IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 January 2025, accepted 23 January 2025, date of publication 28 January 2025, date of current version 3 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3535535

RESEARCH ARTICLE

Improving Sleep Disorder Diagnosis Through Optimized Machine Learning Approaches

MD. ATIQUR RAHMAN[1], ISRAT JAHAN[1], MAHEEN ISLAM[1], (Member, IEEE), TASKEED JABID[1], MD SAWKAT ALI[1], MOHAMMAD RIFAT AHMMAD RASHID[1], MOHAMMAD MANZURUL ISLAM[1], MD. HASANUL FERDAUS[1], MD MOSTOFA KAMAL RASEL[1], MAHMUDA RAWNAK JAHAN[1], SHAYLA SHARMIN[2], TANZINA AFROZ RIMI[2], (Graduate Student Member, IEEE), ATIA SANJIDA TALUKDER[2], MD. MAFIUL HASAN MATIN[1], AND M. AMEER ALI[3]

[1]Department of Computer Science and Engineering, East West University, Dhaka 1212, Bangladesh
[2]Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City 1216, Bangladesh
[3]Tesserent, Melbourne VIC 3000, Australia

Corresponding author: Md. Atiqur Rahman (atick_rasel@yahoo.com)

ABSTRACT Classifying sleep disorders, such as obstructive sleep apnea and insomnia, is crucial for improving human quality of life due to their significant impact on health. The traditional expert-based classification of sleep stages, particularly through visual inspection, is challenging and prone to errors. This fact highlights the need for accurate machine learning algorithms (MLAs) for analyzing, monitoring, and diagnosing sleep disorders. This paper compares the MLAs for sleep disorder classification, specifically targeting None, Sleep Apnea, and Insomnia, using the Sleep Health and Lifestyle Dataset. We conducted two experiments. In the first one, we selected five key features from the feature spaces using the Gradient Boosting Regressor based on the Mean Decrease Impurity (MDI) technique. We chose two key features using the same methodology in the second experiment. We utilized 15 machine learning classifiers, and Gradient Boosting, Voting, Catboost, and Stacking Classifiers achieved an identical classification accuracy of 97.33%, with Precision, Recall, F1-score of 0.9733, and Specificity of 0.9569 in the original feature space. Among these, Gradient Boosting had the highest AUC of 0.9953 and was 3.36, 5.86, and 20.16 times faster than Voting, Catboost, and Stacking Classifiers, respectively. In the second experiment, the Decision Tree achieved the highest accuracy of 96% in the original and engineered feature spaces and was 149.33 times faster in the engineered feature space. Thus, this research proposes Gradient Boosting as the most effective method, outperforming all state-of-the-art techniques by achieving the highest accuracy, precision, recall, F1-score, and AUC, highlighting its superior classification performance and computational efficiency.

INDEX TERMS SMOTEENN, ANOVA test, feature engineering, classification, sleep disorder, machine learning classifiers, ensemble technique.

I. INTRODUCTION

Sleep is essential for physical and mental health because it supports bodily healing and optimal brain function. It is crucial for individuals across all age groups, including children and older adults, who are more vulnerable to the effects of sleep deprivation. While the immediate concerns may include an increased risk of accidents and decreased cognitive function, the more serious implications of inadequate sleep are significant. Chronic sleep deprivation can lead to various health issues, such as heart disease, diabetes, and obesity, affecting overall well-being and quality of life [2], [3], [4].

Medical professionals and experts in sleep categorize sleep stages to assess sleep quality, encompassing five stages: wakefulness, N1, N2, N3, and REM. Wakefulness represents the state of alertness, with high-frequency and irregular brain

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

waves indicating consciousness of the surroundings. In N1, the initial light sleep stage, brain waves slow, and muscles relax. N2 is deeper and less easily disturbed, while N3, known as slow-wave sleep, is the deepest stage from which it is challenging to awaken. REM sleep features rapid eye movements and brain waves resembling wakefulness. Each stage plays a crucial role in physiological processes. During sleep, the brain and body remain active, addressing systemic needs. Polysomnography (PSG) monitors brain and body activity [5], [6], [7], [8], [9], [10].

Scientists have investigated ways to minimize human input in sleep disorder diagnosis, using automated systems for routine tasks. These systems employ classification and predictive algorithms to help identify and categorize disorders, enabling prompt treatment. Traditional methods often depend on expert knowledge and manual sleep data analysis, which can be slow and error-prone. Machine learning presents a promising solution using computational analysis to automate complex processes. This research aims to enhance sleep disorder classification, specifically including None, Sleep Apnea, and Insomnia, through machine-learning techniques applied to real-world medical data. The process involves thorough data preprocessing, hypothesis testing, feature engineering, and feature selection. The study offers five specific contributions to improve classification accuracy:

1) This study uses machine learning algorithms for classification and prediction to automate common tasks, which reduces the need for human involvement in identifying sleep disorders.
2) The research applies machine learning methods to actual datasets, resulting in a more accurate classification of sleep disorders.
3) The study underscores the critical role of feature engineering in optimizing machine learning models for sleep disorder classification, demonstrating its ability to significantly enhance model performance compared to using original feature spaces.
4) It utilizes fewer features to create a lightweight model that enhances computational efficiency and accelerates classification tasks.
5) It proposes future research to apply machine learning models to other classification tasks, develop lightweight models, and improve real-time operational capabilities for better healthcare outcomes.

The paper is organized as follows: Section II reviews the related works, and Section III describes the methodology implemented for evaluation. Section IV presents the experimental results and analysis, and Section V provides a comparative study. Finally, the paper concludes with the planned future work for this application in Section VI.

II. RELATED WORK

Polysomnography (PSG) records, manually evaluated by medical professionals, are subject to human error and can be time-consuming for assessing sleep stages. Philips, a leading company in healthcare technology, addresses challenges in sleep assessment and enhances sleep quality by conducting an annual World Sleep Day survey. The 2021 survey covered over 13,000 adults across 13 countries and found that only 55% of adults were satisfied with their sleep quality. They attributed this dissatisfaction to factors like the COVID-19 pandemic, sleep apnea, and insomnia. The survey highlighted the significant impacts of the pandemic on sleep, with 37% of respondents reporting disrupted sleep patterns. Also, 37% of participants experienced insomnia, 29% reported snoring, 22% had a shift-work sleep disorder, and 12% suffered from sleep apnea. Given these challenges, PSG remains crucial for evaluating sleep stages, including wakefulness, N1, N2, N3, and rapid-eye movement (REM), each of which plays a distinct role in brain and body function. Researchers are developing automated algorithms to improve sleep assessment and reduce the need for human intervention in sleep stage classification and prediction tasks [11], [12].

The paper [13] highlights that in machine learning, having many irrelevant or less relevant features can reduce classification accuracy and run-time performance. Researchers often use feature selection methods to eliminate irrelevant features or transform them into fewer, more relevant ones. However, these methods alone may not be sufficient to enhance performance. This study proposes a new hybrid feature projection model to improve classification performance in sleep disorder diagnosis. We utilize the MCMST Clustering algorithm during the data preprocessing stage with various feature projection methods, including PCA, LDA, SVD, t-SNE, NCA, Isomap, and PR. The proposed model achieved a classification accuracy of 0.9627 using Kernel PCA, demonstrating its superiority over pure KNN and KNN with other feature projection methods.

Shao et al. [8] addressed the challenges of manual sleep staging and the limitations of automatic sleep staging in clinical practice. They proposed a hybrid intelligent model combining data intelligence and knowledge intelligence, using EEG and EOG channels with a temporal fully convolutional network and a multi-task feature mapping structure. Their model performed better than the existing ones, scoring 0.804 on the ISRUC dataset and 0.780 on the Sleep-EDFx dataset using the Macro-F1 measure. The study also investigated how smart algorithms can fix big jumps and illogical changes in sleep stage charts, making clinical sleep staging more accurate and efficient.

A 2024 study by Alshammari examined how machine learning can classify sleep disorders. The research compared deep learning methods with machine learning techniques. It presented an improved approach with the Sleep Health and Lifestyle Dataset [1]. The study used genetic algorithms to optimize machine-learning methods and compared their results to known standards. It found that artificial neural networks (ANNs) performed best, correctly classifying sleep disorders at 92.92%, better than all other methods tested [2].

Using the same Dataset, Hidayat [14] developed an optimized Random Forest classifier, evaluating the split quality in each decision tree with the Gini Index and achieving an accuracy of 88%. Another study by Tareq [15] investigated how machine learning can identify sleep problems such as insomnia and sleep apnea. As traditional methods for detecting these disorders were often costly and slow, there was a need for automated systems that used medical data. This research used a Random Forest Classifier to predict sleep disorders, working with data from the Sleep Health and Lifestyle Dataset. This method correctly identified sleep disorders at 88%, performing better than other algorithms tested on the same dataset [1].

Warunlawan et al. [16] examined how sleep habits and lifestyle choices affect life quality using the Sleep Health and Lifestyle Dataset. The author showed that MRMR was the most effective technique for selecting the three key factors: level of physical activity, systolic blood pressure, and body mass index (BMI). Finally, among the 21 classifiers, Bagged Trees was the most accurate and achieved an accuracy of 91.90%.

SwSleepNet, a deep learning sleep stages classifier proposed by Zhu et al. [17], used Sequential CNN for extracting the most significant features. They applied this method to three datasets: clinical data of Huashan Hospital Fudan University (HSFU), Sleep-EDF Expanded, and Montreal Archive of Sleep Studies, and achieved an accuracy of 81.8%, 84.5%, and 86.7%, respectively. In 2022, Zhang et al. [18] created an automated system that used a radio frequency (RF) sensor built into a bed to identify and forecast breathing problems during sleep. They gathered nighttime recordings from 27 Weill Cornell Center for Sleep Medicine patients. The system used near-field coherent sensing (NCS) to record breathing patterns. A random-forest machine learning model analyzed the data, achieving up to 88.6% sensitivity and 89.0% specificity for detecting apneic events and predicting events up to 90 seconds in advance, with sensitivity and specificity rates of 81.3% and 82.1%, respectively.

In 2023, Wadichar et al. proposed an automated system to identify cyclic alternating patterns (CAP) phases and categorize sleep disorders using EEG data from the CAP sleep database [19]. Their hierarchical method achieved 91.45% accuracy in distinguishing between healthy and unhealthy CAP sequences and 90.55% accuracy in classifying sleep disorders, such as PLM, RBD, NFLE, NARCO, and INS. Focusing specifically on phase B of the CAP sequences improved the model's performance, yielding accuracies of 92.79% and 93.31% for healthy-unhealthy and disease classification, respectively.

