UtilitasMathematica

ISSN 0315-3681 Volume 122 (2), 2025

Transformative Randomized Decision Trees for Sleep Disorder Classification from Health Life-Style Data

Babu Enthoti[1*], G. V. Nanda Kishore Reddy[2], Bujjayola Saivamshi Goud[2], Gattu Rishitha[2], Y. Inna Reddy[2], Akshaya Goud[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, Telangana, India

[2]Department of Computer Science and Engineering (AI & ML), Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, Telangana, India

*Corresponding Author: babuenthoti@mrem.ac.in

Abstract

Sleep disorders, such as insomnia and sleep apnea, impact a substantial portion of the global population, with insomnia affecting approximately 10% of adults and sleep apnea influencing up to 3%. Despite these significant statistics, existing diagnostic approaches face challenges such as incomplete datasets and suboptimal classification accuracy. Traditional methods often struggle with differentiating between various sleep disorders due to limitations in feature extraction and classification techniques. To address these challenges, a novel Sleep Disorders Classification (SDC) framework is proposed. This framework incorporates advanced preprocessing techniques to enhance data quality and uses Incremental Eigen Values Analysis (IEVA) for efficient feature extraction, which dynamically reduces the dimensionality of the dataset. The framework also employs Randomized Decision Trees (RDT) classification to accurately distinguish between insomnia, sleep apnea, and normal sleep conditions. By integrating these advanced methods, the SDC framework aims to improve the accuracy and reliability of sleep disorder diagnoses.

Keywords: sleep disorders, insomnia, sleep apnea, classification, Incremental Principal Component Analysis, RDT, feature extraction, preprocessing

1. Introduction

Sleep disorders are prevalent and affect a significant portion of the global population. Insomnia impacts approximately 10% of adults, leading to persistent difficulties in falling or staying asleep [1]. Sleep apnea, a condition characterized by repeated interruptions in breathing during sleep, affects up to 3% of the population. These statistics underscore the widespread nature of sleep disorders and highlight the need for effective diagnostic and treatment strategies to manage their impact on individuals' health and quality of life [2]. The application of accurate sleep disorder diagnostics is critical in various fields, including clinical medicine, public health, and personal wellness [3]. Early and precise identification of disorders like insomnia and sleep apnea can significantly improve patient outcomes by enabling timely intervention and tailored treatment plans. Moreover, the integration of accurate diagnostic tools into wearable health devices and telemedicine platforms can facilitate continuous monitoring and personalized care, further enhancing overall health management and reducing healthcare costs [4].

Existing manual methods for diagnosing sleep disorders often involve subjective assessments and are time-consuming [5]. Traditional approaches typically rely on self-reported symptoms, sleep diaries, and polysomnography (PSG), which can be both costly and cumbersome [6]. These methods often suffer from inconsistencies and inaccuracies due to the variability in individual reporting and the limitations of manual data interpretation, which can hinder the effectiveness of diagnosis and treatment. The advent of artificial intelligence (AI) [7]and machine learning (ML) [8] has the potential to revolutionize sleep disorder diagnostics by automating and refining the classification process. AI-driven methods can analyse large volumes of data quickly and with high precision, enabling more accurate and consistent diagnosis of sleep disorders [9]. By leveraging advanced algorithms and computational models, AI can provide insights that surpass traditional methods, offering more personalized and effective treatment options.

However, existing AI and ML approaches to sleep disorder classification face several challenges. Many models are limited by the quality and size of available datasets, leading to potential biases and inaccuracies in predictions. Additionally, the complexity of sleep data and the variability in individual sleep patterns can complicate the development of robust models. Issues such as overfitting, underfitting, and the need for extensive computational resources further impact the effectiveness of AI solutions. Addressing these challenges requires ongoing research and refinement of algorithms to enhance their reliability and applicability in real-world scenarios. The novel contributions of this work as follows

- Development of a scalable SDC framework capable of integrating new data seamlessly, ensuring the model remains up-to-date and accurate over time.
- Utilization of IEVA to handle large datasets and enable continuous model updates without retraining, enhancing feature relevance and reducing dimensionality.
- Application of RDT Classification algorithm for its interpretability and ability to manage complex decision boundaries, leading to accurate classification of sleep disorders.

