from OSAS could have made classifying OSAS easier.

In Table 9, we compare our results with similar works. Because sources that seek to classify depression in OSAS patients are scarce, much less those that do so following sleep staging, we compare our work with [30], [32], [36], [40], who

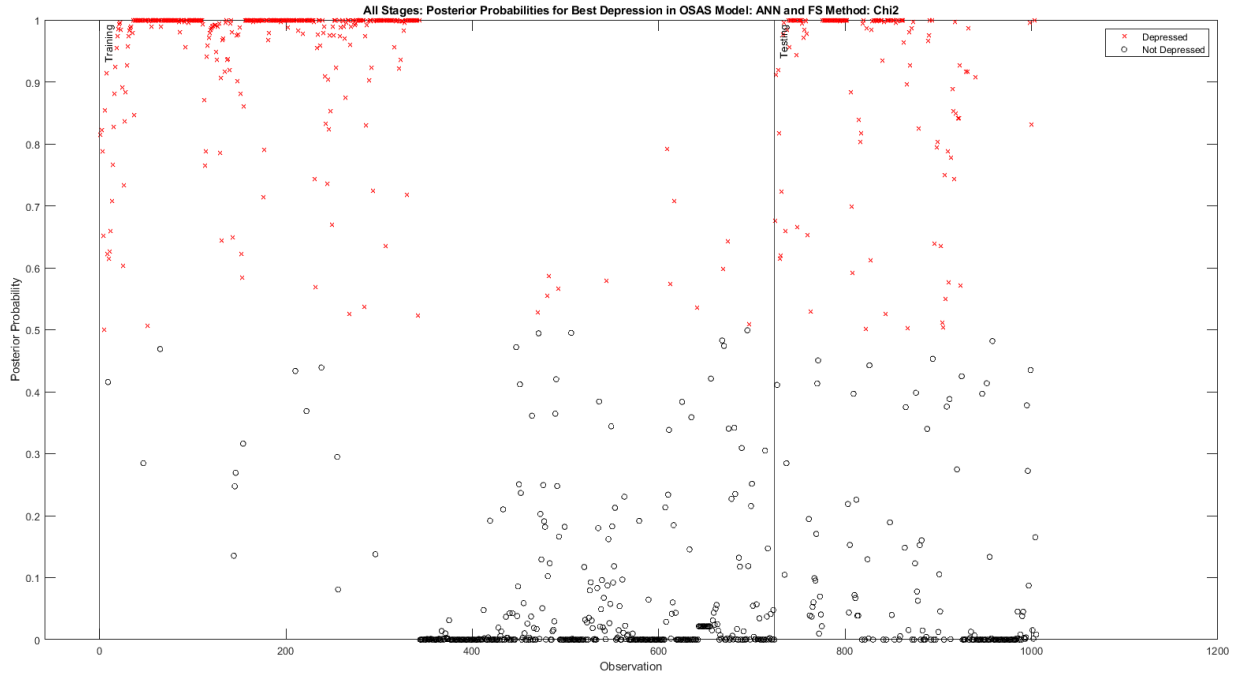

FIGURE 19. Posterior probabilities of the best depression in OSAS patients detection setup (all sleep stages, features selected with Chi^2 and classified with ANN). Posterior probabilities of the training and testing sets are separated as seen in the figure.