Ramesh et al. [20] explored various machine learning models, including SVM, Random Forest, KNN, Logistic Regression, Catboost, and Light Gradient Boosting Machine (LGBM). They utilized Bayesian optimization and genetic algorithms to optimize the models' hyperparameters for identifying individuals with or without obstructive sleep apnea (OSA) with data from the Wisconsin Sleep Cohort dataset. Among the classifiers tested, SVM achieved the highest accuracy at 68.06%, with a sensitivity of 0.8876, a specificity of 0.4074, and an F1-score of 0.7596.

Nyholm et al. proposed an optimized model for classifying sleep disturbances related to dementia using data from the SNAC-B dataset [21], derived from the Swedish National Study on Aging and Care in Blekinge, which involved 3,208 participants aged 60 to 99. The dataset comprised 5,033 samples and focused on participants' sleep disturbances. They employed five machine learning models (Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machine, and Gaussian Naive Bayes) to predict outcomes using 16 features. Among these models, Gradient Boosting achieved the highest accuracy of 92.9%, with an AUC score of 0.974 and an F1-score of 0.926 [22].

III. METHODOLOGY

A. DATASET DESCRIPTION

The dataset Sleep Health and Lifestyle employed in this study comprises 374 samples and 13 features. This dataset is publicly available on Kaggle at [1]. The features in the dataset consist of both categorical and numerical variables. Of the 13 features, five are categorical, representing discrete, non-quantitative attributes that capture qualitative characteristics within the data. Meanwhile, the remaining eight features are numerical, reflecting continuous quantitative measurements. The demographic features and other feature-related information, including category-wise percentages and ranges, are presented in Table 1. Because the 'Person ID' attribute comprises unique values, it does not provide meaningful information for the classification task. Therefore, we removed the 'Person ID' from the dataset at the beginning of the preprocessing phase.

This dataset presents a multiclass classification problem, where the objective is to classify each sample into one of the predefined categories. The dependent feature within this dataset is Sleep Disorder, which encompasses the categories None, Sleep Apnea, and Insomnia. The number of samples for each category in the main dataset appears in Table 2.

Figure 1 shows the flowchart that outlines the proposed optimized approach for sleep disorder classification. This flowchart details the sequence of steps involved in the application of classifiers and methodologies throughout the study. This visual representation provides a step-by-step breakdown of the data processing, model training, and evaluation phases.

B. MISSING VALUES HANDLING AND CATEGORICAL ENCODING

We verified that the dataset X using the equation (1) checks for the presence of null values across all features in the dataset. After running this check, we confirmed that the sum of missing values for each feature was zero, indicating that no missing values were present. Therefore, no data imputation or

TABLE 1. Demographic features and other feature data with percentages and ranges.

Features	Feature description	Range	Ratios (%)
Person ID	Unique ID	-	-
Gender	Male and Female	Male	50.53
		Female	49.47
Age	Person's age	27-35	25.15
		36-45	45.44
		46-52	13.64
		53+	15.77
Occupation	Person's profession	Accountant and Manager	10.16
		Doctor and Nurse	38.50
		Engineer and Software Engineer	17.91
		Lawyer	12.57
		Scientist and Teacher	11.77
		Sales Representative and salesperson	9.09
Sleep Duration	Daily sleep duration (hours)	5.8-6.4	25.65
		6.5-7.2	30.49
		7.3-7.9	24.87
		8.0+	18.99
Quality of Sleep	Quality of sleep, rated on a scale from 1 to 10	4-5	3.22
		6-7	48.66
		8	29.14
		9	18.98
Physical Activity Level	Minutes of physical activity per day	30-45	40.12
		46-60	21.66
		61-75	19.25
		76+	18.97
Stress Level	Stress level rating (1 to 10)	3-4	37.7
		5-6	30.21
		7-8	32.09
BMI Category	Person's BMI category	Normal	57.76
		Obese	2.67
		Overweight	39.57
Blood Pressure	Person's blood pressure (systolic/diastolic)	115/75-125/80	40.91
		125/82-130/85	29.93
		130/86+	29.16
Heart Rate	Resting heart rate (beats per minute)	65-74	83.93
		75+	16.07
Daily Steps	Daily step count	3000-7000	62.26
		7001+	37.74
Sleep Disorder	Presence of sleep disorder	None	58.56
		Insomnia	20.59
		Sleep Apnea	20.86

TABLE 2. Category-wise number of samples in the original dataset.

Sleep Disorder	Number of Samples
None	219
Sleep Apnea	78
Insomnia	77

handling of missing values was necessary. Machine learning algorithms interpret and process coded values of categorical data, transforming the qualitative information into numerical form. Our study employed the label encoding technique to perform categorical encoding [23], [24], [25]. Therefore, the assigned encoded values for sleep disorder as a dependent feature were Insomnia: 0, None: 1, Sleep Apnea: 2.

$$\text{missing\_values\_count} = X.\text{isnull}().\text{sum}() \quad (1)$$

C. DATA SPLITTING

We then divided this dataset into test and training sets. Notably, we used 80% of the data for training, as shown in Figure 2, which presents sample distribution across various categories within the training set. We used the remaining 20% for the test set. This partitioning strategy ensures a balanced data distribution for model training and evaluation.

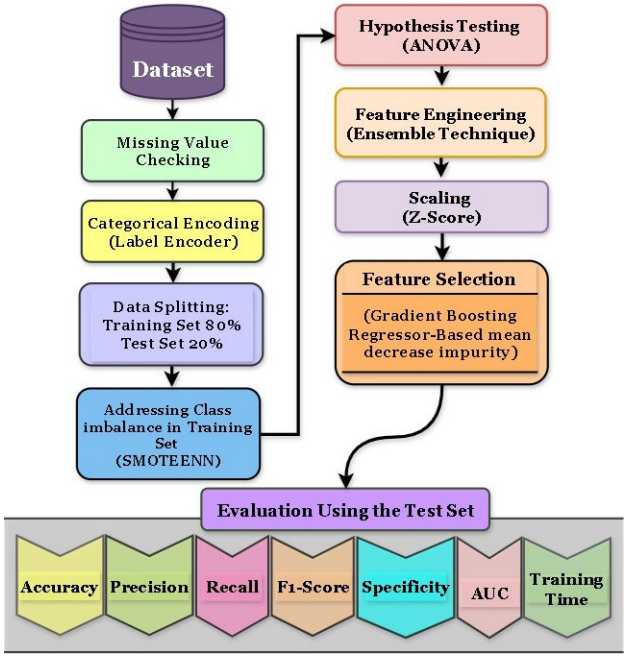

FIGURE 1. Flowchart of the proposed optimized approach for sleep disorder classification.

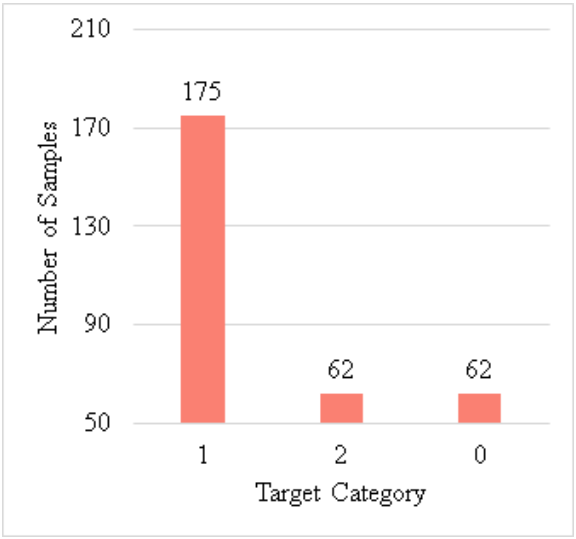

FIGURE 2. Category-wise number of samples in training set.

D. ADDRESSING CLASS IMBALANCE IN TRAINING SETS

The distribution of samples across different categories within the training set appears in Figure 1. Category 1 consists

of 175 samples, comprising approximately 58.52% of the total, while categories 2 and 0 have 62 and 62 samples accounting for about 20.74% and 20.74% of the whole training set, respectively. This disparity in sample distribution highlights an imbalance within the training set [25]. We addressed this problem with the Synthetic Minority Over-sampling Technique + Edited Nearest Neighbors (SMOTEENN), a data balancing technique [26], [27], [28], [29]. It synthetically generates new instances by SMOTE and then removes samples from all the likely noisy classes. SMOTEENN reduces the imbalance compared to the initial sample distribution in the training set. The minority classes have increased in size relative to the majority class, leading to a more balanced dataset, thus enhancing the performance of machine learning models trained on the data.

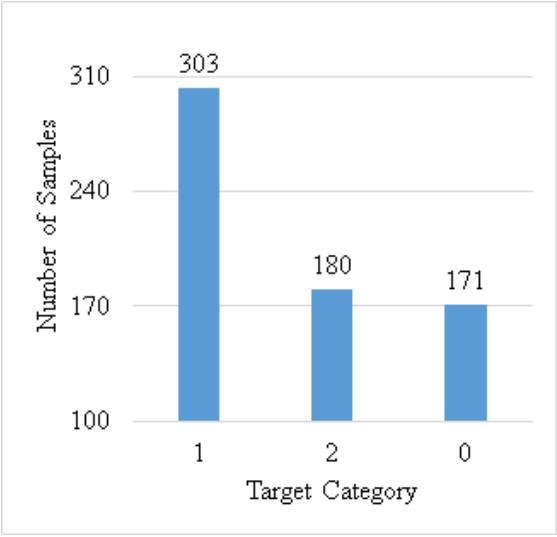

FIGURE 3. Category-wise number of samples in the balanced training set.

Figure 3 illustrates that the training set has reduced the imbalance with categories 1, 2, and 0 containing 303, 180, and 171 instances, respectively. Although the dataset is not perfectly balanced, the reduced imbalance is beneficial, thus increasing predictive outcomes with more accuracy on an aggregate basis.

E. HYPOTHESIS TESTING

Our research performed a one-way Analysis of Variance (ANOVA) test [30], [31], [32] on our dataset to examine several characteristics' statistical significance. We used a criterion of 0.05 to determine whether the features were statistically significant. Table 3 shows p-values per each feature. We observed that all features have p-values less than 0.05, which implies a rejection of the null hypothesis since they are significantly associated with the response variable. Hence, we dropped nothing from further examination to avoid compromising the strength of our analysis and to focus on key predictors only.

TABLE 3. Statistical significance of features: p-Value analysis.

