IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 January 2024, accepted 1 March 2024, date of publication 7 March 2024, date of current version 13 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374408

RESEARCH ARTICLE

Applying Machine Learning Algorithms for the Classification of Sleep Disorders

TALAL SARHEED ALSHAMMARI
Department of Information and Computer Science, College of Computer Science and Engineering, University of Ha'il, Hail 81481, Saudi Arabia
e-mail: talal.alshammari@uoh.edu.sa

ABSTRACT Sleep disorder classification is crucial in improving human quality of life. Sleep disorders and apnoea can have a significant influence on human health. Sleep-stage classification by experts in the field is an arduous task and is prone to human error. The development of accurate machine learning algorithms (MLAs) for sleep disorder classification requires analysing, monitoring and diagnosing sleep disorders. This paper compares deep learning algorithms and conventional MLAs to classify sleep disorders. This study proposes an optimised method for the Classification of Sleep Disorders and uses the Sleep Health and Lifestyle Dataset publicly available online to evaluate the proposed model. The optimisations were conducted using a genetic algorithm to tune the parameters of different machine learning algorithms. An evaluation and comparison of the proposed algorithm against state-of-the-art machine learning algorithms to classify sleep disorders. The dataset includes 400 rows and 13 columns with various features representing sleep and daily activities. The k-nearest neighbours, support vector machine, decision tree, random forest and artificial neural network (ANN) deep learning algorithms were assessed. The experimental results reveal significant performance differences between the evaluated algorithms. The proposed algorithms obtained a classification accuracy of 83.19%, 92.04%, 88.50%, 91.15% and 92.92%, respectively. The ANN achieved the highest classification accuracy of 92.92%, and its precision, recall and F1-score values on the testing data were 92.01%, 93.80% and 91.93%, respectively. The ANN algorithm that achieved high accuracy than other tested algorithms.

INDEX TERMS Machine learning algorithms, deep learning, classification, sleep disorder, genetic algorithm.

I. INTRODUCTION

Sleep is a vital physiological function necessary for physical and mental health. Sleep helps strengthen the body and consolidate the brain and memories. Sleep quality affects cognitive functions, particularly in children and older drivers at increased risk of accidents. Sleep deprivation can affect the human body and cause health problems like heart disease, diabetes and obesity. Physicians, doctors, medical professionals and experts must manually evaluate polysomnography (PSG) records, which can lead to different assessments of sleep stages. Manual classification is prone to human error and is time-consuming for sleep-stage classification [1], [2].

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

Philips conducts an annual World Sleep Day survey on sleep-related attitudes and behaviours. In 2021, the survey polled more than 13,000 adults in 13 countries. Only 55% of adults were satisfied with their sleep, and the rest were dissatisfied with their sleep quality. They suffered from sleep quality because of such factors as the coronavirus disease 2019 (COVID-19) pandemic, sleep apnoea and insomnia. The statistics revealed that 37% said the pandemic negatively influenced their ability to sleep well. Moreover, 37% of participants reported suffering insomnia, while 29% snore, 22% have a shift-work sleep disorder, and 12% experience sleep apnoea [1], [2]. Medical professionals and sleep experts evaluate the quality of sleep by analysing the sleep system classified for various sleep stages. There are five stages of sleep: wakefulness, N1, N2, N3, and rapid-eye

movement (REM). Wakefulness is the stage of alertness when individuals are aware of their surroundings. Brain waves are fast and irregular during consciousness. In N1, the lightest stage of sleep, brain waves are slow, and the muscles relax. In addition, N2 is a deep stage of sleep, and N3 is the deepest stage of sleep, where it is difficult to wake a person during this sleep stage. In REM, the eyes move rapidly back and forth, and brain waves are similar to those exhibited during wakefulness. Every stage of sleep is crucial for different functions. The brain and body stay remarkably active during sleep. Thus, doctors can use PSG to observe the activity state of the brain and body to record electroencephalogram (EEG) and electrocardiogram (ECG) signals [3], [4], [5]. Several researchers have developed techniques to decrease human intervention, involving classification and prediction algorithms to predict patterns or the following actions to automate frequent tasks.

These techniques can be split into conventional (traditional) machine learning algorithms (MLAs) and deep learning algorithms. Traditional MLAs can be employed on a relatively small training dataset, with faster implementation and relative simplicity. The feature engineering process is manual and extracts features of the signals for classification of sleep stages, such as signal entropy and energy. Deep learning algorithms have been introduced as biologically inspired MLAs that attempt to mimic the human brain using neural networks that learns complicated patterns from the data. Deep learning algorithms are a likely replacement for traditional machine intelligence. Deep learning refers to any algorithm that employs layers for data processing, and the process of feature engineering is automatic [6], [7]. Deep learning models are particularly suitable for classification tasks that involve considerable data or complex features. The most common technique for sleep-stage classification is applying an EEG as input [8].

In this study, the authors review research in the field of sleep disorders, focusing on such challenges as data collection, which includes data that are often noisy and uncertain (e.g. missing data) from various hospitals from patients during sleep. The dataset has many limitations due to the data being collected from only one sleep clinic. It is challenging to generalise evaluated results due to the bias of the data towards certain groups of patients, and the biased data can lead to inaccurate results that can influence decision-making. However, there is a lack of natural sleep-stage datasets [9]. Moreover, feature extraction from the dataset is required to train models and select discriminative features, which usually requires more computational effort to select well-suited MLAs from different classifiers [10]. This study is motivated toward the requirement to handle the challenges caused through sleep disorders in the modern lifestyle, especially for people suffering from sleep disorders. Sleep disorders-related diseases are a crucial concern, with the influence of modern lifestyles and people's neglect of this critical need, the dangers associated with the increase in sleep disorders become even more crucial. Sleep is one of the factors most essential to human life. The implement of machine learning techniques to classify sleep disorder is crucial to ensure human's well-being and quality of life.

The MLAs have been implemented for sleep disorder classification, but to the authors? knowledge, there is a lack of comprehensive evaluations of such MLAs in this field. This article makes a two-fold contribution: 1) an overview of the existing studies and research on sleep disorder classification and 2) a comprehensive evaluation of traditional MLAs with deep learning algorithms and evaluation of the performance of the proposed algorithm compared to state-of-the-art machine learning algorithms with default parameters for classification in the context of sleep disorders. The paper is organised as follows. Section II reviews the related work, and Section III provides the evaluation methodology and details the state-of-the-art MLAs reviewed in this paper. Next, Section IV discusses the methods and their performance in sleep disorder classification and shows the results. Finally, Section V discusses the planned future work for this application and concludes the paper.

II. RELATED WORK

The authors in [11] reviewed several studies using consumer sleep technology (CST) with MLAs for sleep classification. They noted that PSG is an essential standard; however, it is expensive and much harder to adjust manual processes requiring specialist controller settings to classify sleep stages. Although CST has been used to track sleep, PSG is more accurate than CST in classifying sleep stages. The article reviewed 27 papers using diverse MLAs, comprising logistic regression (LR), decision tree (DT), support vector machine (SVM), and deep learning. The models may significantly improve the accuracy for classification sleep-stage utilizing CST. However, there are limited ways to apply raw signals with deep learning algorithms.