The rest of the paper is organized as follows: Section 2 presents a survey of related work in the field of sleep disorder classification. Section 3 details the proposed method, while Section 4 discusses the results. Finally, Section 5 provides the conclusion and future directions.

2. Literature Survey

In [10], the authors proposed a method for sleep disorder identification using wavelet scattering on electrocardiogram (ECG) signals. They employed wavelet transformation to extract meaningful features from ECG data, specifically targeting sleep disorders like apnea. While this method captures temporal features effectively, it is limited by its sensitivity to noise in ECG signals, which can reduce accuracy in real-world clinical settings. In [11], the study applied multi-class classification methods to automate sleep disorder prediction. By utilizing machine learning classifiers, the method distinguishes between various sleep disorders. However, its reliance on generalized classifiers without specialized feature extraction techniques can result in lower diagnostic precision, particularly for borderline cases. In [12], the authors developed an automated, explainable wavelet-based sleep scoring system for patients suspected of insomnia, apnea, and periodic leg movement (PLM). Wavelet transformations provided detailed time-frequency analysis, aiding in sleep scoring. Yet, the method struggles with computational efficiency, especially when processing large datasets in real time.

In [13], the researchers employed machine learning algorithms, such as support vector machines (SVM) and decision trees, for the classification of sleep disorders. Although these traditional classifiers offer straightforward implementation, they face challenges with overfitting and scaling when exposed to more complex datasets with high dimensionality. In [14], a deep feature-based metabolism syndrome prediction system was proposed to identify sleep disorder diseases. This intelligent system integrated feature extraction with deep learning for predictive analysis. However, deep learning models often require large datasets and significant computational power, making them less feasible in environments with limited resources. In [15], the authors utilized the Random Forest Classifier (RFC) algorithm for the detection and classification of sleep disorders. While Random Forests offer high accuracy and robustness in handling mixed data types, they are computationally intensive and can be prone to overfitting in cases with an excessive number of decision trees or when applied to small datasets. In [16], RFC was applied to a health and lifestyle dataset to classify sleep disorders. This method performed well in handling non-linear relationships and complex feature interactions. However, it tends to lack interpretability and transparency, which may be a drawback in clinical applications where explainability is critical.

In [17], the study focused on predicting obstructive sleep apnea (OSA) in patients with temporomandibular disorder (TMD) using multidata and machine learning. Although the method combines various datasets for better prediction accuracy, integrating diverse data sources poses a challenge due to data imbalance and inconsistencies, which can hinder model performance. In [18], machine learning techniques were employed to analyze the prevalence of depression in elderly patients with OSA. This approach effectively identified correlations between depression and sleep disorders, but the model struggled with generalization due to limited sample diversity, reducing its applicability across different populations. In [19], Deep Learning Neural Network (DLNN) was applied to actigraphy data for the prediction of sleep disorders. While the deep learning models showed strong

predictive capabilities, they faced limitations due to the sparse and noisy nature of actigraphy data, requiring extensive preprocessing to avoid data distortion.

In [20], the study aimed to improve sleep apnea diagnoses using machine learning based on the STOP-BANG questionnaire. This method provided an efficient screening tool, but the questionnaire-based approach may introduce subjective biases and rely on patient self-reporting, which can lead to misclassification. In [21], the authors applied machine learning techniques to diagnose obstructive sleep apnea/hypopnea syndrome (OSAHS). The approach yielded accurate results, yet it was limited by the availability of high-quality, annotated datasets, and the performance of the model could degrade when faced with insufficient training data. In [22], a novel hybrid feature reduction method using the MCMST-Clustering algorithm was introduced for sleep disorder diagnosis. This technique reduced feature dimensionality while maintaining critical data, improving classification efficiency. However, its complexity and reliance on clustering can make the method less adaptable to dynamically evolving datasets. In [23], the researchers developed a machine learning model to predict the risk of depression in adults with OSAHS. Although the model provided valuable insights into comorbid conditions, it suffered from limitations related to feature interpretability, making it difficult to translate predictions into actionable clinical insights. In [24], the authors used ResNet-50 and Gradient Boosting to detect sleep apnea and rapid eye movement (REM) stages. While the combination of convolutional neural networks (CNNs) and boosting techniques improved accuracy, the model's complexity increased computational costs, making it unsuitable for real-time or large-scale applications. In [25], the detection of sleep apnea was performed using an ECG spectrogram and a novel machine-learning framework based on bag-of-features. This approach was effective in handling ECG data, yet it faced challenges in accurately predicting apnea events during continuous positive airway pressure (CPAP) titration due to the variability in signal patterns during treatment.