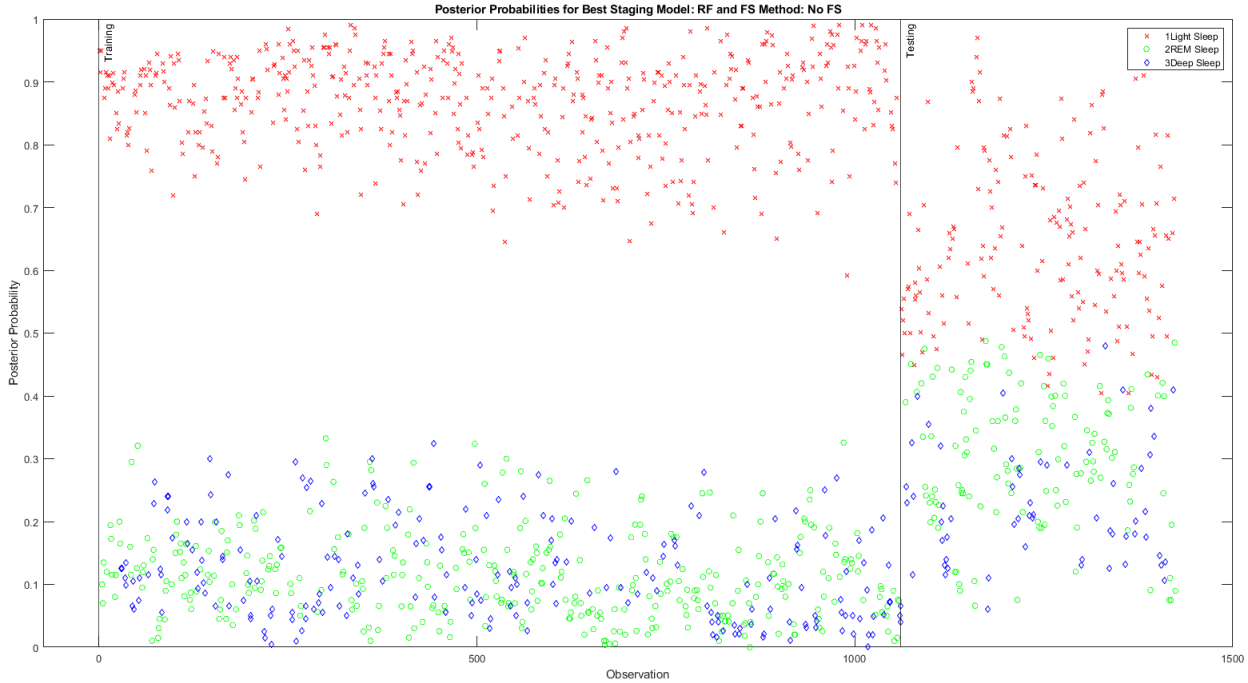

FIGURE 20. Posterior probabilities of the best sleep stage detection setup (all sleep stages, all features used and classified with subspace KNN). Posterior probabilities of the training and testing sets are separated as seen in the figure.

perform at least one of our three classification tasks (OSAS, sleep stage, depression).

We can see from Figures 16-18 that the mean powers of EEG channels are most significant in each classification problem, more apparently in classification of depression. Lambda and respiratory frequency are noted among the top features in classification of OSAS, and very low-frequency ECG power is noted among the top features in sleep staging, but the remaining features all pertain to EEG. We also note no general patterns in the significance of one brain wave/frequency band over another, likewise with channels. Although we notice that frontal channel features do not appear among the top features for sleep staging, are exclusively the top (and only) features in depression classification, and one of the occipital channels' features (O2) do not show up among top features in OSAS classification.

Though works that center around detecting OSAS in depressed patients are limited, we compare the performance of our separate models for sleep staging, classifying OSAS, and classifying depression with works that deal with these individual problems [30], [32], [36], [40] as seen in Table 9. While there is a relatively major difference between our classification of depression and Hosseinifard's *et. al* [40] due to the involvement of OSAS in our work, both works do still classify depression. The first two works, as stated in Section I, classify OSAS alone. Our work shows an improvement in classification of OSAS but a reduction in sleep staging and depression detection performance when pit against the compared works shown in Table 9. However, our dataset is unique in the coincidence of OSAS and depression in approximately a third of our subjects.

V. CONCLUSION

To sum up, the objective of this work was mainly to classify OSAS and depression in patients with OSA. The dataset included overnight EEG and ECG recordings from 118 subjects, 40 were depressed with OSAS, 40 were not depressed but had OSAS, and the remaining 38 were not depressed and did not suffer from OSAS with a total of 1,424 observations. Afterward, we process the data to ensure no NaNs are present and that the data is normal before feature selection. After processing, 14 classifiers are trained and 10-fold cross-validated with the data of 75 % of the patients, or 89 patients, and tested with the remaining 25 %, or 29 patients. For OSAS detection, we obtained a maximum accuracy of 98.36 %, an F1-score of 98.82 %, a κ of 0.96, and an AUC of 1.00. For detecting depression in patients with OSAS, all sleep stages together with Chi^2 for feature selection and ANN for classification yielded the best performance with an accuracy of 72.95 %, F1-score of 73.43 %, a κ of 0.46, and an AUC of 0.84. Finally, random forest without feature selection was the best method for sleep stage classification, yielding an accuracy of 70.52 %, an F1-score of 70.78 %, a κ of 0.34, and an AUC of 0.79. Future work includes using these same methods, in addition to more advanced deep learning methods, such as long short-term memory (LSTM) networks and convolutional neural networks (CNNs) with the raw polysomnography (EEG and ECG) data of these patients.