Another article [12] reviewed 48 papers and discussed the significance of sleep apnoea and its challenges. In addition, MLAs, such as SVM, random forest (RF) and deep learning algorithms, can be used for detecting sleep apnoea from ECG signals. However, they noted some challenges of applying MLAs for sleep apnoea classification: the difference in ECG signals and the availability limitations of datasets for training the models. In their study, the SVM and deep learning-based neural networks performed best in detecting sleep apnoea from ECG signals.

The authors in [13] used MLAs to classify the sleep stage using an EEG spectrogram. The classification of sleep stages takes more time. It is prone to error and uses MLAs with EEG signals for classification. Moreover, the accuracy is low because the data are unbalanced. They used four public datasets to evaluate their models. The results revealed that the proposed algorithms obtained a classification accuracy of 94.17%, 86.82%, 83.02% and 85.12% for the four datasets. They used deep learning algorithms to classify sleep stages and designed a deep learning model. Convolutional neural networks (CNNs) were

applied to extract frequency features and time from the EEG spectrogram. The model contains multiple hidden layers of bidirectional long short-term memory (LSTM) to recognise prediction sequences, a significant method for classifying sleep stages from EEG spectrograms.

Researchers in [14] used MLAs to predict the severity of obstructive sleep apnoea (OSA) syndrome using actual data collected from 4,014 patients, which are not publicly available. The authors conducted supervised and unsupervised learning techniques, such as gradient boosting, RF and K-means. Their proposed methods obtained a good classification accuracy of 88%, 88% and 91%. However, their study has several limitations. Data were collected from a single centre that may be biased, and the data have some missing values. They developed an MLA model that can be used effectively to predict OSA severity that is not time-consuming and is effortless.

The authors in [15] used MLAs, CNNs, LSTM, bidirectional LSTM and gated recurrent units to detect sleep apnoea from a single-lead ECG. The apnoea-ECG dataset was used to validate the proposed algorithms that contain the total number of records, which is 70. Their proposed hybrid models obtained a classification accuracy of 80.67%, 75.04%, 84.13% and 74.72%. The CNN achieved a higher accuracy than the other algorithms, and the results revealed that the best-performing algorithm was the hybrid CNN and LSTM network. They analysed the performance of several deep learning algorithms and pointed out that deep learning algorithms can learn automatic sleep apnoea detection, which differs from conventional MLAs.

Another study [16] proposed a system that used DT, k-nearest neighbours (KNN) and RF algorithms to classify sleep stages from the ECG. They used the publicly available ISRUC?Sleep dataset collected from adults, which had two states: healthy and sleep disorders. Every recording was randomly chosen from the PSG via the Hospital of Coimbra University Sleep Medicine Centre. They used statistical features to analyse the sleep attributes. The DT, KNN and RF algorithms performed the best in automated sleep stages. The RF algorithm obtained a classification accuracy above 90%, which is better than the DT and KNN algorithms.

In addition, researchers [16] proposed a model that used conventional MLAs, such as DT, KNN, RF and deep algorithms, for sleep apnoea detection using a single-lead ECG. The authors used the PhysioNet ECG Sleep Apnoea v1.0.0 dataset that contains 70 records and applied hybrid convolutional-recurrent CNNs to extract features and deep recurrent neural networks (DRRNs) to capture the time pattern of the data. They applied principal component analysis to reduce the dimensions. The accuracy detection of the hybrid CNN-DRNN architecture was better than that of the other algorithms. They recommended hybrid deep neural networks for sleep apnoea detection from the ECG.

In addition, researchers [17] proposed a model that used conventional MLAs, such as DT, KNN, RF and deep algorithms, for sleep apnoea detection using a single-lead ECG. The authors used the PhysioNet ECG Sleep Apnoea v1.0.0 dataset that contains 70 records and applied hybrid convolutional-recurrent CNNs to extract features and deep recurrent neural networks (DRRNs) to capture the time pattern of the data. They applied principal component analysis to reduce the dimensions. The accuracy detection of the hybrid CNN-DRNN architecture was better than that of the other algorithms. They recommended hybrid deep neural networks for sleep apnoea detection from the ECG.

The authors [18] used several MLAs, extreme gradient boosting (XGB), light gradient boosting machine (LGBM), CB, RF, KNN, LR and SVM, for the early detection of individuals with high pretest OSA to recognise whether they have OSA or non-OSA. They used the Wisconsin Sleep Cohort database to evaluate the proposed algorithms, and the clinical data include 1,479 records. These features comprise blood reports, physical measurements, and others. Bayesian optimisation and genetic algorithms have been implemented to tune model hyperparameters, and they suggested that regularly collected clinical parameters can be used to address the limitations. The SVM algorithm obtained a high accuracy of 68.06%, a sensitivity of 88.76%, a specificity of 40.74%, and an F1-score of 75.96%.

Other researchers [19] proposed a sleep-staging model using conventional machine learning and a deep learning approach to automate sleep-stage classification using multi-modal signals. In their study, the RF, KNN, SVM and deep learning algorithms combined CNN and LSTM algorithms and implemented the CNN to extract special features from EEG signals, using LSTM to model the temporal dynamics of the signals. The proposed system was evaluated using public databases (sleep-edf). The CNN combined with LSTM achieved an accuracy of 87.4%, which was better than the other algorithms. Moreover, the data recordings from the patients had noise. However, they used the Butterworth filter to clean the data.

Another paper [20] proposed a system using a deep learning model to automatically classify sleep stages using raw PSG signals. The model extracts features from a one-dimensional CNN. To evaluate the proposed model, they used databases (sleep-edf and sleep-edfx) that are publicly available online. The proposed model obtained high accuracy for two to six sleep classes at 98.06%, 94.64%, 92.36%, 91.22% and 91.00%. The authors suggested that deep learning is a promising approach for automated sleep-stage classification that can replace the job of classical methods to avoid manual experts. This approach is prone to human error. A summary of the algorithm, dataset and accuracy in some of the reviewed studies is presented in **Table 1**.

The authors [21] have developed an efficient method that integrated a heterogeneous feature representation and a genetic algorithm-based ensemble learning model to predict antitubercular peptides to help in the search for a new treatment to strive tuberculosis. Two independent anti-tubercular

TABLE 1. A summary of the algorithm, dataset and accuracy in some of the reviewed studies is presented.