3. Proposed Methodology

The proposed SDC framework introduces a novel combination of IEVA and RDT for the classification of sleep disorders, specifically insomnia and sleep apnea. Figure 1 shows the proposed SDC system model. This approach, not presented in existing surveys, overcomes the limitations of traditional methods by integrating dynamic feature reduction with a robust classification mechanism. IEVA addresses the challenge of high-dimensional data by incrementally reducing features while retaining essential information, thus improving computational efficiency and accuracy. Meanwhile, RDT provides a flexible and efficient classification framework that adapts well to complex, non-linear relationships in the data. This innovative combination not only enhances diagnostic precision but also mitigates the problems associated with conventional methods, such as overfitting and inadequate generalization.

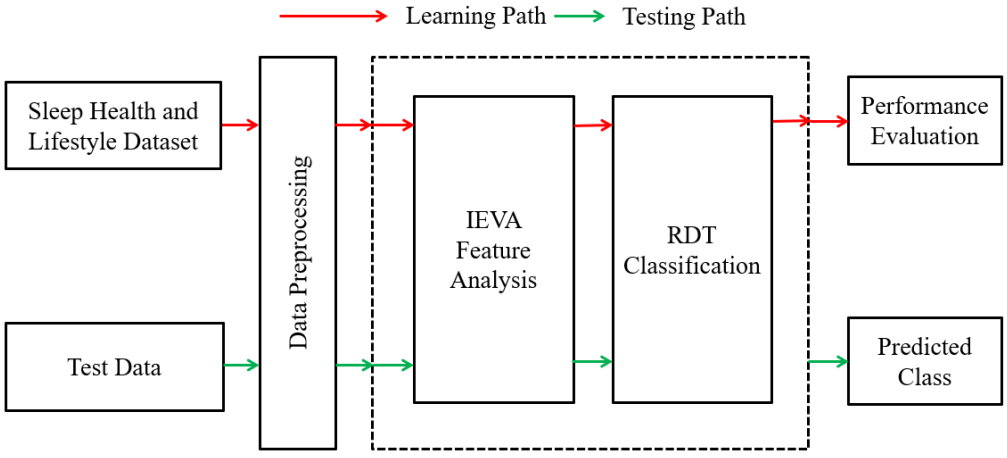

Figure 1: Proposed SDC system model.

The first step in our methodology is data preprocessing, which involves cleaning and transforming raw data into a format suitable for analysis. This process begins with handling missing values by either imputing them with statistical measures like mean or median or removing them altogether if they are deemed too significant. Standardization is then applied to numeric features, which scales data to a uniform range, preventing variables

with larger scales from dominating those with smaller scales. This step is crucial as it helps in maintaining the integrity of the data and avoids biasing the subsequent analysis. Following data preprocessing, IEVA is employed to reduce the dimensionality of the dataset. The IEVA identifies patterns and correlations among variables, condensing them into a smaller set of principal components that retain the most relevant information. The incremental approach to Principal Component Analysis (PCA) is advantageous for handling large datasets that do not fit into memory all at once. By processing data in smaller batches, this method incrementally computes principal components, making it computationally feasible and efficient. Dimensionality reduction through IEVA simplifies the dataset while preserving its essential features, which enhances the performance of subsequent classification algorithms. This step is critical in our methodology as it facilitates the extraction of meaningful insights from complex lifestyle data related to sleep patterns. The final stage of our methodology involves RDT classification. The decision trees are intuitive models that segment data into hierarchical structures based on a series of yes-no questions. Random forests extend this concept by aggregating multiple decision trees, each trained on different subsets of the data and features. The RDT are employed to classify individuals into categories of normal sleep, insomnia, or sleep Apnea based on their lifestyle-related data. This ensemble method improves robustness and accuracy by reducing overfitting and capturing diverse patterns present in the dataset.