Feature	Average p-value
BMI Category	$2.9349 \times e^{-220}$
Blood Pressure	$1.2415 \times e^{-142}$
Occupation	$5.4008 \times e^{-96}$
Physical Activity Level	$1.24168 \times e^{-58}$
Age	$3.94005 \times e^{-55}$
Daily Steps	$9.21527 \times e^{-39}$
Heart Rate	$1.34324 \times e^{-38}$
Sleep Duration	$2.09984 \times e^{-38}$
Gender	$2.05423 \times e^{-36}$
Quality of Sleep	$3.48851 \times e^{-26}$
Stress Level	$3.70292 \times e^{-11}$

F. FEATURE ENGINEERING

The most important part of machine learning, which is called feature engineering [33], [34], [35], [36], is to create new features or alter existing ones. This process enhances the performance and explanations of such models by making accurate predictions and perceiving underlying data patterns more deeply in our studies. In this work, following hypothesis testing, we engineered seven new features (created_feature1, created_feature2, created_feature3, created_feature4, created_feature5, created_feature6, and created_feature7) comprising the seven classifiers: Random Forest, Gradient Boosting, Gaussian Naive Bayes, K-nearest neighbors (KNN), Decision Tree, Logistic Regression, and Support Vector Machine, [37], [38]. We trained each classifier C_i on the training dataset (X_{train}, y_{train}) and generated predictions for the training set using the equation (2). We combined these predictions $\hat{y}_{i,\text{train}}$ with the original training features X_{train} to create an augmented training set X'_{train} using the equation (3).

$$\hat{y}_{i,\text{train}} = C_i(X_{\text{train}}) \tag{2}$$

$$X'_{\text{train}} = \left[X_{\text{train}}, \hat{y}_{1,\text{train}}, \hat{y}_{2,\text{train}}, \hat{y}_{3,\text{train}}, \hat{y}_{4,\text{train}}, \hat{y}_{5,\text{train}}, \hat{y}_{6,\text{train}}, \hat{y}_{7,\text{train}}\right] \tag{3}$$

Similarly, we independently trained the same classifiers, denoted as D_i, from scratch on the test dataset (X_{test}, y_{test}) and generated predictions for the test set using the equation (4). We combined these predictions $\hat{y}_{i,\text{test}}$ with the original test features X_{test} to create an augmented test set X'_{test} using the equation (5).

$$\hat{y}_{i,\text{test}} = D_i(X_{\text{test}}) \tag{4}$$

$$X'_{\text{test}} = \left[X_{\text{test}}, \hat{y}_{1,\text{test}}, \hat{y}_{2,\text{test}}, \hat{y}_{3,\text{test}}, \hat{y}_{4,\text{test}}, \hat{y}_{5,\text{test}}, \hat{y}_{6,\text{test}}, \hat{y}_{7,\text{test}}\right] \tag{5}$$

Since we independently trained the classifiers on the training and test sets, the models are not biased for classification. This approach enriches the training and test sets with additional features derived from classifier predictions, enhancing the model's ability to capture complex patterns and relationships in the data. The number of features in the feature space becomes 18 after adding seven new features.

G. SCALING

We improved our feature set by applying Z-score transformation to standardize the features and adjusting them to have an average of 0 and a standard deviation of 1 [39], [40]. This scaling process prevented some features from influencing the model prediction too much, leading to a more balanced and reliable learning process. We can calculate the Z-score transformation using equation (6).

$$Z = \frac{x - \mu}{\sigma} \tag{6}$$

where x, μ, σ, and Z represent the original value, mean, standard deviation, and standardized value of the feature.

H. FEATURE SELECTION

Feature selection is a process in machine learning models that helps prevent overfitting, improve generalization, and simplify model structure [41]. We employed the gradient boosting regressor-based mean decrease impurity (MDI) for calculating the importance scores of the features [42], [43]. Figures 4 and 5 illustrate the distribution of the importance scores across all features within these engineered and original feature spaces. We observed that some new features generated from classifier predictions exhibited higher importance scores than some original features. This outcome is consistent with our feature augmentation strategy. By incorporating predictions from various classifiers, we introduce additional dimensions of information that capture complex patterns not represented by the original features alone. For instance, these new features can capture nuanced variations in sleep patterns and behavioral characteristics, which are crucial for accurately diagnosing and treating sleep disorders. We selected the top five most significant features in the first experiment based on the gradient-boosting regressor-based mean decrease impurity (MDI) scores from both feature spaces. Therefore, we simplified the classification process by eliminating the bottom 13 and six features from the engineering and original feature spaces, respectively. In the second experiment, we applied the same technique to select the two most critical features from both spaces.

In the first experiment, the five selected features from the engineered and original feature spaces were: created_feature1, created_feature5, created_feature2, Sleep Duration, and Age; and Blood Pressure, BMI Category, Daily Steps, Sleep Duration, and Occupation, respectively. Within the original feature space, the rank of the selected feature (Age) from the extended feature space was 6. This result provides insights into the relative importance of features within two different feature spaces and how they can aid in determining crucial factors for successful classification models.

I. MACHINE LEARNING CLASSIFIERS

The classifiers used in this study for the multiclass classification problem include Logistic Regression, a linear model

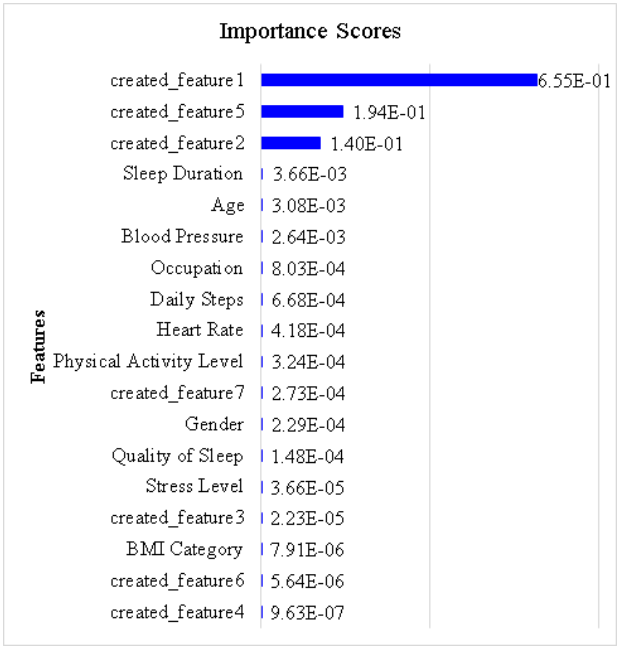

FIGURE 4. Feature importance on the extended feature space.

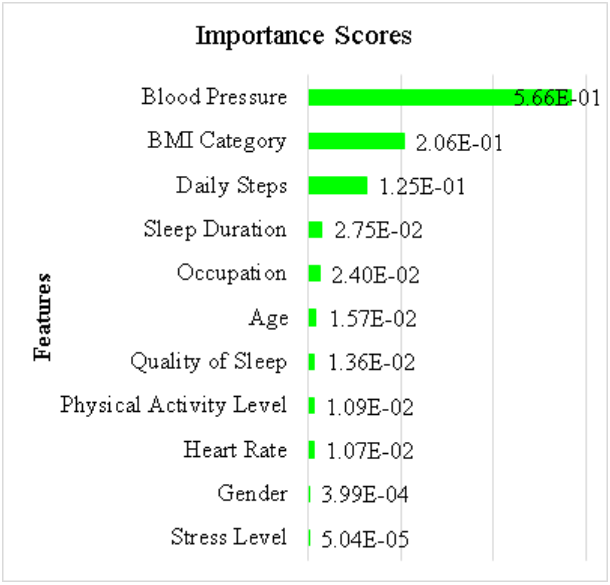

FIGURE 5. Feature importance on the original feature space.

that predicts probabilities by applying the logistic function. It is extended to Multinomial Logistic Regression when multi_class='auto' with the solver set to 'lbfgs' [44], [45]. Decision Tree constructs a tree-like model of decisions based on feature splits [46], [47], [48], [49]. In high-dimensional spaces, Support Vector Machines (SVMs) find an optimal hyperplane for class separation. Extra Trees is an ensemble of unpruned decision trees with random feature selection and splitting [47], [50]. Random Forest is an ensemble

method that aggregates multiple decision trees trained on random subsets of data [15], [51], [52]. eXtreme Gradient Boosting (XGBoost) is an optimized gradient boosting algorithm that incorporates regularization and second-order derivatives [53], [54]. Gradient Boosting is a boosting method that iteratively improves model accuracy by correcting residual errors [10], [54]. Light Gradient Boosting Machine (LightGBM) is a faster gradient boosting model utilizing histogram-based methods and leaf-wise tree growth [10], [54]. Voting Classifier combines the predictions of multiple base models using majority voting or averaging [55], [56]. A Catboost is a gradient-boosting algorithm optimized for handling categorical data [54], [57]. A Stacking Classifier is a meta-model that combines the predictions of base models using a meta-classifier [58]. A Model Optimized by GridSearchCV tunes hyperparameters using exhaustive search and cross-validation [54], [59]. Tree-based Pipeline Optimization Tool (TPOT) is an automated machine learning tool that optimizes pipeline performance using genetic algorithms [54], [56]. Adaptive Boosting (AdaBoost) [60], [61] focuses on improving the performance of weak classifiers by emphasizing misclassified instances. Bernoulli Naive Bayes is a probabilistic model designed for binary features based on Bayes' theorem [15].

J. PERFORMANCE STANDARDS

When dealing with an imbalanced dataset, achieving a higher accuracy may not appropriately reflect the model's performance [62]. Conversely, a slower model might not be suitable for real-world applications. Therefore, we employed seven evaluation metrics: accuracy, precision, recall, F1-score, specificity, AUC (Area Under the Curve), training time, and testing time to evaluate our model.

Regarding sleep disorder classification, accuracy, precision, recall, F1-score, specificity, AUC score, and training time assess the model's performance and practical viability. Accuracy measures the overall correctness of the classification, providing a general sense of the model's performance. Precision and recall evaluate the model's ability to identify and capture the presence of specific sleep disorders correctly. Precision ensures that predicted positive cases are truly positive, while recall ensures the identification of most of the actual positive cases. The F1-score balances precision and recall, which is useful when the distribution of disorders is imbalanced. Specificity is crucial for understanding how well the model identifies the absence of sleep disorders, which reduces false positives. The AUC score provides insight into the model's ability to distinguish between sleep disorders across various decision thresholds, reflecting its overall discriminative power. Training time indicates the efficiency of the model's learning process, which is crucial for deploying practical solutions responsive to real-time needs. These metrics ensure that the classification model effectively and efficiently diagnoses sleep disorders, balancing accuracy with practical considerations.