Ref.	Year	Algorithm used	Accuracy	Datasets	Available	Real
[13]	2022	CNN	94.17%, 86.82%, 83.02%, 85.12%	Sleep-EDFX-8, Sleep-EDFX-20, Sleep-EDFX-78, and SHHS	Yes	Yes.
[14]	2023	gradient-boost and RFand KNN.	88%, 88%, 91%.	Medical Centre	No	Yes.
[15]	2021	CNN, LSTM, Bidirectional LSTM and Gated recurrent unit (GRU) .	80.67%, 75.04%, 84.13%, 74.72%.	PhysioNet Apnoea-ECG Database	Yes	Yes.
[16]	2021	DT, KNN, RF.	89.10%, 89.10%, 94.46%.	ISRUC%-Sleep database .	Yes	Yes.
[17]	2022	CNN, LSTM, MLP.	the highest is hybrid deep models 88.13%.	The-PhysioNet ECG Sleep Apnoea v1.0.0 dataset.	Yes	Yes.
[18]	2021	XGB, LGBM, CB, RF, KNN, LR and SVM.	the highest is SVM 68.06%.	the The Wisconsin Sleep Cohort dataset.	Yes	Yes.
[19]	2023	CNN+LSTM, RF, KNN and SVM.	87.4%, 74.07%, 83.65%, 76.04% .	the The Wisconsin Sleep Cohort dataset.	Yes	Yes.
[20]	2019	CNN.	98.06%	sleep-edf and sleep-edfx.	Yes	Yes.

peptides (AtbPs) datasets were used to evaluate the proposed algorithm. Their proposed ''iAtbP-Hyb-EnC'' algorithm obtained a prediction accuracy of 94.47% and 92.68%, respectively, better than other algorithms.

III. METHODOLOGY

A. MATERIALS AND METHODS

This section focuses on implementing deep learning algorithms and conventional MLAs to classify sleep disorders. The following sections describe the datasets to assess the proposed algorithms, the performance metrics to evaluate the models, and the feature importance technique to calculate a score for inputted features. In addition, the classification algorithm used in this research is briefly explained.

B. REAL SLEEP HEALTH AND LIFESTYLE DATASET

The dataset used in this study is the Sleep Health and Lifestyle Dataset downloaded from the Kaggle website [22]. The original dataset includes 400 observations and 13 columns of various data types. Each observation represents the actual sleep state. These data can be categorised into 13 variables relevant to sleep and daily habits, such as gender, age, occupation, sleep duration and sleep quality. Column 13 presents the sleep disorder for each person. This dataset groups the data into three sleep disorder categories, none, sleep apnoea and insomnia, pre-processing step was performed to replace the labels namely: None, Sleep Apnoea and Insomnia into 1, 2 and 3. **Table 2** presents an example of the dataset.

TABLE 2. Detailed information about the sleep health and lifestyle database records in this study.

ID	Gen	Age	Occu	Sle Dur	Q of Sle	Phys Act	Str Lev	BMI Cat	Blood Pr	HR	DS	Sleep Dis-or-der
1	M	27	SW	6.1	6	42	6	Overw	126/83	77	4200	None
2	M	28	DR	6.2	6	60	8	Normal	125/80	75	10000	None
3	M	28	DR	6.2	6	60	8	Normal	125/80	75	10000	None
4	M	28	Sal	5.9	4	30	8	Obese	140/90	85	3000	Apnoea
5	M	28	Sal	5.9	4	30	8	Obese	140/90	85	3000	Apnoea
6	M	28	SW	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
7	M	29	Teac	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
8	M	29	DRr	7.8	7	75	6	Normal	120/80	82	8000	None

C. EXPERIMENT DESIGN

This section proposes an assessment algorithm for the classification of sleep disorders. The methodology comprises of two approaches. The first approach, the model learns from data that are not scored on model performance, and the remaining 30% of the dataset is the testing set. Model performance is evaluated on unseen testing sets, where the model learns from the data without tuning and optimising the parameters. The diagram of the machine learning model used to classify sleep disorders is illustrated in **Figure 1**. The motive behind this step is to evaluate the performance of machine learning algorithms without feature selection and optimisation to comprehend all of the aspects and determine their weaknesses and limitations. In the second approach. The dataset was prepared, and each record was input into the models. The dataset was divided into 70% training data. The proposed models were trained and tested using the optimisation method. The Genetic Algorithm (GA+MLAs) approach was implemented and define a fitness function which, combines the GA and MLAs. GA was used to find an optimal set of parameters, apply feature selection to the training and testing sets. The proposed models learnt from the data using the selected features or parameters in a training phase and performed classification using the trained GA with MLA, evaluating the classification performance of the

models. GA was used for feature selection, to rectify the classifiers optimisation shortcomings. The classifiers have several parameters that require to be tuned and optimised. GA was applied for tuning the best values of model parameters to achieve the best performance of the proposed model. **Figure 2** shows an overview of the implementation of a genetic algorithm. The proposed algorithm runs as follows: Step1: The initial population is randomly generated. Step2: Evaluate a fitness value that evaluates the performance of a candidate solution (a set of parameters). Step3: Select parents for reproduction that have individuals with higher fitness. Step4: Conduct crossover that combines two parents to create new individuals (offspring). Step5: Perform mutation to make random changes to the genetic material. Step 6: Repeat Step 2-5 until the stop criteria are met. with MLAs to conduct feature selection and classification of sleep disorders.

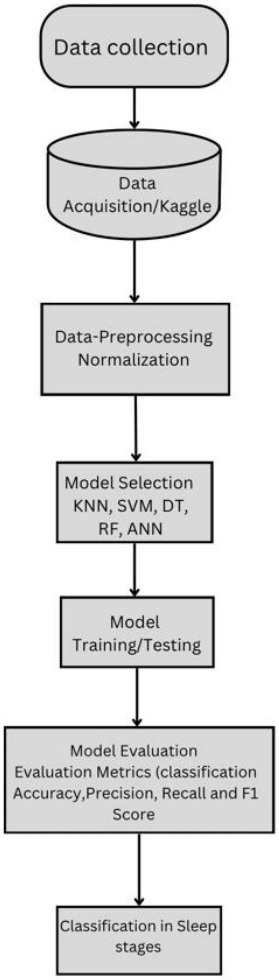


FIGURE 1. Diagram of the machine learning model to classify sleep disorders.

D. PERFORMANCE METRICS

This study evaluates and validates the performance of the proposed model subject to the classification of sleep disorders. In addition, the ratio of the performed activities

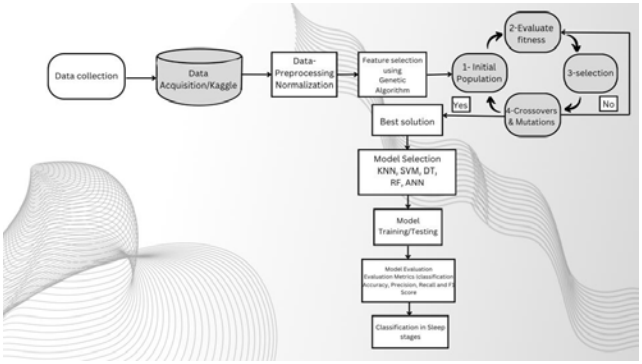


FIGURE 2. The proposed optimised model for sleep disorder classification.

for every person is commonly not identical. For example, sleep apnoea may account for much of the total activity space. The classification accuracy metric is inappropriate for this kind of dataset with unbalanced labels, and the majority class can obtain a higher accuracy [23]. For example, the accuracy metric is suitable when the label class is well-balanced; however, it is not helpful with unbalanced classes. Therefore, this research used four evaluation metrics: classification accuracy, precision, recall and the F1-score [24]. The mathematical expressions for these statistical indices are defined in the equations below. Accuracy was used as a metric to evaluate the classification algorithms, that is, the ratio of correct predictions to the total number of predictions, as presented in (1), where TP denotes a true positive, TN indicates a true negative, FP represents a false positive, and FN denotes a false negative:

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

Precision is the ratio of the number of predicted TPs to the total number of predicted positives (2):

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

Recall is the ratio of the number of predicted TPs to the total number of actual TPs (3):

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

The F1-score provides a weighted average for the precision and recall of a number. A perfect F1-score provides low FPs and low FNs (4):

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

E. CLASSIFICATION ALGORITHMS

1) SUPPORT VECTOR MACHINE

An SVM is a supervised learning algorithm that can be used for classification or regression [25]. The SVM aims to make the best line, called a decision boundary, and is based on the class of hyperplanes that is the line with the highest

margin between two classes. The margin is the vertical distance between the decision boundary and closest data points. Moreover, the SVM is efficient when the number of samples is less than the number of dimensions in the dataset. In addition, SVMs can be used with various kernel functions, such as the RBF and linear functions, allowing the model to learn complicated decision functions [26].

2) K-NEAREST NEIGHBOURS

The KNN is a nonparametric supervised learning algorithm that can be used for classification and regression [25] to classify a data point associated with its closest neighbour. The KNN algorithms are based on feature similarity. The value of k is a process called parameter tuning that refers to the number of nearest neighbour data points to include in the majority voting process. There are various types of distance metrics, such as Euclidean, Manhattan and Minkowski [27].

3) DECISION TREE

A DT is another nonparametric supervised learning algorithm that can be used for classification and regression problems [25]. The DT algorithms are simple and easy to understand and interpret. Thus, the DT model learns simple state rules inferred from the labelled data. The DT can be employed with categorical and numerical data. Moreover, the model achieves good performance, even with noisy data. However, the DT has some disadvantages. For example, the DT cannot handle missing values and may be unstable due to slight variations in the dataset that lead to generating complex trees that do not generalise well to new data, and it is prone to overfitting [28].

4) RANDOM FOREST

An RF classifier is an ensemble learning algorithm that creates multiple random DTs to combine the predictions, improve model predictive accuracy and manage overfitting [25]. The model can use two random processes: Bootstrapping and random selection of features. Bootstrapping guarantees that the model does not utilize the like data for each tree, so that the model is minimally sensitive to conversions in the training data data. Random feature selection reduces the correlation between the trees and aggregates them [29].

5) ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is a supervised learning algorithm that mimics the human brain. It is a combination of interconnected nodes called artificial neurons. The ANN comprises multiple hidden layers that are between the input and output layers. Every entry contains a neural weight. Each input is fed to each neuron of the first layer, and every layer is completely linked to the next layer and is assigned a weight. A weighted sum is sent via a threshold function to an activation function. The output of the activation function determines whether a neuron is activated, and the activated neuron is passed to the output neuron of the next layer, called feed-forward propagation [30], [31].

F. FEATURE IMPORTANCE

Feature importance is a technique to calculate the score for each input feature passed to the model. The maximum score of features has a significant influence on model accuracy. In this paper, which involves the body mass index (BMI), blood pressure, sleep duration, occupation and age features, feature importance highly influences model accuracy, as depicted in **Figure 3**.

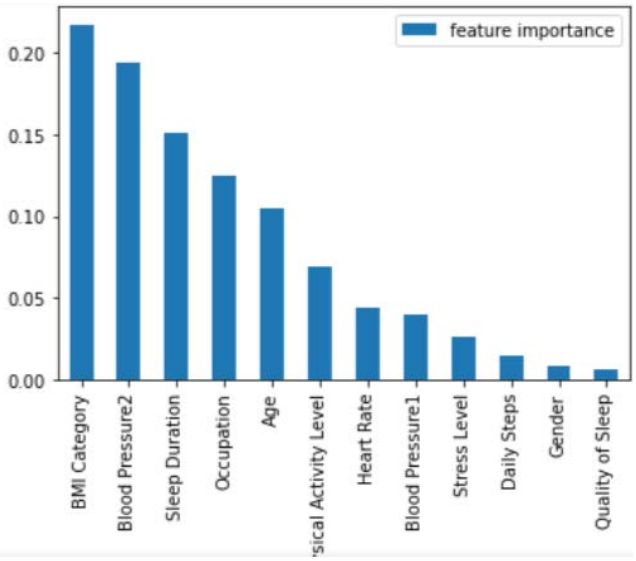


FIGURE 3. Feature importance.

G. CORRELATION COEFFICIENT

The correlation coefficient is a statistical measure which, shows the correlations between variables relevant to sleep and daily habits. It is a number between -1 and 1. The quality of sleep has a higher correlation coefficient with sleep duration than the other variables. The calculated correlation between the features is summarised as illustrated in **Figure 4**.

H. EXPORT GRAPHVIZ IN PYTHON

The export graphviz in Python was used to export a DT in the DOT format, which is a text-based format presented in **Figure 5**.

I. GENETIC ALGORITHM

Genetic algorithm (GA) is a kind of evolutionary algorithm that is optimization algorithms inspired by the process of natural selection and genetics. GA is used to tune the parameters and solve optimization problems for which there are several of candidate solutions. The genetic algorithm follows a several of steps, as shown in **Figure 6**. [32].

IV. RESULTS AND DISCUSSION

This study determined that MLAs can be used to accurately classify sleep disorders. The experiments were conducted

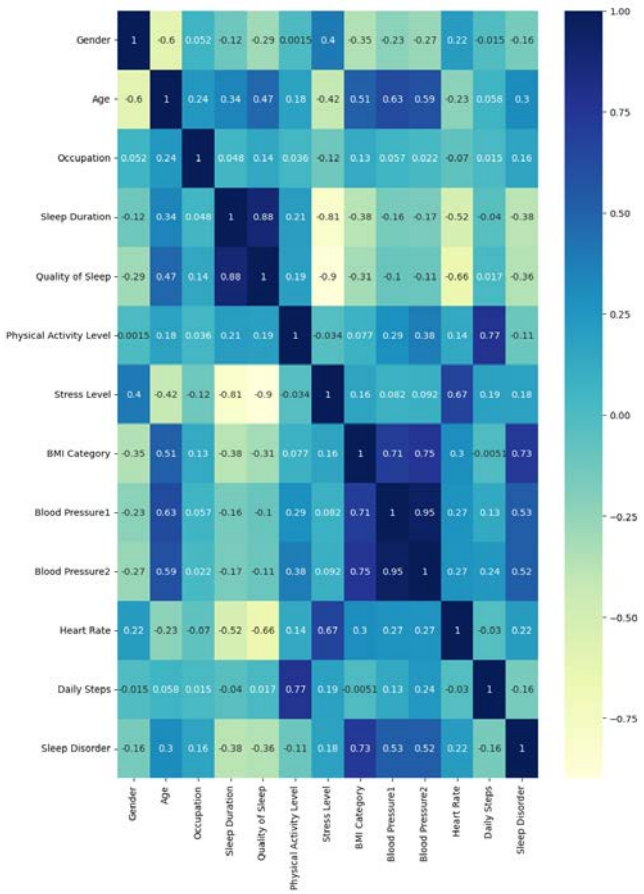


FIGURE 4. Correlation coefficient.