3.1 IEVA feature extraction

The IEVA is an extension of traditional PCA that processes data in a batch-by-batch manner, making it suitable for large datasets or streaming data. Table 1 presents the proposed IEVA algorithm steps for feature extraction. Figure 2 presents the proposed IEVA architecture. The goal is to reduce the dimensionality of the data while preserving as much variance as possible. The IEVA operation starts with equation (1), which represents the data matrix (A), where (m) is the number of samples and (n) is the number of features. Each element a_{ij} in the matrix denotes the value of the (j)-th feature for the (i)-th sample. The data matrix is constructed by stacking each sample a_i as a row vector in the matrix. For example, in the context of sleep disorder classification, each row vector a_i might represent an individual's lifestyle statistics, including age, BMI, exercise frequency, diet, and sleep duration. Organizing the data in this matrix form allows for efficient mathematical manipulation and analysis.

$$A_{m*n} = \begin{matrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ a_{m1} & \dots & a_{mn} \end{matrix} = a_1, a_2, \dots, a_m \quad (1)$$

Equation (2) calculates the mean vector $\bar{A}$ of the dataset, where $\bar{A}$ is an (n)-dimensional vector. Each element of $\bar{A}$ is computed by taking the average of the corresponding feature across all (m) samples. For instance, if the dataset includes sleep duration as one of the features, the corresponding element in $\bar{A}$ would represent the average sleep duration across all individuals in the dataset. Mean Centering the data is a critical preprocessing step in IEVA, as it ensures that the principal components reflect the directions of maximum variance around the mean of the dataset.

$$\bar{A} = \frac{1}{m}\sum_i^m A_i \quad (2)$$

According to Equation (3), the standard deviation vector (S) is computed to measure the dispersion of the dataset. For each feature, (S) calculates the square root of the average squared difference between each sample and the mean. This normalization process ensures that all features contribute equally to the analysis, preventing features with larger numerical ranges from dominating the results. For example, if the dataset includes both age (which might range from 18 to 90) and sleep duration (which might range from 4 to 10 hours), standardizing these features ensures that the IEVA does not give undue weight to the age variable simply because of its larger range of values.

$$S = \sqrt{\frac{1}{m}\sum_{i=1}^m (A_i - \bar{A})^2} \quad (3)$$

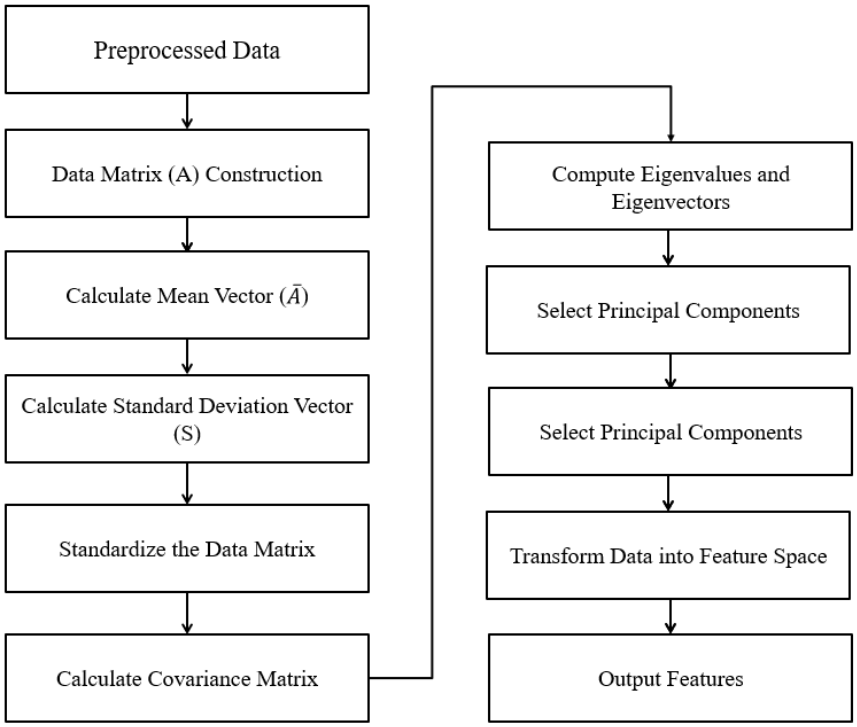

Figure 2: Proposed IEVA architecture.