ACKNOWLEDGMENT

The authors would like to thank the American Center for Psychiatry and Neurology (ACPN), 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), 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

[1] M. R. Mannarino, F. Di Filippo, and M. Pirro, "Obstructive sleep apnea syndrome," *Eur. J. Int. Med.*, vol. 23, no. 7, pp. 586–593, Oct. 2012.
[2] M. Li, X. Li, and Y. Lu, "Obstructive sleep apnea syndrome and metabolic diseases," *Endocrinology*, vol. 159, no. 7, pp. 2670–2675, Jul. 2018.
[3] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 37–48, Jan. 2009.
[4] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," *New England J. Med.*, vol. 328, pp. 1230–1235, Apr. 1993.
[5] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea: A population health perspective," *Amer. J. Respiratory Crit. Care Med.*, vol. 165, no. 9, pp. 1217–1239, May 2002.
[6] R. N. Aurora, N. A. Collop, O. Jacobowitz, S. M. Thomas, S. F. Quan, and A. J. Aronsky, "Quality measures for the care of adult patients with obstructive sleep apnea," *J. Clin. Sleep Med.*, vol. 11, no. 3, pp. 357–383, Mar. 2015.
[7] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: From classification to treatment outcome prediction," *CNS Neurosci. Therapeutics*, vol. 24, no. 11, pp. 1037–1052, Nov. 2018.
[8] C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, and V. Aboyans, "Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010," *Lancet*, vol. 380, no. 9859, pp. 2197–2223, Dec. 2012.
[9] M. Strock, "Depression. National institute of mental health," NIH Publication, Bethesda, MD, USA, Tech. Rep., 2002. [Online]. Available: https://www.nimh.nih.gov/health/topics/depression
[10] I. Hussain and S. J. Park, "HealthSOS: Real-time health monitoring system for stroke prognostics," *IEEE Access*, vol. 8, pp. 213574–213586, 2020.
[11] I. Hussain and S. J. Park, "Big-ECG: Cardiographic predictive cyber-physical system for stroke management," *IEEE Access*, vol. 9, pp. 123146–123164, 2021.
[12] W. Yue, W. Hao, P. Liu, T. Liu, M. Ni, and Q. Guo, "A case—Control study on psychological symptoms in sleep apnea-hypopnea syndrome," *Can. J. Psychiatry*, vol. 48, no. 5, pp. 318–323, 2003.
[13] E. Björnsdóttir, B. Benediktsdóttir, A. I. Pack, E. S. Arnardottir, S. T. Kuna, T. Gíslason, B. T. Keenan, G. Maislin, and J. F. Sigurdsson, "The prevalence of depression among untreated obstructive sleep apnea patients using a standardized psychiatric interview," *J. Clin. Sleep Med.*, vol. 12, no. 1, pp. 105–112, Jan. 2016.
[14] S. M. Ejaz, I. S. Khawaja, S. Bhatia, and T. D. Hurwitz, "Obstructive sleep apnea and depression: A review," *Innov. Clin. Neurosci.*, vol. 8, no. 8, p. 17, 2011.
[15] F. Bozkurt, M. K. Uçar, M. R. Bozkurt, and C. Bilgin, "Detection of abnormal respiratory events with single channel ECG and hybrid machine learning model in patients with obstructive sleep apnea," *IRBM*, vol. 41, no. 5, pp. 241–251, Oct. 2020.
[16] A. Sheta, H. Turabieh, T. Thaher, J. Too, M. Mafarja, M. S. Hossain, and S. R. Surani, "Diagnosis of obstructive sleep apnea from ECG signals using machine learning and deep learning classifiers," *Appl. Sci.*, vol. 11, no. 14, p. 6622, Jul. 2021.
[17] D. Dey, S. Chaudhuri, and S. Munshi, "Obstructive sleep apnoea detection using convolutional neural network based deep learning framework," *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 95–100, Feb. 2018.
[18] U. Erdenebayar, Y. J. Kim, J.-U. Park, E. Y. Joo, and K.-J. Lee, "Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram," *Comput. Methods Programs Biomed.*, vol. 180, Oct. 2019, Art. no. 105001.
[19] D. Padovano, A. Martinez-Rodrigo, J. M. Pastor, J. J. Rieta, and R. Alcaraz, "An experimental review on obstructive sleep apnea detection based on heart rate variability and machine learning techniques," in *Proc. Int. Conf. e-Health Bioeng. (EHB)*, Oct. 2020, pp. 1–4.
[20] B. Nazli and V. H. Altural, "Evaluation of different machine learning algorithms for classification of sleep apnea," in *Proc. 29th Signal Process. Commun. Appl. Conf. (SIU)*, Jun. 2021, pp. 1–4.
[21] A. Srinivasulu, S. Mohan, T. Harika, P. Srujana, and Y. Revathi, "Apnea event detection using machine learning technique for the clinical diagnosis of sleep apnea syndrome," in *Proc. 3rd Int. Conf. Signal Process. Commun. (ICPSC)*, May 2021, pp. 490–493.