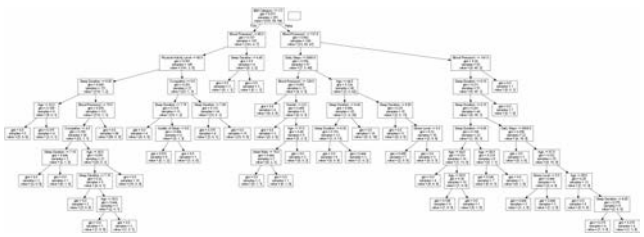


FIGURE 5. Export graphviz flow chart.

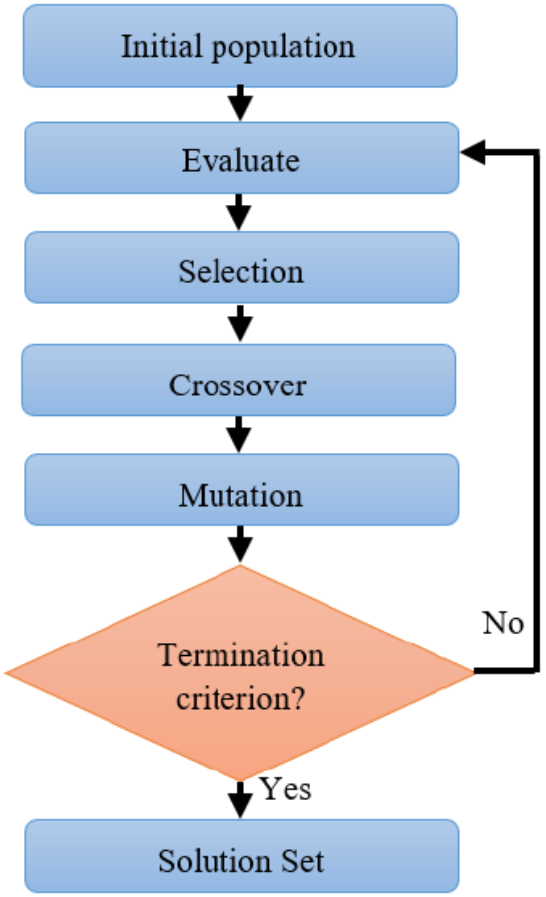


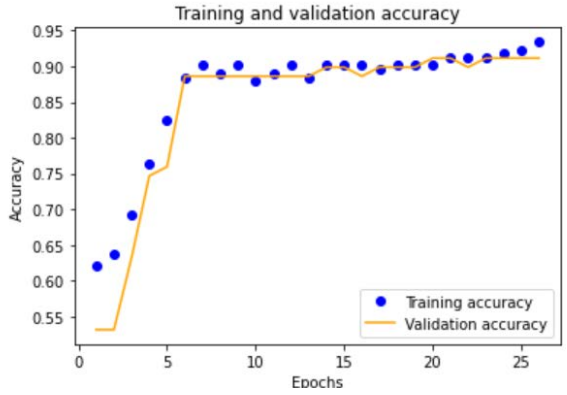
FIGURE 6. Basic architecture of the genetic algorithm [32].

without the use of GA. The KNN, SVM, DT, RF and ANN classifiers achieved an accuracy of 84.96%, 64.6%, 86.73%, 88.5% and 91.15% respectively. The highest-performing algorithms achieved an accuracy of over 90%. This article used the default configuration for each classifier that applied various kernels of the SVM on training data to determine the best-performing kernel. The RBF kernel obtained good performance for the SVM algorithm, whereas the linear and polynomial kernels produced the worst accuracy. However, there is a challenge to find an optimal parameter for each classifier. Due to the lack of an optimisation algorithm appropriate for MLAs in high dimensional datasets. **Figure 7-8** present a performance plot during training and validation. These plots were generated while attempting to classify the experimental values. Despite demonstrating similar loss curves, the points might not all be identical due to model weight changes. However, this training and validation loss provides a good comprehension of the learning performance changes over the number of epochs. This method assists in determining problems for the model during the learning phase to prevent overfitting the model and identify whether adding more training patterns improves the validation score.

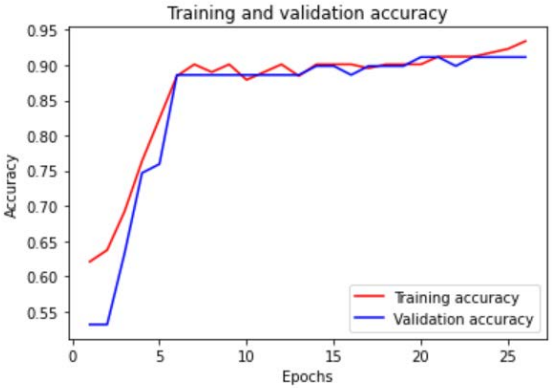
Table 3 shows the results of the performance of all evaluated MLAs by training phase and **Table 4** presents the results were obtained using a 5-fold cross-validation.

The overall performance of all the evaluated MLAs through the datasets in testing phase using accuracy, precision, recall and the F1-score are summarised in **Figure 9** and **Table 5**. The results reveal the competitive performance of the evaluated algorithms. However, the deep learning algorithms based on neural networks obtained the highest accuracy over the other conventional machine learning techniques and achieved a classification accuracy of 91.15%.

Although, some of models have showed good accuracy, due to the insufficiency of an optimisation algorithm appropriate for every classifier in high dimensional datasets. The models have diverse parameters that require to be tuned and optimised to obtain the best possible results. In this



(a) Figure A



(b) Figure B

FIGURE 7. Training and validation accuracy.

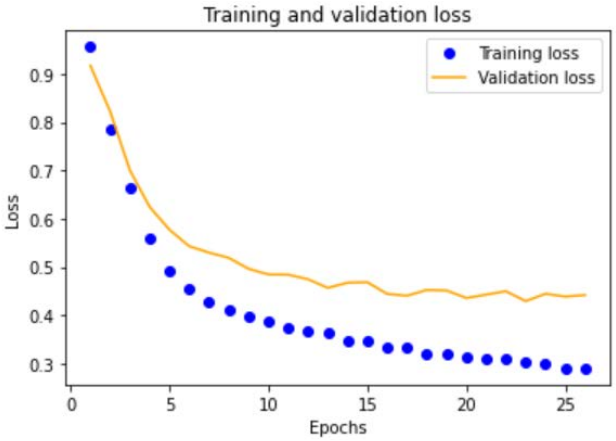


FIGURE 8. Training and validation loss.

experiment, the GA applied to search for an optimal and tune parameters of classifiers that achieved good results as shown in **Figure 10** and **Table 6**.