Table 1. Proposed IEVA algorithm.

Input: Preprocessed dataset **Output:** IEVA Features
Step 1: Input Data Matrix (A) Construction: Given a dataset with m samples and n features. Construct the data matrix A where $A \in R^{m\times n}$. Each element A_{ij} represents the value of the j-th feature for the i-th sample. **Step 2: Calculate Mean Vector ($\bar{A}$):** Compute the mean vector $A \in R^{m\times n}$. Each element of $\bar{A}$ is the average of the corresponding feature across all samples. **Step 3: Calculate Standard Deviation Vector (S):** Compute the standard deviation vector $S \in R^{n}$ for each feature. **Step 4: Standardize the Data Matrix:** Normalize each feature to have zero mean and unit variance. Standardized matrix $A' = \frac{A-\bar{A}}{S}$. **Step 5: Calculate Covariance Matrix:** Compute the covariance matrix Σ from the standardized data matrix $\bar{A}$. **Step 6: Compute Eigenvalues and Eigenvectors:** Perform eigenvalue decomposition on the covariance matrix Σ. Find eigenvalues λ and corresponding eigenvectors. **Step 7: Select Principal Components:** Order the eigenvalues λ in descending order. Select the top k eigenvalues and their corresponding eigenvectors. Calculate the proportion of variance captured by the first k principal components. **Step 8: Transform Data into Feature Space:** Project the original standardized data onto the new feature space defined by the top k principal components. For each sample A_i, compute the transformed feature vector Y_i.

Equation (4) calculates the covariance between two features (A) and (B). Covariance measures the degree to which two variables change together. A positive covariance indicates that the two features tend to increase or decrease together, while a negative covariance indicates that one feature tends to increase when the other decreases. In IEVA, the covariance matrix is crucial as it encapsulates the relationships between all pairs of features in the dataset. This matrix is then used to identify the principal components—directions in which the data varies the most. For example, in a dataset of lifestyle statistics, the covariance between physical activity and sleep duration might reveal how these two factors are related.

$$Cov(A,B) = \frac{\sum_{i=1}^{m}(A_i-\bar{A})(B_i-\bar{B})}{m} \quad (4)$$

Equation (5) quantifies the proportion of the total variance captured by the first (k) principal components out of the total (p) principal components. Here, λ_i represents the eigenvalues of the covariance matrix, which correspond to the amount of variance explained by each principal component. By summing the largest eigenvalues and dividing by the sum of all eigenvalues, we can determine how much of the total variance is retained by the top (k) components. This helps in selecting an appropriate number of principal components that capture a significant portion of the dataset's variability, thereby reducing the dimensionality of the data while preserving most of its information.

$$\frac{\lambda_1+\lambda_2+\cdots\lambda_k}{\lambda_1+\lambda_2+\cdots\lambda_p} \quad (5)$$

Equation (6) represents the transformation of the original data into the new feature space defined by the principal components. Here, Y_i is the projection of the (i)-th sample onto the principal components, adjusted by the inverse of the standard deviation. The term Y_i centers the data, while S^{-1} normalizes it. This transformation ensures that the new features are scaled appropriately, making them comparable and interpretable. For example, in the context of sleep disorder classification, this transformed data can be used to train machine learning models that leverage the most significant patterns in the lifestyle statistics to predict sleep disorders.

$$Y_i = \sqrt{((Y_i - \bar{Y}S^{-1})(Y_i - \bar{Y}))} \quad (6)$$

3.2 RDT classification

The RDT are an ensemble method, part of the broader category of random forests and decision trees. They improve classification accuracy by reducing overfitting and variance. Table 2 presents the proposed RDT classification algorithm steps for feature extraction. Figure 3 presents the proposed RDT architecture. The IEVA operation starts with Equation (7), the Gini impurity Gi_m measures the impurity of a node m in a decision tree. Here, P_c represents the proportion of samples belonging to class ccc at node m. The Gini impurity quantifies the probability of incorrectly classifying a randomly chosen element if it was randomly labelled according to the distribution of labels in the node. The second part of the equation, $1 - \sum_{c=1}^{c} P_c^2$, simplifies the calculation by directly using the squared probabilities of class proportions. This impurity measure helps in deciding the optimal split at each node of the tree by choosing the split that reduces impurity the most.