The mathematical formulations for the accuracy, precision, recall, F1-score, and specificity appear in the equations (7-11), where TP, TN, FP, and FN indicate true positive, true negative, false positive, and false negative, respectively [63], [64].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F1-score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{11}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In our experiment, we evaluated 15 classifiers for this task, including 13 default machine learning classifiers with parametric fine-tuning and two classifiers optimized by the usage of hyperparameter tuning strategies (GridSearchCV and Tree-based Pipeline Optimization Tools (TPOT)). The initial parameter settings used for GridSearchCV and TPOT appear in Tables 4 and 5, respectively. We applied these classifiers to both engineered and original feature spaces, and the experimental outcomes of the first and second experiments appear in Tables 6, 7, 9, and 10. The tables represent Accuracy as Acc, Precision as Pre, Recall as Rec, F1-Score as F1, Specificity as Spe, AUC Score as AUC, and Training Time as TT.

TABLE 4. GridSearchCV initial parameters.

Parameter Name	Value
cv	3
estimator__loss_function	MultiClass
iterations	[50, 100]
learning_rate	[0.003, 0.001, 0.0003]
depth	[4, 6, 8]
refit	True
return_train_score	False
scoring	accuracy

TABLE 5. TPOT initial parameters.

Parameter Name	Value
crossover_rate	0.1
cv	5
disable_update_check	False
generations	5
max_eval_time_mins	5
mutation_rate	0.9
n_jobs	1
population_size	20
subsample	1.0
use_dask	False
verbosity	2
warm_start	False

A. FIRST EXPERIMENT (PREDICTION USING FIVE KEY FEATURES)

In this subsection, we detail the results obtained from the classifiers using only five key features. The selected features—Blood Pressure, BMI Category, Daily Steps, Sleep Duration, and Occupation—are clinically relevant to sleep disorder diagnosis. Elevated blood pressure is linked to sleep apnea, while high BMI is a known risk factor for sleep disturbances. Daily steps reflect physical activity, which impacts sleep quality. Sleep duration directly measures sleep quality, and abnormalities are indicators of sleep disorders. Occupation, including stress and shift work, also affects sleep patterns. These features were chosen for their predictive power and clinical significance in diagnosing sleep disorders. Table 6 summarizes the outcomes obtained from the engineered feature space. Here, the Logistic Regression attains the highest performance across the evaluation metrics: accuracy (94.67%), precision (.9459), recall (.9467), F1-score (.9452), specificity (.9153), and AUC score (.9147). Conversely, Table 7 presents the overall performance of the classifiers on the original feature space. The top four Classifiers show identical scores for the metrics: accuracy (97.33%), precision (0.9733), recall (0.9733), F1-score (0.9733), and specificity (0.9569). However, the Gradient Boosting shows the highest AUC score (0.9953) among all the classifiers on the original feature space. The tables indicate that the models demonstrated significant performance improvements compared to their counterparts in the engineered feature space. This result suggests that the unique capabilities of the models effectively captured the data complexities present in the original feature space, thereby enhancing model generalization.

Comparing outcomes between the engineered and original feature spaces reveals significant insights. Interestingly, the original feature space consistently outperformed the engineered feature space across all models, emphasizing the robust and inherent predictive strength of the original feature space. Models trained on the original features consistently demonstrated higher performance and reliability, achieving top scores. In contrast, models trained on the engineered feature space exhibited slightly lower performance but were faster than those trained on the original feature space.

Tables 6 and 7 demonstrate that the Gradient Boosting classifier performed best in the original feature spaces, achieving 2.66% better accuracy than the top model (Logistic Regression) in the engineered feature space. For training time, Logistic Regression and Gradient Boosting took 0.0156 and 0.2189 seconds in the engineered and original feature spaces, respectively. Despite Gradient Boosting's superior accuracy in the original feature space, Logistic Regression is 14.03 times quicker in the engineered feature space. As Gradient Boosting demonstrates higher accuracy, it is recommended as the superior model in the original feature space using five key features. Table 8 lists the optimized parameters for training the classifier in both spaces. We identified these parameters as optimal for the classification task and applied them to train the model accordingly.

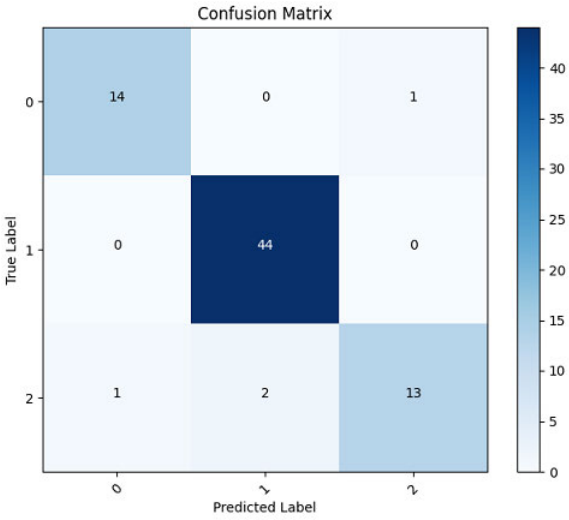

FIGURE 6. The confusion matrix of top classifier (Logistic Regression) trained on the engineering feature space using five features.

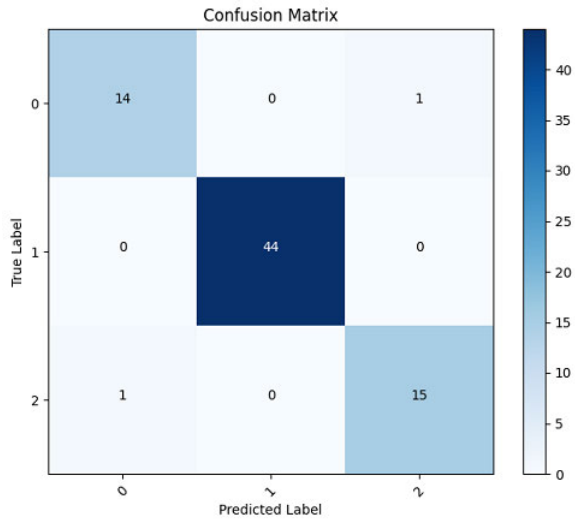

FIGURE 7. The confusion matrix of top classifier (Gradient Boosting) trained on the original feature space using five features.

This analysis utilized confusion matrices to evaluate the performance of the top classifier from each feature space. These matrices appear in Figures 6 and 7, which indicate the performance of the classifiers by showing the number of events distributed in each class and emphasizing the misclassifications.

Figures 6 and 7 show the confusion matrices of the Logistic Regression and Gradient Boosting classifiers applied to the engineered and original feature spaces, respectively. Figure 6 shows that Class 0 was misclassified as Class 2 once, Class 2 was misclassified as Class 0 once, and

TABLE 6. The results on engineered features space using five features.

Model Name	Acc (%)	Pre	Rec	F1	Spe	AUC	TT
Logistic Regression	94.67	0.9459	0.9467	0.9452	0.9153	0.9147	0.0156
Decision tree	93.33	0.9337	0.9333	0.9324	0.9077	0.9493	0.0023
Support Vector Machine	93.33	0.9403	0.9333	0.9345	0.9077	0.9064	0.1089
Extra Trees	93.33	0.9403	0.9333	0.9345	0.9077	0.9561	0.1095
Random Forest	93.33	0.9403	0.9333	0.9345	0.9077	0.9735	0.11
eXtreme Gradient Boosting	93.33	0.9403	0.9333	0.9345	0.9077	0.9235	0.2187
Gradient Boosting	93.33	0.9403	0.9333	0.9345	0.9077	0.9514	0.2298
Light Gradient Boosting Machine	93.33	0.9337	0.9333	0.9324	0.9077	0.9233	0.4555
Voting Classifier	93.33	0.9337	0.9333	0.9324	0.9077	0.9613	0.6098
Catboost	93.33	0.9403	0.9333	0.9345	0.9077	0.9722	1.5323
Stacking Classifier	93.33	0.9337	0.9333	0.9324	0.9077	0.9653	3.5999
Model Optimized by GridSearchCV	93.33	0.9403	0.9333	0.9345	0.9077	0.9928	4.0976
Tree-based Pipeline Optimization Tool (TPOT)	93.33	0.9337	0.9333	0.9324	0.9077	0.9691	73.6066
Adaptive Boosting	92.00	0.9198	0.9200	0.9188	0.8855	0.9053	0.0877
Bernoulli Naive Bayes	76.00	0.5981	0.7600	0.6692	0.6369	0.7534	0.0031

TABLE 7. The results on original features space using five features.

Model Name	Acc (%)	Pre	Rec	F1	Spe	AUC	TT
Gradient Boosting	97.33	0.9733	0.9733	0.9733	0.9569	0.9953	0.2189
Voting Classifier	97.33	0.9733	0.9733	0.9733	0.9569	0.9924	0.735
Catboost	97.33	0.9733	0.9733	0.9733	0.9569	0.9866	1.2822
Stacking Classifier	97.33	0.9733	0.9733	0.9733	0.9569	0.9906	4.4133
Decision Tree	94.67	0.9495	0.9467	0.9466	0.9153	0.9335	0.002
Support Vector Machine	94.67	0.9495	0.9467	0.9466	0.9153	0.8631	0.0937
Extra Trees	94.67	0.9495	0.9467	0.9466	0.9153	0.9631	0.1093
eXtreme Gradient Boosting	94.67	0.9495	0.9467	0.9466	0.9153	0.9816	0.2343
Light Gradient Boosting Machine	94.67	0.9495	0.9467	0.9466	0.9153	0.9758	0.3297
Tree-based Pipeline Optimization Tool (TPOT)	94.67	0.9495	0.9467	0.9466	0.9153	0.9781	58.9508
Model Optimized by GridSearchCV	92.00	0.9263	0.9200	0.9222	0.8987	0.9924	7.0523
Random Forest	90.67	0.9236	0.9067	0.9096	0.8793	0.9866	0.1249
Logistic Regression	89.33	0.8993	0.8933	0.8954	0.8543	0.9664	0.0313
Adaptive Boosting	89.33	0.9051	0.8933	0.8964	0.8822	0.9273	0.053
Bernoulli Naive Bayes	72.00	0.7815	0.7200	0.6908	0.6359	0.8575	0.0156

TABLE 8. The parameters used for training the best classifier (Gradient Boosting) using five features.