The results compare the best performance for the MLA models with the best performance for the GA+MLA models. To then show that the best accuracy of the two models has a statistically significant difference, another test is performed using the t-test. The test results of all machine learning algorithms evaluated on the dataset using Precision, Recall,

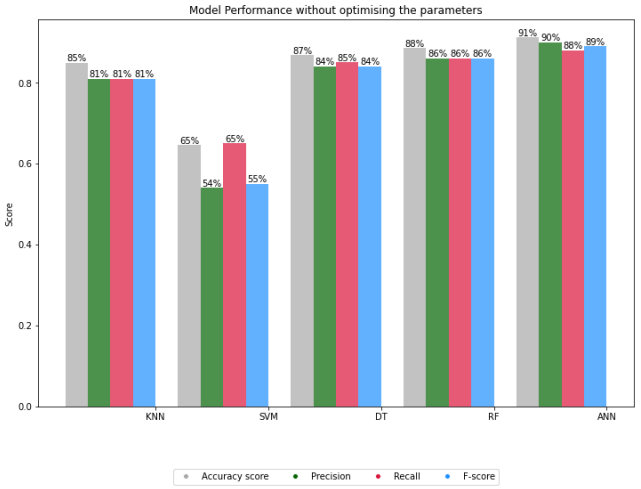


FIGURE 9. Results of the performance of all evaluated MLAs (As default parameters).

TABLE 3. Results of the performance of all evaluated MLAs by training phase. (without optimisation of the parameters.).

Evaluation metrics	KNN	SVM	DT	RF	ANN
Precision	87.22%	54.33%	93.49%	93.49%	91.33%
Recall	87.35%	66.28%	93.48%	93.48%	91.18%
F-score	87.25%	57.46%	93.47%	93.48%	91.23%
Accuracy score	87.35%	66.28%	93.48%	93.48%	91.18%

TABLE 4. Results of the performance of all evaluated MLAs by 5-fold cross-validation. (without optimisation of the parameters.).

Evaluation metrics	KNN	SVM	DT	RF	ANN
Precision	87.22%	54%	93.49%	93.49%	92.25%
Recall	87.35%	65%	93.48%	93.48%	91.58%
F-score	87.25%	55%	93.47%	93.48%	91.55%
Accuracy score	83.94%	64.6%	86.99%	88.14%	91.58%

TABLE 5. Results of the performance of all evaluated MLAs by testing phase (without optimisation of the parameters.).

Evaluation metrics	KNN	SVM	DT	RF	ANN
Precision	81%	54%	84%	86%	90%
Recall	81%	65%	85%	86%	88%
F-score	81%	55%	84%	86%	89%
Accuracy score	84.96%	64.6%	86.73%	88.5%	91.15%

F-Score and the t-test are shown in **Table 6**. Based on the test, we can see that while not all algorithms show

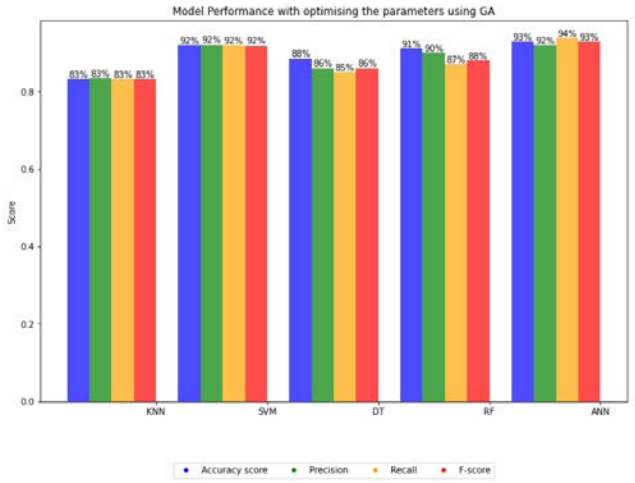


FIGURE 10. Results of the performance of all evaluated MLAs +GA (model performance with optimisation of the parameters using GA).

TABLE 6. Results of the performance of all evaluated MLAs (model performance with optimisation of the parameters using GA).

Evaluation metrics	KNN	SVM	DT	RF	ANN
Precision	83.42%	92.11%	86%	90.00%	92.01%
Recall	83.18%	92.03%	85%	87.00%	93.80%
F-score	83.21%	91.88%	86%	88.00%	91.93%
Accuracy score	83.19%	92.04%	88.50%	91.15%	92.92%

significant differences, the results obtained are the proposed method (GA+MLAs) outperforming MLAs with default parameters. To improve the performance of the classifiers, GA was implemented to determine the optimal values for the MLAs. The GA was performed for 5 generations and different parameter settings were used, as shown in **Table 7**, and the best parameters and solutions were selected based on the fitness score as shown in **Table 8**. For example, the optimal parameters generated by the GA for optimising the KNN model were k = 2 and the Euclidean distance metric. These optimised parameters were used for training and testing the KNN model on the entire dataset. The KNN, SVM, DT, RF and ANN obtained a classification accuracy of 83.19%, 92.04%, 88.50%, 91.15% and 92.92% respectively. This article applied a grid search technique, instead of the GA, which attempts to optimising the hyperparameters of the SVM. This technique can search the hyperparameter space and find the optimal hyperparameter values for the SVM classifier even more effectively and achieve better results in a short period and reduce the training time. Furthermore, a comparison was conducted with the previous study which, used the same SleepHealth and Lifestyle Dataset [33]. The proposed algorithm with optimisation of the parameters using (GA) obtined better results than recent study.

TABLE 7. parameter of the GA settings used.

Parameter	Value
Population size	12
Generations	5
Elite percentage	0.2
Mutation rate	0.8
Crossover rate	0.8

TABLE 8. Best-optimised parameters of models.

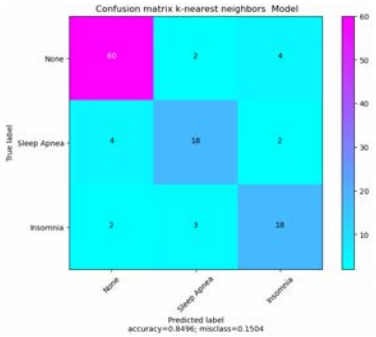
Model	Best- Optimised Parameters
KNN	(k = 2, the Euclidean distance metric)
SVM	GridSearchCV(cv=5, estimator=SVC(), param-grid='C': [0.1, 1, 10], 'gamma': [0.001, 0.01, 0.1], scoring='f1-weighted')
DT	(max-depth=4, min-samples-split=3)
RF	(max-depth=9, min-samples-split=6, n-estimators=33)
ANN	('num-hidden-layers': 1, 'num-units-per-layer': 24, 'learning-rate': 0.004068331104981341)

TABLE 9. The estimating of p values and t-tests.