$$Gi_m = \sum_{c=1}^{c} P_c(1 - P_c) = 1 - \sum_{c=1}^{c} P_c^2 \quad (7)$$

Equation (8) calculates the Gini gain, Vm_{JM}^{GINI}, at node M. It is the difference between the Gini impurity of the parent node Gi_M and the sum of the Gini impurities of the left child node Gi_L and the right child node Gi_R. The Gini gain is used to select the best split for a node. A higher Gini gain indicates a better split because it means a greater reduction in impurity, thus leading to more homogeneous child nodes.

$$Vm_{JM}^{GINI} = Gi_M - Gi_L - Gi_R \quad (8)$$

Equation (9) defines the set y, which consists of predictions from m different decision trees $Y_i(X)$ in the ensemble. Here, X represents the input features. Each $Y_i(X)$ is a decision tree trained on a subset of the data, possibly with a subset of the features. This collection of predictions will be used to make the final classification decision through majority voting or averaging.

$$y = \{Y_1(X), Y_2(X), \ldots, Y_m(X)\} \quad (9)$$

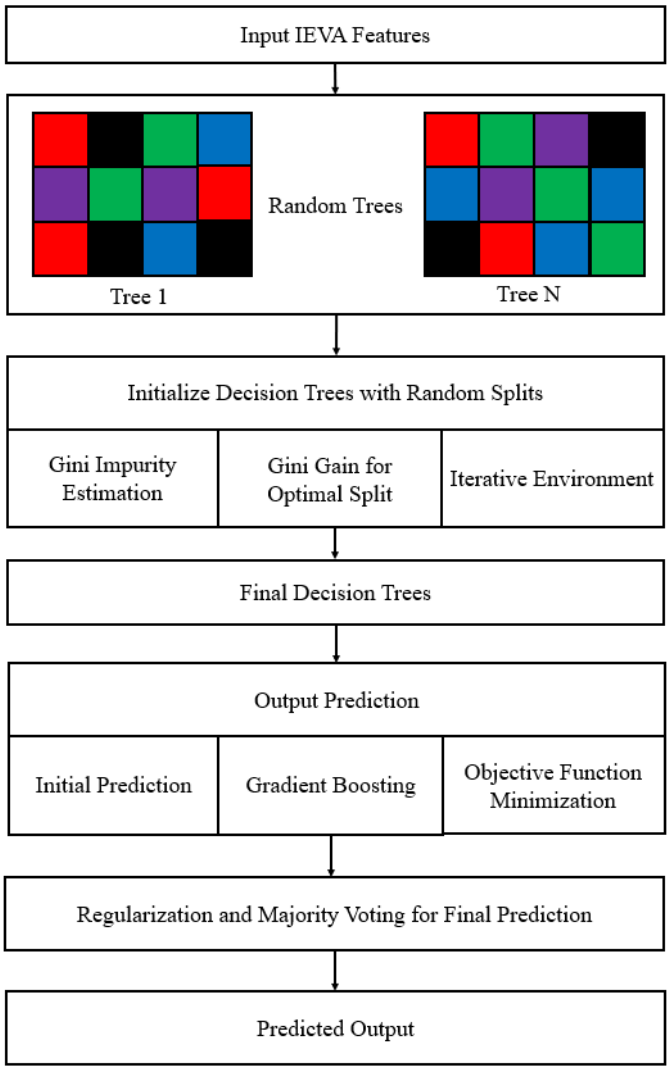

Figure 3: Proposed RDT architecture.