Parameter Names	Value
ccp_alpha	0.013
criterion	friedman_mse
init	None
learning_rate	0.001
loss	log_loss
max_depth	28
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	15
min_samples_split	10
min_weight_fraction_leaf	0.0
n_estimators	47
n_iter_no_change	None
random_state	0
subsample	0.8653
tol	0.0001
validation_fraction	0.1
verbose	0
warm_start	False

Class 2 was misclassified as Class 1 twice. Similarly, the Gradient Boosting applied to the original feature space shows that Class 0 was misclassified as Class 2 once, and Class 2 was misclassified as Class 0 once. These matrices play an important role in assessing the accuracy of each classifier in different classes. Figures 8 and 9 show the ROC curves of the top classifier from each feature space. These figures reveal higher AUC values of 0.9950, 1.0, and 0.9910 achieved for classes 0, 1, and 2, respectively, in the original feature space. This result demonstrates excellent classification performance, accurately classifying samples and emphasizing its usefulness in applications requiring accurate classification between groups.

B. SECOND EXPERIMENT (PREDICTION USING TWO KEY FEATURES)

Tables 9 and 10 present the outcomes from the engineered and original feature spaces, respectively, using only two key features for prediction. The tables indicate that the Decision Tree achieved equal accuracy in both spaces. However, the classifier showed slightly better AUC performance in the original feature space. Despite this result, it operated 14.93 times faster in the engineered feature space. Therefore, the Decision Tree is the optimal classifier using two features in the engineered feature space.

TABLE 9. The Outcomes on engineering feature space using two key features.

Model Name	Acc	Pre	Rec	F1	Spe	AUC	TT
Decision tree	0.9600	0.9617	0.9600	0.9606	0.9494	0.9488	0.0015
Adaptive Boosting	0.9600	0.9617	0.9600	0.9606	0.9494	0.9493	0.0312
Light Gradient Boosting Machine	0.9600	0.9617	0.9600	0.9606	0.9494	0.9075	0.061
Support Vector Machine	0.9600	0.9617	0.9600	0.9606	0.9494	0.9633	0.0679
Random Forest	0.9600	0.9617	0.9600	0.9606	0.9494	0.9481	0.12
Gradient Boosting	0.9600	0.9617	0.9600	0.9606	0.9494	0.9493	0.1939
Catboost	0.9600	0.9617	0.9600	0.9606	0.9494	0.9481	0.7499
Model Optimized by GridSearchCV	0.9600	0.9617	0.9600	0.9606	0.9494	0.9481	2.9072
Logistic Regression	0.9333	0.9337	0.9333	0.9324	0.9077	0.9224	0.0158
Extra Trees	0.9333	0.9403	0.9333	0.9345	0.9077	0.9481	0.0785
eXtreme Gradient Boosting	0.9333	0.9403	0.9333	0.9345	0.9077	0.9075	0.125
VotingClassifier	0.9333	0.9337	0.9333	0.9324	0.9077	0.9481	0.2343
StackingClassifier	0.9333	0.9337	0.9333	0.9324	0.9077	0.9481	1.6419
Tree-based Pipeline Optimization Tool (TPOT)	0.9333	0.9403	0.9333	0.9345	0.9077	0.9481	71.2079
Bernoulli Naive Bayes	0.7600	0.5981	0.7600	0.6692	0.6369	0.7534	0.0025

TABLE 10. The outcomes on original feature space using two key features.

Model Name	Acc	Pre	Rec	F1	Spe	AUC	TT
Decision Tree	0.9600	0.9606	0.9600	0.9599	0.9347	0.9549	0.0224
Support Vector Machine	0.9600	0.9606	0.9600	0.9599	0.9347	0.9612	0.0312
Extra Trees	0.9600	0.9606	0.9600	0.9599	0.9347	0.9538	0.0937
Catboost	0.9600	0.9606	0.9600	0.9599	0.9347	0.9685	0.8282
Tree-based Pipeline Optimization Tool (TPOT)	0.9600	0.9606	0.9600	0.9599	0.9347	0.9711	62.576
Adaptive Boosting	0.9333	0.9341	0.9333	0.9333	0.8931	0.9595	0.0966
Gradient Boosting	0.9333	0.9341	0.9333	0.9333	0.8931	0.9622	0.2248
Light Gradient Boosting Machine	0.9333	0.9341	0.9333	0.9333	0.8931	0.9931	0.2372
eXtreme Gradient Boosting	0.9333	0.9341	0.9333	0.9333	0.8931	0.9709	0.2656
Voting Classifier	0.9333	0.9341	0.9333	0.9333	0.8931	0.9902	0.5003
Stacking Classifier	0.9333	0.9341	0.9333	0.9333	0.8931	0.9942	2.7199
Model Optimized by GridSearchCV	0.9067	0.9067	0.9067	0.9067	0.8765	0.9869	3.9388
Logistic Regression	0.8933	0.8960	0.8933	0.8945	0.8689	0.9511	0.0156
Random Forest	0.8667	0.8756	0.8667	0.8691	0.8273	0.9789	0.1507
Bernoulli Naive Bayes	0.7067	0.6765	0.7067	0.6624	0.6136	0.8474	0.0083

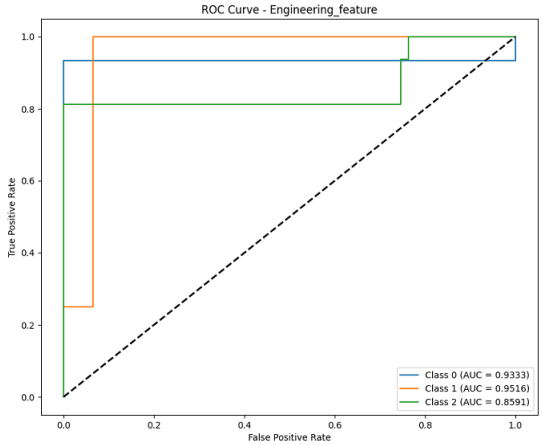

FIGURE 8. The ROC curve of Logistic Regression trained on the engineering feature space using five features.

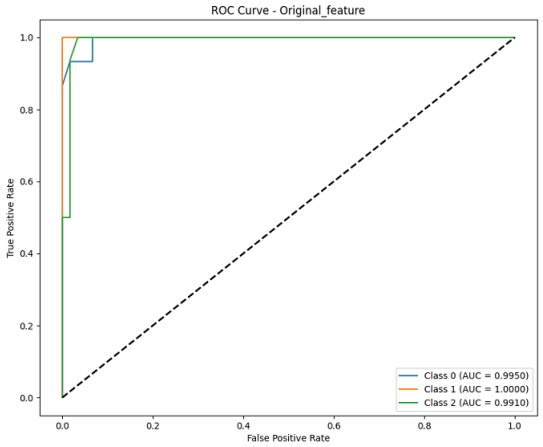

FIGURE 9. The ROC curve of Gradient Boosting trained on the original feature space using five features.

Table 11 documents the optimal parameters for training the Decision tree in both scenarios. These parameters were selected for their effectiveness in this classification task, utilizing only two key features. In this scenario, the confusion matrices, ROC curves, and trained models for the engineered and original feature spaces appear in Figures 10 to 15. The confusion matrices show that in the engineered feature space, the Decision Tree misclassified Class 0 as Class 2, Class 1 as Class 0, and Class 2 as Class 0 once each. Conversely, in the original feature space, the same classifier misclassified Class

0 as Class 2 twice and Class 2 as Class 0 once. The ROC curves demonstrate that Class 1 achieved an AUC score of 1.0 in the original feature space.

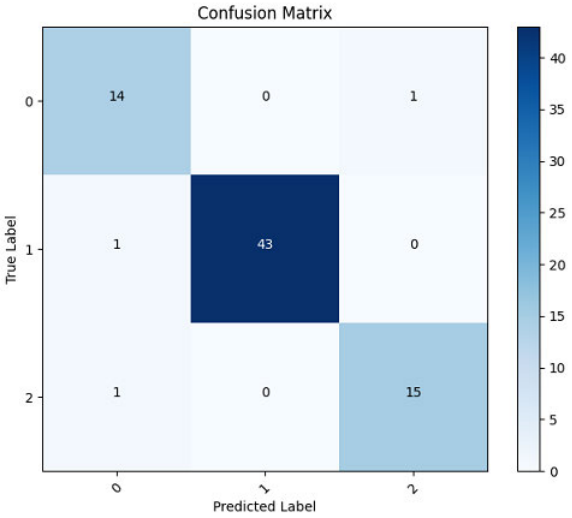

FIGURE 10. The confusion matrix of top classifier (Decision Tree) trained on the engineering feature space using two key features.

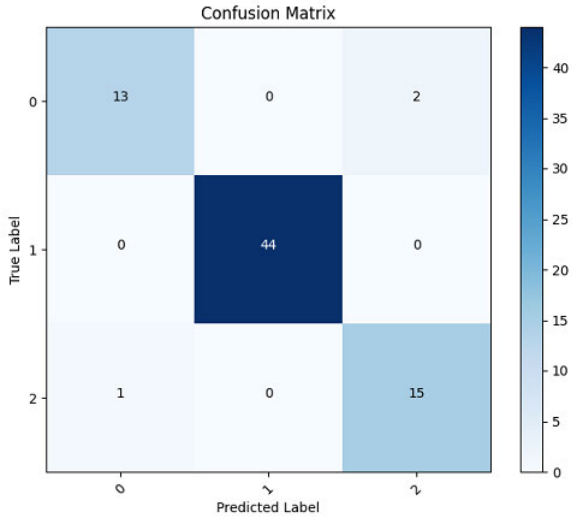

FIGURE 11. The confusion matrix of top classifier (Decision Tree) trained on the original feature space using two key features.

Figures 14 and 15 demonstrate the trained Decision Tree on both spaces. The heights of the decision trees in Figures 14 and 15 are 2 and 9, respectively. This result indicates that the decision tree classifier trained on the engineered feature space is approximately 4.5 times faster than the one trained on the original feature space. The research underscores the crucial role of data preparation in optimizing model performance. It enables models to learn more effectively and apply their knowledge to new situations. Conversely, poor preparation can diminish model effectiveness.

C. COMPARISON OF TRAINING AND TESTING TIME EFFICIENCY

Figure 16 compares the training and testing times of the top-performing two models (Gradient Boosting and Decision Tree) in these experiments. Gradient Boosting achieved the

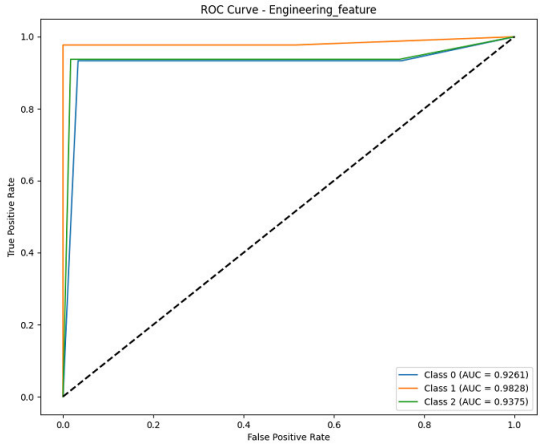

FIGURE 12. The ROC curve of Decision Tree trained on the engineering feature space using two key features.