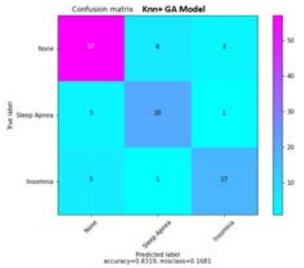
Model		t-test result	Conclusion
KNN	t-stat.	0.3375263702777991	No significant
—	p-value	0.7396243325450431	improvement
SVM	t-stat.	-1.5212776585113266	significant
—	p-value	0.14556309281759117	improvement
DT	t-stat.	0.3629888130588261	No significant
—	p-value	0.7208408332545125	improvement
RF	t-stat.	-1.3416407864997872	significant
—	p-value	0.19639447228341037	improvement
ANN	t-stat.	-1.186884852112364	significant
—	p-value	0.2507018182951046	improvement

A. T-TEST ANALYSIS

To evaluate the statistical significance of the improvement obtained through the GA-optimised MLAs, t-test for samples were performed. The null hypothesis stated that the average accuracy of some of the GA-optimised MLAs models differs significantly from the baseline accuracy. The t-test revealed a significant improvement in accuracy with the GA-optimised MLAs classifiers as shown in **Table 9**.

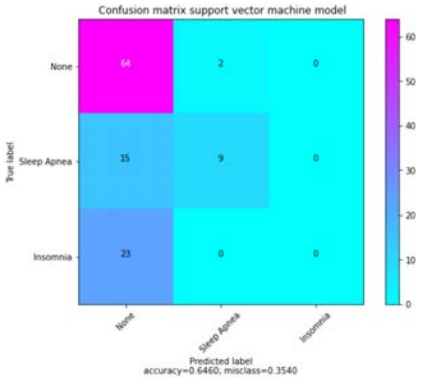


(a) KNN model with default parameters

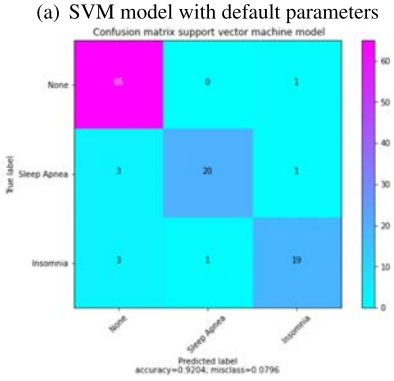


(b) KNN model with optimisation of the parameters

FIGURE 11. Confusion matrix for KNN model.

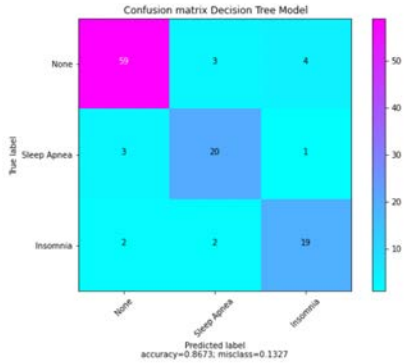


(a) SVM model with default parameters

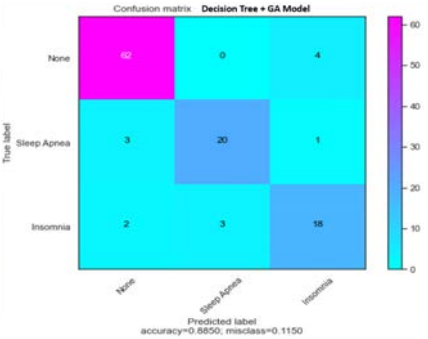


(b) SVM model with optimisation of the parameters

FIGURE 12. Confusion matrix for SVM model.

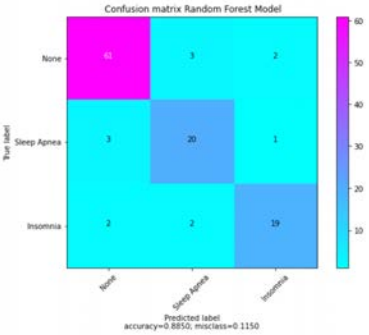


(a) DT model with default parameters

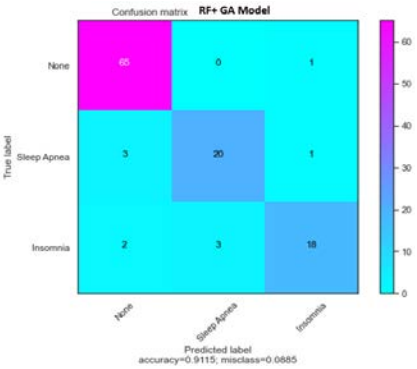


(b) DT model with optimisation of the parameters

FIGURE 13. Confusion matrix for DT model.



(a) RF model with default parameters

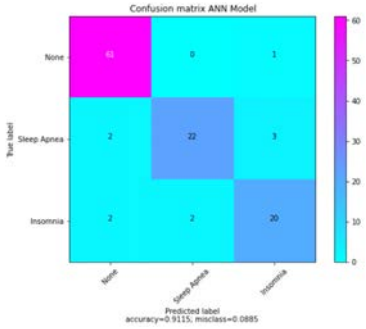


(b) RF model with optimisation of the parameters

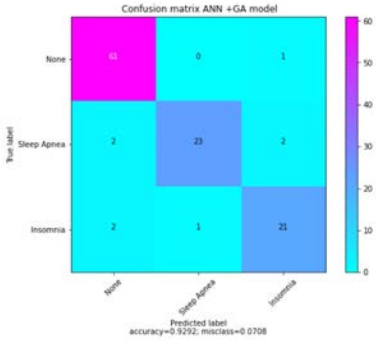
FIGURE 14. Confusion matrix for RF model.

B. CONFUSION MATRIX

The performance of the MLAs classifiers was evaluated using a confusion matrix that summarises the classification results. **Figure 11-15** shows the confusion matrix for the multi-class classification task. The confusion matrix allows insights into the performance of the model by appearance the number of instances that were classified into each class and



(a) ANN model with default parameters



(b) ANN model with optimisation of the parameters

FIGURE 15. **Confusion matrix for ANN model.**

the misclassifications. For example, in ANN+ GA model, in Class 1, 61 instances were correctly classified, while 0 instances were misclassified as Class 2 and 1 instance as Class 3. Similarly, in Class 2, 23 instances were correctly classified, however, 2 instances were misclassified as Class 1 and 3, in Class 3, 21 instances were correctly classified, but 2 instances were misclassified as Class 1 and 1 instance as Class 2. The confusion matrix reveals crucial knowledge for the performance of the RF classifier in the multi-class classification task. Using the matrix. The model obtained high accuracy in Class 1, with 96% correctly classified instances. However, it showed lower accuracy in Class 2 and Class 3, with misclassifications of 20% and 26%, respectively.

V. CONCLUSION

In this paper, an optimised model for sleep disorder classification is proposed that implements MLAs with a genetic algorithm to explore optimal hyperparameter values for each model and obtain good results. This paper analysed the performance of MLAs for sleep disorder classification and evaluated many state-of-the-art MLAs on the real-world Sleep Health and Lifestyle Dataset. In addition, MLAs can learn from high-dimensional sleep data and attempt to classify sleep disorders without depending on expert-defined features. The proposed optimised ANN with GA achieved the highest accuracy over the other MLAs at 92.92%. The precision, recall, and F1-score values on the testing data were 92.01%, 93.80% and 91.93%, respectively. Even with a limitation in the amount of data. This study addressed the challenges in implementing MLAs for classification sleep disordering. However, large datasets are still needed for training and evaluating models in this field. The MLAs with GA can significantly improve the accuracy of sleep disorder classification. Future work will focus on developing MLAs using unsupervised learning in addition to assessing the dataset on a new model and comparing its performance against existing state-of-the-art models.