In this equation (10), $t(X)$ represents the final predicted class for input X. The $ArgMax$ function selects the class A that has the highest sum of $Z(Y_I(X))$, where Z is an indicator function those outputs 1 if the predicted class matches A and 0 otherwise. Essentially, this equation performs majority voting, where the class that appears most frequently among the m tree predictions is chosen as the final output.

$$t(X) = ArgMax \sum_{I-1}^{m} Z(Y_I(X)) = A \tag{10}$$

Equation (11) describes an update step in gradient boosting, a technique often used to enhance the performance of decision tree ensembles. Here, $\hat{x}^n$ represents the prediction at iteration n, $\hat{x}^{n-1}$ is the prediction from the previous iteration, β is the learning rate, and $g_n(Y, \phi_n)$ is the gradient at iteration n. This update rule incrementally improves the model by adding the gradient of the loss function with respect to the model parameters, scaled by the learning rate.

$$\hat{x}^n = \hat{x}^{n-1} + \beta\, g_n(Y, \phi_n) = \beta \sum_{i=1}^{n} g_i(Y, \phi_i) \tag{11}$$

Equation (12) defines the objective function M to be minimized during training. It sums the loss m over all samples j, where $\hat{y}_j$ is the predicted output, and x_j is the actual input feature. The objective function evaluates how well the model predictions match the actual data and guides the optimization process. The equation incorporates the update rule from gradient boosting, showing how each prediction is iteratively refined.

$$M = \sum_j m(\hat{y}_j, x_j) = \sum_j m[\hat{x}_j^{\,n-1} + \beta g_i(Y_j, \phi_i), x_j] \tag{12}$$

Equation (13) expands the objective function to include a regularization term ϕ_i). The regularization term penalizes the complexity of the model, helping to prevent overfitting. Here, ϕ_i represents the parameters of the model, and $\phi_i$is a regularization function (such as L1 or L2 regularization). The first part of the equation sums the loss over all samples, while the second part adds the regularization penalty.

$$M = \sum_j m\ \left(\hat{x}_j . x_j\right) + \sum_i \varphi\left(\phi_i\right) = \sum_j m\left[\hat{x}_j^{\,i-1} + \beta g_i\left(Y_j, \phi_i\right), x_j\right] + \sum_i \varphi(\phi_i) \tag{13}$$

Equation (14) defines the regularization term ϕ_i in more detail. It consists of two components: $\delta\ S_i$ and $\frac{1}{2} v\|u_i\|$. The term $\delta\ S_i$ represents a penalty proportional to the size of the model (or the number of parameters), where δ is a hyperparameter controlling the strength of this penalty. The second term, $\frac{1}{2} v\|u_i\|$, is a quadratic penalty on the magnitude of the parameters, with v controlling the penalty's strength. This regularization helps keep the model parameters small, reducing the risk of overfitting and improving generalization.

$$\varphi(\phi_i) = \delta\ S_i + \frac{1}{2} v\|u_i\| = \delta\ \ S_i + \frac{1}{2} v \sum_{l=1}^{S_i}\left[u_l^{(i)}\right]^2 \tag{14}$$

Table 2. Proposed RDT algorithm.

Input: IEVA Features **Output:** Predicted Output
Step 1: Random Trees: Consider input IEVA features, then apply them to multiple decision trees over random space. **Step 2: Gini Impurity:** Compute the Gini impurity Gi_m for each node m using Equation (7). P_c is the proportion of samples belonging to class ccc at node m. **Step 3: Gini Gain for Optimal Split:** For each potential split at node M, calculate the Gini gain Vm_{JM}^{GINI} using Equation (8). Gi_m is the Gini impurity of the parent node. Gi_L and Gi_R are the Gini impurities of the left and right child nodes, respectively. **Step 3: Final Decision Trees:** Train multiple decision trees (m) using different subsets of the data and features. Each decision tree $Y_I(X)$ is trained independently as per Equation (9). **Step 4: Initial Prediction:** For a given input X, collect predictions from all m decision trees. Use majority voting to determine the final predicted class $t(X)$ as per Equation (10). The Z is an indicator function those outputs 1 if the predicted class matches A and 0 otherwise. **Step 5: Gradient Boosting:** Initialize the prediction $\hat{x}^{(0)}$. For each iteration n, update the prediction using Equation (11). β is the learning rate and g_n is the gradient at iteration n. **Step 6: Objective Function Minimization:** Evaluate the objective function M as per Equation (12) to minimize the error over all samples j. **Step 7: Regularization:** Incorporate regularization term ϕ_i into the objective function to penalize model complexity, as per Equation (13). **Step 8: Majority Voting for Final Prediction:** For a given input X, collect predictions from all m decision trees with prevented overfitting data generated by regularization.