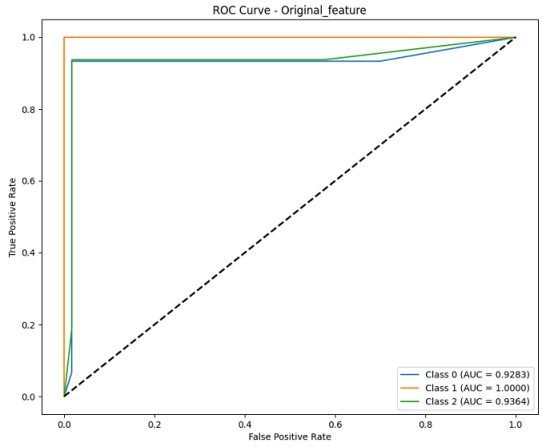

FIGURE 13. The ROC curve of Decision Tree trained on the original feature space using two key features.

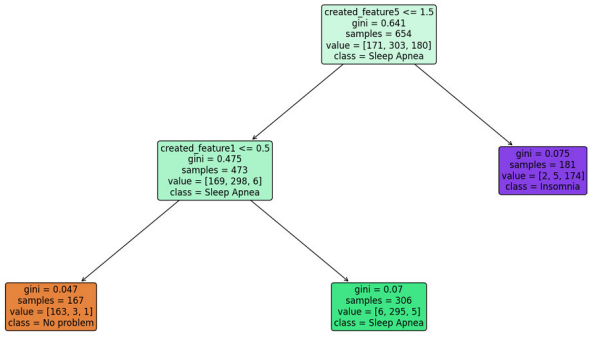

FIGURE 14. Trained Decision Tree on the engineering feature space using two key features.

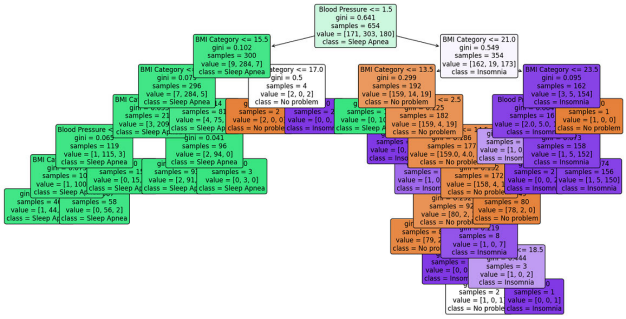

FIGURE 15. Trained Decision Tree on the original feature space using two key features.

TABLE 11. The parameters used for training the best classifier (Decision Tree) using two features.

Parameter Names	Value
ccp_alpha	0.013
class_weight	None
criterion	gini
max_depth	13
max_features	2
max_leaf_nodes	None
min_impurity_decrease	0.0
min_samples_leaf	1
min_samples_split	14
min_weight_fraction_leaf	0.0
monotonic_cst	None
random_state	0
splitter	best

highest results using five selected features in the original feature space. By comparison, the Decision Tree achieved the best performance with two selected features in the engineered feature space. However, Figure 16 shows that Gradient Boosting takes 145.93 times longer in training and 2.41 times longer in testing than the Decision Tree. This trade-off emphasizes the potential benefit of using faster models in real-time or resource-constrained environments, where computational efficiency is crucial.

D. OVERFITTING AVOIDANCE

We applied ccp_alpha = 0.013 in both Gradient Boosting and Decision Tree to mitigate overfitting. The ccp_alpha parameter controls tree pruning in both models, serving as a regularization technique by penalizing model complexity. This approach promotes better generalization for unseen data and maintains a balance between model fit and complexity, resulting in robust performance across diverse datasets.

In summary, our comprehensive experiments demonstrate that models trained using only five key features (Blood Pressure, BMI Category, Daily Steps, Sleep Duration, and Occupation) from the original feature space, particularly the Gradient Boosting classifier, performed the best for this task. This model consistently achieved outstanding scores across all evaluation metrics except training and testing time, underscoring its reliability and effectiveness in making accurate predictions. However, when computational efficiency is essential, the Decision Tree model, using two selected features in the engineered feature space, proves more effective. It offers faster training and testing times while maintaining competitive performance, though it is slightly less accurate than Gradient Boosting.

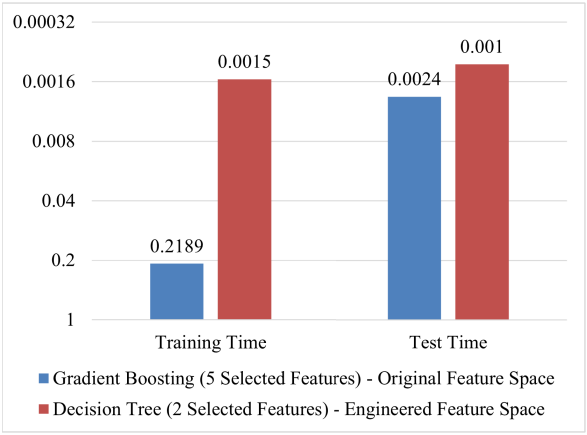

FIGURE 16. Comparison of Training and Testing Times for Top Models in Original and Engineered Feature Spaces.

These findings can be applied in real-world industry settings, such as creating AI-based diagnostic tools for hospitals and sleep clinics to automatically detect sleep disorders. For example, the developed models can be integrated into wearable devices or healthcare management systems to monitor patients in real-time, enabling accurate and timely predictions for conditions like sleep apnea or insomnia.

V. COMPARATIVE ANALYSIS

Table 12 presents a comparative analysis of the recent studies focusing on sleep health datasets and employing various models, detailing their respective accuracies. In this study, we achieved an outstanding accuracy of 97.33% using the Gradient Boosting classifier on the Sleep Health and Lifestyle Dataset despite using only five features from the original feature space for classification. This result surpasses other studies using the same dataset. According to a study [2] published in IEEE Access in 2024, an ANN model achieved the highest accuracy of 92.92% using five features on the same dataset. Our proposed model outperforms this model by 4.41% in accuracy, using the same number of features. Tareq et al. in [15] and Warunlawan et al. in [16] achieved an accuracy of 88% and 91.90% by Random Forests and Bagged Trees, respectively, on the same dataset. These findings underscore the superior performance of the Gradient Boosting classifier in the present study compared to alternative models and methodologies applied to the identical dataset in recent research. Additionally, our model obtained higher accuracy than the other studies in the table.

TABLE 12. Comparative analysis of model accuracies in sleep health datasets.

Ref	year	Dataset Type	Model	Accuracy
This study	-	Sleep Health and Lifestyle Dataset	Gradient Boosting	97.33%
[2]	2024	Sleep Health and Lifestyle Dataset	ANN	92.92%
[15]	2024	Sleep Health and Lifestyle Dataset	Random Forests	88%
[16]	2024	Sleep Health and Lifestyle Dataset	Bagged Trees	91.90%
[57]	2023	Medical Centre	KNN	91%
[65]	2021	ISRUC-SLEEP	RF	94.46%
[17]	2023	Montreal Archive of Sleep Studies (MASS)	SwSleepNet	86.7%
[66]	2024	MIT-BIH Polysomnographic and Sleep-EDF expanded databases	K-fold cross validation	97.14% for three classes
[67]	2023	polysomnography (PSG) data	biLSTM	71.2±5.8%
[68]	2023	PhysioNet Apnea-Electrocardiogram	MobileNet V1 + GRU	90.29%
[69]	2022	Sleep-EDFX-8	EEGSNet	94.17%
[70]	2020	Cyclic Alternating Pattern (CAP) sleep data	Bagged tree	86.27%
[71]	2021	The Wisconsin Sleep Cohort (WSC)	CNN+ LSTM	87.4%

VI. CONCLUSION

This paper introduces an optimized classifier for sleep disorder classification, leveraging machine learning algorithms (MLAs) to achieve superior performance. The primary objective is to enhance classification accuracy by evaluating and optimizing various MLAs using the Sleep Health and Lifestyle Dataset without relying on expert-defined features. The novelty of this work lies in evaluating and optimizing MLAs within the original feature space of the dataset, without relying on expert-defined features.

Our findings demonstrate that the proposed approach, which applied effective preprocessing strategies, significantly improved model performance. Specifically, Gradient Boosting was identified as the most effective method, outperforming all state-of-the-art techniques by achieving the highest accuracy, precision, recall, F1-score, and AUC, highlighting its superior classification performance and computational efficiency. The optimized Gradient Boosting classifier achieved an outstanding accuracy of 97.33%, with Precision, Recall, and F1-score of 0.9733, and Specificity and AUC scores of 0.9569 and 0.9953, respectively. These results confirm the effectiveness of our approach in accurately classifying sleep disorders.

While the study highlights the potential of MLAs in this domain, it also acknowledges limitations due to the dataset size. Further research directions include integrating unsupervised learning methods and advanced feature extraction techniques to improve classification accuracy and robustness. Additionally, expanding the dataset and exploring state-of-the-art techniques will further enhance the model's generalizability and real-world applicability. From a practical perspective, the proposed methodology can be applied in real-world health diagnostics, such as automated systems in sleep clinics or wearable health monitoring devices, to facilitate early detection and classification of sleep disorders. These systems can significantly reduce the reliance on manual evaluations, streamline diagnosis, and improve accessibility to care, particularly in resource-constrained settings.

REFERENCES

[1] L. Tharmalingam. (2024). *Sleep Health and Lifestyle Dataset*. Accessed: Jun. 26, 2024. [Online]. Available: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

[2] T. S. Alshammari, "Applying machine learning algorithms for the classification of sleep disorders," *IEEE Access*, vol. 12, pp. 36110–36121, 2024.

[3] X. Xiao, Y. Rui, Y. Jin, and M. Chen, "Relationship of sleep disorder with neurodegenerative and psychiatric diseases: An updated review," *Neurochemical Res.*, vol. 49, no. 3, pp. 568–582, Mar. 2024.

[4] M. Braun, S. Dietz-Terjung, U. Sommer, C. Schoebel, and C. Heiser, "Stated patient preferences for overnight at-home diagnostic assessment of sleep disorders," *Sleep Breathing*, vol. 28, no. 5, pp. 1939–1949, Oct. 2024.