REFERENCES

[1] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, and A. G. Ravelo-García, "A portable wireless device for cyclic alternating pattern estimation from an EEG monopolar derivation," *Entropy*, vol. 21, no. 12, p. 1203, Dec. 2019.

[2] Y. Li, C. Peng, Y. Zhang, Y. Zhang, and B. Lo, "Adversarial learning for semi-supervised pediatric sleep staging with single-EEG channel," *Methods*, vol. 204, pp. 84–91, Aug. 2022.

[3] E. Alickovic and A. Subasi, "Ensemble SVM method for automatic sleep stage classification," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1258–1265, Jun. 2018.

[4] D. Shrivastava, S. Jung, M. Saadat, R. Sirohi, and K. Crewson, "How to interpret the results of a sleep study," *J. Community Hospital Internal Med. Perspect.*, vol. 4, no. 5, p. 24983, Jan. 2014.

[5] V. Singh, V. K. Asari, and R. Rajasekaran, "A deep neural network for early detection and prediction of chronic kidney disease," *Diagnostics*, vol. 12, no. 1, p. 116, Jan. 2022.

[6] J. Van Der Donckt, J. Van Der Donckt, E. Deprost, N. Vandenbussche, M. Rademaker, G. Vandewiele, and S. Van Hoecke, "Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104429.

[7] H. O. Ilhan, "Sleep stage classification via ensemble and conventional machine learning methods using single channel EEG signals," *Int. J. Intell. Syst. Appl. Eng.*, vol. 4, no. 5, pp. 174–184, Dec. 2017.

[8] Y. Yang, Z. Gao, Y. Li, and H. Wang, "A CNN identified by reinforcement learning-based optimization framework for EEG-based state evaluation," *J. Neural Eng.*, vol. 18, no. 4, Aug. 2021, Art. no. 046059.

[9] 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.

[10] Z. Mousavi, T. Y. Rezaii, S. Sheykhivand, A. Farzamnia, and S. N. Razavi, "Deep convolutional neural network for classification of sleep stages from single-channel EEG signals," *J. Neurosci. Methods*, vol. 324, Aug. 2019, Art. no. 108312.

[11] S. Djanian, A. Bruun, and T. D. Nielsen, "Sleep classification using consumer sleep technologies and AI: A review of the current landscape," *Sleep Med.*, vol. 100, pp. 390–403, Dec. 2022.

[12] N. Salari, A. Hosseinian-Far, M. Mohammadi, H. Ghasemi, H. Khazaie, A. Daneshkhah, and A. Ahmadi, "Detection of sleep apnea using machine learning algorithms based on ECG signals: A comprehensive systematic review," *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 115950.

[13] C. Li, Y. Qi, X. Ding, J. Zhao, T. Sang, and M. Lee, "A deep learning method approach for sleep stage classification with EEG spectrogram," *Int. J. Environ. Res. Public Health*, vol. 19, no. 10, p. 6322, May 2022.

[14] H. Han and J. Oh, "Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity," *Sci. Rep.*, vol. 13, no. 1, p. 6379, Apr. 2023.

[15] M. Bahrami and M. Forouzanfar, "Detection of sleep apnea from single-lead ECG: Comparison of deep learning algorithms," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2021, pp. 1–5.

[16] S. Satapathy, D. Loganathan, H. K. Kondaveeti, and R. Rath, "Performance analysis of machine learning algorithms on automated sleep staging feature sets," *CAAI Trans. Intell. Technol.*, vol. 6, no. 2, pp. 155–174, Jun. 2021.

[17] M. Bahrami and M. Forouzanfar, "Sleep apnea detection from single-lead ECG: A comprehensive analysis of machine learning and deep learning algorithms," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[18] J. Ramesh, N. Keeran, A. Sagahyroon, and F. Aloul, "Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning," *Healthcare*, vol. 9, no. 11, p. 1450, Oct. 2021.
[19] S. K. Satapathy, H. K. Kondaveeti, S. R. Sreeja, H. Madhani, N. Rajput, and D. Swain, "A deep learning approach to automated sleep stages classification using multi-modal signals," *Proc. Comput. Sci.*, vol. 218, pp. 867–876, Jan. 2023.
[20] O. Yildirim, U. Baloglu, and U. Acharya, "A deep learning model for automated sleep stages classification using PSG signals," *Int. J. Environ. Res. Public Health*, vol. 16, no. 4, p. 599, Feb. 2019.
[21] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "IAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104778.
[22] (2023). *Sleep Health and Lifestyle Dataset*. [Online]. Available: http://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
[23] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.
[24] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[25] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[26] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, Jun. 2022, Art. no. 100071.
[27] M. Q. Hatem, "Skin lesion classification system using a K-nearest neighbor algorithm," *Vis. Comput. Ind., Biomed., Art*, vol. 5, no. 1, pp. 1–10, Dec. 2022.
[28] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: An updated survey," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4765–4800, May 2023.
[29] P. Tripathi, M. A. Ansari, T. K. Gandhi, R. Mehrotra, M. B. B. Heyat, F. Akhtar, C. C. Ukwuoma, A. Y. Muaad, Y. M. Kadah, M. A. Al-Antari, and J. P. Li, "Ensemble computational intelligent for insomnia sleep stage detection via the sleep ECG signal," *IEEE Access*, vol. 10, pp. 108710–108721, 2022.
[30] Y. You, X. Zhong, G. Liu, and Z. Yang, "Automatic sleep stage classification: A light and efficient deep neural network model based on time, frequency and fractional Fourier transform domain features," *Artif. Intell. Med.*, vol. 127, May 2022, Art. no. 102279.
[31] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. R. Rao, "Cognitive analysis of working memory load from eeg, by a deep recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2576–2580.
[32] A. Hichri, M. Hajji, M. Mansouri, K. Abodayeh, K. Bouzrara, H. Nounou, and M. Nounou, "Genetic-algorithm-based neural network for fault detection and diagnosis: Application to grid-connected photovoltaic systems," *Sustainability*, vol. 14, no. 17, p. 10518, Aug. 2022.
[33] I. A. Hidayat, "Classification of sleep disorders using random forest on sleep health and lifestyle dataset," *J. Dinda : Data Sci., Inf. Technol., Data Anal.*, vol. 3, no. 2, pp. 71–76, Aug. 2023.

TALAL SARHEED ALSHAMMARI received the B.Sc. degree in computer science from Al Jouf University, Saudi Arabia, in 2007, the M.Sc. degree in computer science from California Lutheran University, Thousand Oaks, CA, USA, in 2012, and the Ph.D. degree from Staffordshire University, U.K., in 2019. He is currently an Assistant Professor with the College of Computer Science and Engineering, University of Ha'il. His main research interests include artificial intelligence, machine learning, and data science.