[22] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[23] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. Machine Learning*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.
[24] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, and J. H. Peter, "The apnea-ECG database," in *Proc. Comput. Cardiol.*, vol. 27, Sep. 2000, pp. 255–258.
[25] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May 2001.
[26] M. K. Uçar, M. R. Bozkurt, C. Bilgin, and K. Polat, "Automatic detection of respiratory arrests in OSA patients using PPG and machine learning techniques," *Neural Comput. Appl.*, vol. 28, no. 10, pp. 2931–2945, Oct. 2017.
[27] M. Hafezi, N. Montazeri, S. Saha, K. Zhu, B. Gavrilovic, A. Yadollahi, and B. Taati, "Sleep apnea severity estimation from tracheal movements using a deep learning model," *IEEE Access*, vol. 8, pp. 22641–22649, 2020.
[28] F. Hajipour, M. J. Jozani, and Z. Moussavi, "A comparison of regularized logistic regression and random forest machine learning models for daytime diagnosis of obstructive sleep apnea," *Med. Biol. Eng. Comput.*, vol. 58, no. 10, pp. 2517–2529, Oct. 2020.
[29] M. Schätz, A. Procházka, J. Kuchyňka, and O. Vyšata, "Sleep apnea detection with polysomnography and depth sensors," *Sensors*, vol. 20, no. 5, p. 1360, Mar. 2020.
[30] Z. Keshavarz, R. Rezaee, M. Nasiri, and O. Pournik, "Obstructive sleep apnea: A prediction model using supervised machine learning method," in *Studies in Health Technology and Informatics*, vol. 272. Bethesda, MD, USA: PubMed, 2020, pp. 387–390. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32604683/
[31] J. F. Rodrigues, J.-L. Pepin, L. Goeuriot, and S. Amer-Yahia, "An extensive investigation of machine learning techniques for sleep apnea screening," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2709–2716.
[32] Y. J. Kim, J. S. Jeon, S.-E. Cho, K. G. Kim, and S.-G. Kang, "Prediction models for obstructive sleep apnea in Korean adults using machine learning techniques," *Diagnostics*, vol. 11, no. 4, p. 612, Mar. 2021.
[33] W. S. Almuhammadi, K. A. Aboalayon, and M. Faezipour, "Efficient obstructive sleep apnea classification based on EEG signals," in *Proc. Long Island Syst., Appl. Technol.*, May 2015, pp. 1–6.
[34] V. Vimala, K. Ramar, and M. Ettappan, "An intelligent sleep apnea classification system based on eeg signals," *J. Med. Syst.*, vol. 43, no. 2, p. 36, Feb. 2019.
[35] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.
[36] H. Korkalainen, J. Aakko, S. Nikkonen, S. Kainulainen, A. Leino, B. Duce, I. O. Afara, S. Myllymaa, J. Töyräs, and T. Leppänen, "Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 7, pp. 2073–2081, Dec. 2019.
[37] I. Hussain, M. A. Hossain, R. Jany, M. A. Bari, M. Uddin, A. R. M. Kamal, Y. Ku, and J.-S. Kim, "Quantitative evaluation of EEG-biomarkers for prediction of sleep stages," *Sensors*, vol. 22, no. 8, p. 3079, Apr. 2022.
[38] D. Alvarez-Estevez and R. M. Rijsman, "Inter-database validation of a deep learning approach for automatic sleep scoring," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256111.
[39] W. Mumtaz, S. S. A. Ali, M. A. M. Yasin, and A. S. Malik, "A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD)," *Med. Biol. Eng. Comput.*, vol. 56, pp. 233–246, Feb. 2018.
[40] 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.
[41] X. Zang, B. Li, L. Zhao, D. Yan, and L. Yang, "End-to-end depression recognition based on a one-dimensional convolution neural network model using two-lead ECG signal," *J. Med. Biol. Eng.*, vol. 42, no. 2, pp. 1–9, 2022.
[42] E. W. McGinnis, S. P. Anderau, J. Hruschak, R. D. Gurchiek, N. L. Lopez-Duran, K. Fitzgerald, K. L. Rosenblum, M. Muzik, and R. S. McGinnis, "Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2294–2301, Nov. 2019.
[43] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Proc. Comput. Sci.*, vol. 167, pp. 1258–1267, Jan. 2020.
[44] M. J. Patel, A. Khalaf, and H. J. Aizenstein, "Studying depression using imaging and machine learning methods," *NeuroImage, Clin.*, vol. 10, pp. 115–123, Jan. 2016.
[45] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: Towards a sleep data commons," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018.
[46] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bio. Comput. Biology*, vol. 3, no. 2, pp. 185–205, 2005.
[47] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, Jan. 1997.