[5] 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.

[6] S. K. Satapathy and D. Loganathan, "Automated classification of multi-class sleep stages classification using polysomnography signals: A nine-layer 1D-convolution neural network approach," *Multimedia Tools Appl.*, vol. 82, no. 6, pp. 8049–8091, Mar. 2023.

[7] G. Kong, C. Li, H. Peng, Z. Han, and H. Qiao, "EEG-based sleep stage classification via neural architecture search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1075–1085, 2023.

[8] Y. Shao, B. Huang, L. Du, P. Wang, Z. Li, Z. Liu, L. Zhou, Y. Song, X. Chen, and Z. Fang, "Reliable automatic sleep stage classification based on hybrid intelligence," *Comput. Biol. Med.*, vol. 173, May 2024, Art. no. 108314.

[9] X. Ji, Y. Li, P. Wen, P. Barua, and U. R. Acharya, "MixSleepNet: A multi-type convolution combined sleep stage classification model," *Comput. Methods Programs Biomed.*, vol. 244, Feb. 2024, Art. no. 107992.

[10] C.-H. Tai, T.-Y. Liao, S.-P. Chen, and M.-H. Chung, "Sleep stage classification using light gradient boost machine: Exploring feature impact in depressive and healthy participants," *Biomed. Signal Process. Control*, vol. 88, Feb. 2024, Art. no. 105647.

[11] 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.

[12] 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.

[13] A. Şenol, T. Talan, and C. Aktürk, "A new hybrid feature reduction method by using MCMSTClustering algorithm with various feature projection methods: A case study on sleep disorder diagnosis," *Signal, Image Video Process.*, vol. 18, no. 5, pp. 4589–4603, Jul. 2024.

[14] 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.

[15] W. Z. T. Tareq, "Sleep disorders detection and classification using random forests algorithm," in *Decision Making in Healthcare Systems*. Cham, Switzerland: Springer, 2024, pp. 257–266.

[16] M. Warunlawan, P. Homsud, P. Sapphaphab, O. Rinthon, and S. Pechprasarn, "Identification of crucial factors in sleep quality using machine learning models and MRMR feature selection technique," in *Proc. 15th Biomed. Eng. Int. Conf. (BMEiCON)*, Oct. 2023, pp. 1–5.

[17] H. Zhu, Y. Wu, Y. Guo, C. Fu, F. Shu, H. Yu, W. Chen, and C. Chen, "Towards real-time sleep stage prediction and online calibration based on architecturally switchable deep learning models," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 1, pp. 470–481, Jan. 2024.

[18] Z. Zhang, T. B. Conroy, A. C. Krieger, and E. C. Kan, "Detection and prediction of sleep disorders by covert bed-integrated RF sensors," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 4, pp. 1208–1218, Apr. 2023.

[19] A. Wadichar, S. Murarka, D. Shah, A. Bhurane, M. Sharma, H. S. Mir, and U. R. Acharya, "A hierarchical approach for the diagnosis of sleep disorders using convolutional recurrent neural network," *IEEE Access*, vol. 11, pp. 125244–125255, 2023.

[20] 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.

[21] *Snac—Swedish National Study on Aging and Care*. Accessed: Jul. 8, 2024. [Online]. Available: https://snd.se/en/catalogue/dataset/ext0124-1

[22] J. Nyholm, A. N. Ghazi, S. N. Ghazi, and J. S. Berglund, "Prediction of dementia based on older adults' sleep disturbances using machine learning," *Comput. Biol. Med.*, vol. 171, May 2024, Art. no. 108126.

[23] U. Michelucci, *Fundamental Mathematical Concepts for Machine Learning in Science*. Cham, Switzerland: Springer, 2024.

[24] G. J. Simon and C. Aliferis, *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*. Cham, Switzerland: Springer, 2024.

[25] G. Hackeling, *Mastering Machine Learning With Scikit-learn*. Birmingham, U.K.: Packt Publishing, 2017.

[26] R. Bounab, K. Zarour, B. Guelib, and N. Khlifa, "Enhancing medicare fraud detection through machine learning: Addressing class imbalance with SMOTE-ENN," *IEEE Access*, vol. 12, pp. 54382–54396, 2024.

[27] S. Dhanalakshmi, S. Das, and R. Senthil, "Speech features-based Parkinson's disease classification using combined SMOTE-ENN and binary machine learning," *Health Technol.*, vol. 14, no. 2, pp. 393–406, Mar. 2024.

[28] P. Sun, Z. Wang, L. Jia, and Z. Xu, "SMOTE-kTLNN: A hybrid re-sampling method based on SMOTE and a two-layer nearest neighbor classifier," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121848.

[29] M. H. Jamal, N. Naz, M. A. K. Khattak, F. Saeed, S. N. Altamimi, and S. N. Qasem, "A comparison of re-sampling techniques for detection of multi-step attacks on deep learning models," *IEEE Access*, vol. 11, pp. 127446–127457, 2023.

[30] Y. Xia and J. Sun, "Hypothesis testing and statistical analysis of microbiome," *Genes Diseases*, vol. 4, no. 3, pp. 138–148, Sep. 2017.

[31] C. M. Judd, G. H. McClelland, and C. S. Ryan, *Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond*. Evanston, IL, USA: Routledge, 2017.

[32] D. J. Denis, *SPSS Data Analysis for Univariate, Bivariate, and Multivariate Statistics*. Hoboken, NJ, USA: Wiley, 2018.

[33] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," *J. Inf. Intell.*, vol. 3, no. 2, pp. 113–153, Jan. 2024.

[34] P. Kumar and B. Pratap, "Feature engineering for predicting compressive strength of high-strength concrete with machine learning models," *Asian J. Civil Eng.*, vol. 25, no. 1, pp. 723–736, Jan. 2024.

[35] E. Hossain, *Machine Learning Crash Course for Engineers*. Cham, Switzerland: Springer, 2024.

[36] P. Gupta and N. K. Sehgal, "Practical aspects in machine learning," in *Introduction To Machine Learning With Security: Theory and Practice Using Python in the Cloud*. Cham, Switzerland: Springer, Jan. 2021, pp. 115–142.

[37] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Gener. Comput. Syst.*, vol. 117, pp. 47–58, Apr. 2021.

[38] D. Xibin, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers Comput. Sci.*, vol. 14, no. 2, pp. 241–258, Aug. 2020.

[39] A. Curtis, T. Smith, B. Ziganshin, and J. Elefteriades, "The mystery of the z-score," *AORTA*, vol. 4, no. 4, pp. 124–130, Aug. 2016.

[40] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 142–151.

[41] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Treviño, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv. (CSUR)*, vol. 50, no. 6, p. 94, Dec. 2017.

[42] M. Al-Sarem, F. Saeed, W. Boulila, A. H. Emara, M. Al-Mohaimeed, and M. Errais, "Feature selection and classification using catboost method for improving the performance of predicting parkinson's disease," in *Advances on Smart and Soft Computing: Proceedings of ICAC*. Cham, Switzerland: Springer, 2021, pp. 189–199.

[43] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "Genetic programming for feature selection based on feature removal impact in high-dimensional symbolic regression," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 3, pp. 2269–2282, Jun. 2024.

[44] T. Soni, D. Gupta, and M. Uppal, "Enhancing accuracy of sleep disorder with logistic regression model," in *Proc. IEEE 2nd Int. Conf. Ind. Electronics, Develop. Appl. (ICIDeA)*, Sep. 2023, pp. 292–295.

[45] T. Zhang, *Mathematical Analysis of Machine Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2023.

[46] K. Khosravi, N. Attar, S. M. Bateni, C. Jun, D. Kim, M. J. S. Safari, S. Heddam, A. Farooque, and S. Abolfathi, "Daily river flow simulation using ensemble disjoint aggregating M5-prime model," *Heliyon*, vol. 10, no. 20, Oct. 2024, Art. no. e37965.

[47] M. Gori, A. Betti, and S. Melacci, *Machine Learning: A Constraint-based Approach*. Amsterdam, The Netherlands: Elsevier, 2023.

[48] K. Khosravi, A. A. Farooque, A. Naghibi, S. Heddam, A. Sharafati, J. Hatamiafkoueieh, and S. Abolfathi, "Enhancing pan evaporation predictions: Accuracy and uncertainty in hybrid machine learning models," *Ecol. Informat.*, vol. 85, Mar. 2025, Art. no. 102933.

[49] A. Yeganeh-Bakhtiary, H. EyvazOghli, N. Shabakhty, and S. Abolfathi, "Machine learning prediction of wave characteristics: Comparison between semi-empirical approaches and DT model," *Ocean Eng.*, vol. 286, Oct. 2023, Art. no. 115583.

[50] M. A. Habib, J. J. O'Sullivan, S. Abolfathi, and M. Salauddin, "Enhanced wave overtopping simulation at vertical breakwaters using machine learning algorithms," *PLoS ONE*, vol. 18, no. 8, Aug. 2023, Art. no. e0289318.

[51] A. Mohammadpour, E. Gharehchahi, M. A. Gharaghani, E. Shahsavani, M. Golaki, R. Berndtsson, A. M. Khaneghah, H. Hashemi, and S. Abolfathi, "Assessment of drinking water quality and identifying pollution sources in a chromite mining region," *J. Hazardous Mater.*, vol. 480, Dec. 2024, Art. no. 136050.

[52] M. A. Habib, S. Abolfathi, J. J. O'Sullivan, and M. Salauddin, "Efficient data-driven machine learning models for scour depth predictions at sloping sea defences," *Frontiers Built Environ.*, vol. 10, Feb. 2024, Art. no. 1343398.

[53] Y. Wang, S. Ye, Z. Xu, Y. Chu, J. Zhang, and W. Yu, "Research on sleep staging based on support vector machine and extreme gradient boosting algorithm," *Nature Sci. Sleep*, vol. 16, pp. 1827–1847, Nov. 2024.

[54] T. Geetha and S. Sendhilkumar, *Machine Learning: Concepts, Techniques and Applications*. London, U.K.: Chapman & Hall, 2023.

[55] M. Irfan, H. A. Siddiqa, A. Nahliis, C. Chen, Y. Xu, L. Wang, A. Nawaz, A. Subasi, T. Westerlund, and W. Chen, "An ensemble voting approach with innovative multi-domain feature fusion for neonatal sleep stratification," *IEEE Access*, vol. 12, pp. 206–218, 2024.

[56] G. Kunapuli, *Ensemble Methods for Machine Learning*. New York, NY, USA: Simon and Schuster, 2023.

[57] 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.

[58] M. Nouman, S. Y. Khoo, M. A. P. Mahmud, and A. Z. Kouzani, "Advancing mental health predictions through sleep posture analysis: A stacking ensemble learning approach," *J. Ambient Intell. Humanized Comput.*, vol. 15, no. 9, pp. 3493–3507, Sep. 2024.

[59] G. Sravani, B. Lavanya, K. Mithila, and R. Surendran, "Exploring sleep disorder and lifestyle analysis through data preprocessing and ensemble learning techniques," in *Proc. 2nd Int. Conf. Sustain. Comput. Smart Syst. (ICSCSS)*, Jul. 2024, pp. 791–795.

[60] A. Taher and W. I. Z. Ayon, "Exploring sleep disorders: A comparative analysis of machine learning algorithms on sleep health and lifestyle data," in *Proc. IEEE Int. Conf. Power, Electr., Electron. Ind. Appl. (PEEIACON)*, Sep. 2024, pp. 71–75.

[61] M. Kumar, B. Ahmed, H. M. Mishra, A. K. Jha, P. K. Sikarwal, and S. Rampal, "Predictive sleep disorder modelling: Using machine learning with lifestyle and sleep health data," in *Proc. Int. Conf. Adv. Comput., Commun. Appl. Informat. (ACCAI)*, May 2024, pp. 1–7.

[62] 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.

[63] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation," 2020, *arXiv:2010.16061*.
[64] G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in *Proc. Comput. Sci. On-line Conf.* Cham, Switzerland: Springer, Jan. 2023, pp. 15–25.
[65] 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.
[66] S. Rashidi and B. M. Asl, "Strength of ensemble learning in automatic sleep stages classification using single-channel EEG and ECG signals," *Med. Biol. Eng. Comput.*, vol. 62, no. 4, pp. 997–1015, Apr. 2024.
[67] S. Morokuma, T. Hayashi, M. Kanegae, Y. Mizukami, S. Asano, I. Kimura, Y. Tateizumi, H. Ueno, S. Ikeda, and K. Niizeki, "Deep learning-based sleep stage classification with cardiorespiratory and body movement activities in individuals with suspected sleep disorders," *Sci. Rep.*, vol. 13, no. 1, p. 17730, Oct. 2023.
[68] P. Hemrajani, V. S. Dhaka, G. Rani, P. Shukla, and D. P. Bavirisetti, "Efficient deep learning based hybrid model to detect obstructive sleep apnea," *Sensors*, vol. 23, no. 10, p. 4692, May 2023.
[69] 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.
[70] E. R. Widasari, K. Tanno, and H. Tamura, "Automatic sleep disorders classification using ensemble of bagged tree based on sleep quality features," *Electronics*, vol. 9, no. 3, p. 512, Mar. 2020.
[71] 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, May 2023.

MD. ATIQUR RAHMAN received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Rajshahi, Bangladesh, and the Ph.D. degree in computer and information systems from the School of Computer Science and Engineering, The University of Aizu, Japan. Currently, he is an Assistant Professor with the Department of Computer Science and Engineering, East West University, Bangladesh. He possesses strong skills in complex programming and problem-solving. He has several publications in international journals and conferences. His research interests include machine learning, deep learning, and computer vision. In addition, he serves as a reviewer for several journals and conferences.

ISRAT JAHAN received the B.Sc. degree in computer science and engineering from East West University, Dhaka, Bangladesh. Her academic journey and professional interests encompass data science-based research and projects, focusing on machine learning, deep learning, and computer vision. Her dedication to research and continuous learning drives her to stay updated on the latest data science and AI advancements.

MAHEEN ISLAM (Member, IEEE) received the B.S. and M.S. degrees from the University of Dhaka, Bangladesh, in 1998 and 1999, respectively, and the Ph.D. degree in wireless mesh networking from the Department of Computer Science and Engineering, University of Dhaka, in 2017. She is currently an Associate Professor with the Department of Computer Science and Engineering, East West University, Bangladesh. Her research interests include wireless mesh networks, wireless sensor networks, cognitive radio networks, modeling and analysis of network communication protocols, and machine learning. She is a member of the Green Networking Research (GNR) Group.

TASKEED JABID received the bachelor's degree in computer science and engineering from East West University, Bangladesh, and the Ph.D. degree in computer vision and image processing from Kyung Hee University, South Korea. He is currently an Associate Professor at East West University. His research interests include machine learning, deep learning, image processing, facial image analysis, texture analysis, bioinformatics, computer graphics, and algorithms.

MD SAWKAT ALI received the B.Sc. degree in electrical and electronic engineering (EEE) from the Ahsanullah University of Science and Technology, Bangladesh, the M.Eng. (by Research) degree in electrical engineering from the University of New South Wales, Australia, and the Ph.D. degree in electrical engineering from Central Queensland University, Rockhampton, Australia. After completing the Ph.D. degree, he was a Postdoctoral Research Fellow at Central Queensland University and Deakin University, Australia. He is currently an Associate Professor with the Department of Computer Science and Engineering, East West University, Bangladesh. His research interests include the IoT, machine learning, power electronics, linear and nonlinear control theory and its applications, microgrids, and renewable energy.

MOHAMMAD RIFAT AHMMAD RASHID received the B.Sc. degree in computer science and engineering (CSE) from Khulna University, Bangladesh, the M.Sc. degree in computer science and engineering from The University of Pavia, Italy, and the Ph.D. degree in computer and control engineering from the Polytechnic University of Turin, Italy. He was a Researcher in the pervasive technologies research area at LINKS Foundation, working within the IoT Service Management Unit. He is currently an Assistant Professor with the Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh. His research interests include computer vision, the IIoT, blockchain technology, and the semantic web.

MOHAMMAD MANZURUL ISLAM received the B.Sc. degree in computer science and information technology from the Islamic University of Technology, Bangladesh, in 2005, the M.Sc. degree in internetworking from the University of Technology Sydney, Australia, in 2013, and the Ph.D. degree in cybersecurity from Federation University, Australia, in 2021. He is currently an Assistant Professor at East West University, Bangladesh. He served over 16 years in many organizations, including universities (Monash University, Australia; Federation University, Australia; Canadian University of Bangladesh; and Stamford University Bangladesh) and IT industries (ZTE Corporation and Robi-Axiata Ltd.). His research interests include cybersecurity, machine learning, computer vision, and the IoT.

MD. HASANUL FERDAUS received the B.Sc. (Engg.) degree in CSE from Bangladesh University of Engineering and Technology (BUET), Bangladesh, in 2004, the M.Sc. (Engg.) degree in ICT from Karlsruhe Institute of Technology, Germany, and the Polytechnic University of Turin, Italy, in 2009, and the Ph.D. degree in computer science from the Faculty of IT, Monash University, Australia, in 2016. Since September 2022, he has been an Assistant Professor with the Department of CSE, East West University, Bangladesh. He has taught at Australian universities for seven years, including Monash University, Central Queensland University, Melbourne Institute of Technology, and VIT. Also, he has over five years of research experience in Australia (CLOUDS Laboratory, Melbourne University, and Monash University) and Germany (Telematics Institute, KIT, and FZI Research Center for IT). Furthermore, he was a System Analyst for two years at Robi Axiata, Bangladesh. His research interests include cloud computing, cybersecurity, the IoT, and software engineering.

MD MOSTOFA KAMAL RASEL received the Ph.D. degree from Kyung Hee University, South Korea. He was a Postdoctoral Researcher at Kyung Hee University. He is currently an Assistant Professor at the Department of Computer Science and Engineering, East West University, Bangladesh. His research interests include big data compression and management, SQL script optimization, and application decentralization.

MAHMUDA RAWNAK JAHAN received the B.Sc. degree in computer science and engineering from the Military Institute of Science and Technology (MIST), Dhaka, Bangladesh, and the Erasmus Mundus Joint Master's degree in intelligent field robotic systems from Eötvös Loránd University, Hungary, and the University of Girona, Spain. She is currently a Lecturer with the Department of Computer Science and Engineering, East West University, Bangladesh. She has published in several international conferences. Her research interests include artificial intelligence, robotics, and human–robot interaction.

SHAYLA SHARMIN received the B.Sc. and master's degrees in CSE from JU Bangladesh. She is currently a Senior Lecturer with Daffodil International University. Before working in academia, she worked in various roles for leading ICT companies for more than four years. She has published many peer-reviewed academic papers in well-reputed conferences and journals in this area as an active researcher. Her research interests include deep learning and image processing, particularly in medical imaging and healthcare analysis. As an Academician, she is an active member of several professional bodies.

TANZINA AFROZ RIMI (Graduate Student Member, IEEE) is a passionate Researcher in artificial intelligence and machine learning, focusing on medical image analysis and disease detection. Currently, she is with Daffodil International University. With a solid background in computer science and a keen interest in healthcare technology, she has dedicated her efforts to developing innovative solutions to address real-world challenges. Her goal is to leverage the power of AI to enhance healthcare outcomes and contribute to the betterment of society.

ATIA SANJIDA TALUKDER has been a dedicated teacher working as a Lecturer with the Department of Computer Science and Engineering, Daffodil International University, since July 2023. She was a Lecturer with the Department of GED. During this period, she has taught various courses in statistics and probability, introduction to statistics, statistics for communication research with laboratory (using SPSS), and statistics for environment science. She is also a co-module leader and a module leader during this period. In addition, she is an active member of the committee to develop the OBE-based BSC syllabus for computational statistics under the Faculty of Science and Information Technology.

MD. MAFIUL HASAN MATIN received the bachelor's and master's degrees from Jahangirnagar University, Bangladesh. He is a Lecturer with the Department of Computer Science and Engineering, East West University, Bangladesh. With a strong passion for machine learning and deep learning, he dedicates his academic career to advancing knowledge in these fields. His work reflects a commitment to both teaching and research, contributing to the growing body of knowledge in computer science and its real-world applications.

M. AMEER ALI received the Ph.D. degree in IT from Monash University. He was the Managing Director of IT Company and a Cyber Security Consultant. He is a seasoned IT professional with over 22 years of experience. He has authored more than 55 research publications and has held various roles in academia, including the Department Head and the Dean of Faculty. His extensive expertise spans software development, network design, and cyber security, working with both private and public organizations. He holds several certifications, including PCI QSA, CEH, and ISO27001. Currently, he is a Senior Security Consultant, specializing in PCI and ISO27001 projects. His strong problem-solving skills, decision-making abilities, and proficiency in cybersecurity tools, make him a valuable asset to any organization.