MOSTAFA M. MOUSSA received the B.S. degree in electrical engineering and the M.S. degree in biomedical engineering from the American University of Sharjah, United Arab Emirates, in 2018 and 2020, respectively. He has worked as a Research Assistant for the duration of his master's and the following two years with AUS then began working as a Research Associate with the Biomedical Engineering Department, Khalifa University. His current research interests include human augmentation and rehabilitation, biomimicry, human–computer interaction, cognition, neuroscience, medical devices and robotics, computer-aided diagnosis (CAD), and deep learning in medical applications.

YAHYA ALZAABI, photograph and biography not available at the time of publication.

AHSAN H. KHANDOKER (Senior Member, IEEE) is currently a Theme Leader of the Healthcare Engineering Innovation Center (HEIC), Khalifa University. He has multidisciplinary research accomplishments in bio-signal processing, bioinstrumentation, nonlinear modeling, and artificial intelligence techniques applied to the area of sleep apnea, autonomic dysfunctions in cardiovascular diseases, diabetic autonomic neuropathy, fetal cardiology, and psychiatry. A number of ideas proposed in his work have influenced the efforts of the biosignal processing platforms developed by companies such as ResMed Sydney, Compumedics Melbourne Australia, Atom Medical Co Tokyo, and the startup company MARP Abu Dhabi (Twinkle Heart Fetal Phonogram device). He extended his research collaborative network to Tohoku University Medical School, Japan, the University of Applied Sciences Jena, Germany, the University of Rochester Medical Center, USA. His research projects are funded by the Abu Dhabi Education Council, the Bill and Melinda Gates Foundation, the Australian Research Council, the Abu Dhabi Education Council, and the Khalifa University Internal Funds in cardiac and mental health monitoring research area in collaboration with Cleveland Clinic Abu Dhabi and several key international medical research facilities in Australia, Germany, and Japan. He has developed his teaching philosophy through teaching a variety of engineering courses (from first year to final year undergraduate courses and graduate level courses) in various parts of the world (Bangladesh, Malaysia, Japan, Australia, and United Arab Emirates) over the last 24 years (since 1996). He particularly emphasizes collaborative and student-centered learning styles which stimulates analytical and critical thinking, and thus enhances problem-solving abilities. He designed the classes in such a way so that it can empower and motivate students to continue lifelong learning.