4. Results and Discussion

This section evaluates the performance of multiple methods by applying them to the same Sleep Health and Lifestyle Dataset. The comparison highlights differences in predictive accuracy and efficiency, providing insights into the most effective techniques for classifying sleep disorders.

4.1 Dataset

The Sleep Health and Lifestyle Dataset consists of 40,000 rows and 13 columns, offering comprehensive information on various factors that influence sleep patterns and lifestyle habits. The dataset includes demographic details like gender and age, as well as occupational data, providing context for the subjects' daily routines. Sleep-related variables include sleep duration and quality, while physical activity is represented through daily steps and

overall activity levels. Health metrics such as BMI category, blood pressure, and heart rate further enrich the dataset, capturing the physiological aspects linked to sleep. Stress levels are included to account for psychological factors affecting sleep quality. The dataset is divided into an 80% training set (32,000 rows) and a 20% testing set (8,000 rows) to develop and validate machine learning models. This division allows for robust training of the model on a wide range of lifestyle and health variables, ensuring that the model can generalize well and accurately predict the presence or absence of sleep disorders.

4.2 Data analytics

Figure 4 presents a sample dataset focused on sleep health and lifestyle factors, containing key variables for everyone, identified by a unique Person ID. The dataset includes demographic information such as gender and age, along with the individual's occupation. Sleep-related data, such as sleep duration in hours and a subjective quality of sleep rating on a scale of 1 to 10, are captured to assess sleep patterns. Physical activity is measured by the number of minutes per day spent exercising, and stress levels are self-reported on a 1 to 10 scale. The dataset also incorporates health metrics like BMI category, blood pressure (systolic/diastolic), and resting heart rate (bpm), providing insight into each person's overall physical condition. Additionally, daily steps are tracked as a measure of activity level. The presence or absence of a sleep disorder, categorized as none, insomnia, or sleep apnea, is also recorded, making this dataset valuable for analyzing the impact of various lifestyle and health factors on sleep disorders.

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
...	...	...	...	...	...	...	...	...	...	...	...	...
369	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
370	Female	59	Nurse	8.0	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
371	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
372	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
373	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea

Figure 4. Sample dataset.

	Age	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	Heart Rate	Daily Steps	High_pressure	Low_pressure	Gender_encoded	Occupation_encoded	BMI Category_encoded
0	27	6.1	6	42	6	77	4200	126	83	1	7	1
1	28	6.2	6	60	8	75	10000	125	80	1	1	0
2	28	6.2	6	60	8	75	10000	125	80	1	1	0
3	28	5.9	4	30	8	85	3000	140	90	1	7	1
4	28	5.9	4	30	8	85	3000	140	90	1	7	1
...	...	...	...	...	...	...	...	...	...	...	...	...
369	59	8.1	9	75	3	68	7000	140	95	0	4	1
370	59	8.0	9	75	3	68	7000	140	95	0	4	1
371	59	8.1	9	75	3	68	7000	140	95	0	4	1
372	59	8.1	9	75	3	68	7000	140	95	0	4	1
373	59	8.1	9	75	3	68	7000	140	95	0	4	1

Figure 5. Sample dataset after preprocessing.

Figure 5 illustrates the dataset after preprocessing, where label encoding and scaling techniques have been applied. Categorical variables like gender, occupation, and sleep disorder are converted into numerical labels for machine learning compatibility. Continuous variables such as sleep duration, physical activity level, and heart rate are scaled to ensure uniform data distribution, enhancing model accuracy and performance. Figure 6 presents the IEVA-based correlation matrix, which visualizes the relationships between different variables in the dataset. Each

cell in the matrix represents the strength and direction of the correlation between two features, with positive or negative values indicating the nature of the relationship. High correlations suggest strong linear dependencies, while low correlations indicate weaker or no associations. This matrix helps identify influential factors contributing to sleep disorders, enabling more effective feature selection for predictive modelling.

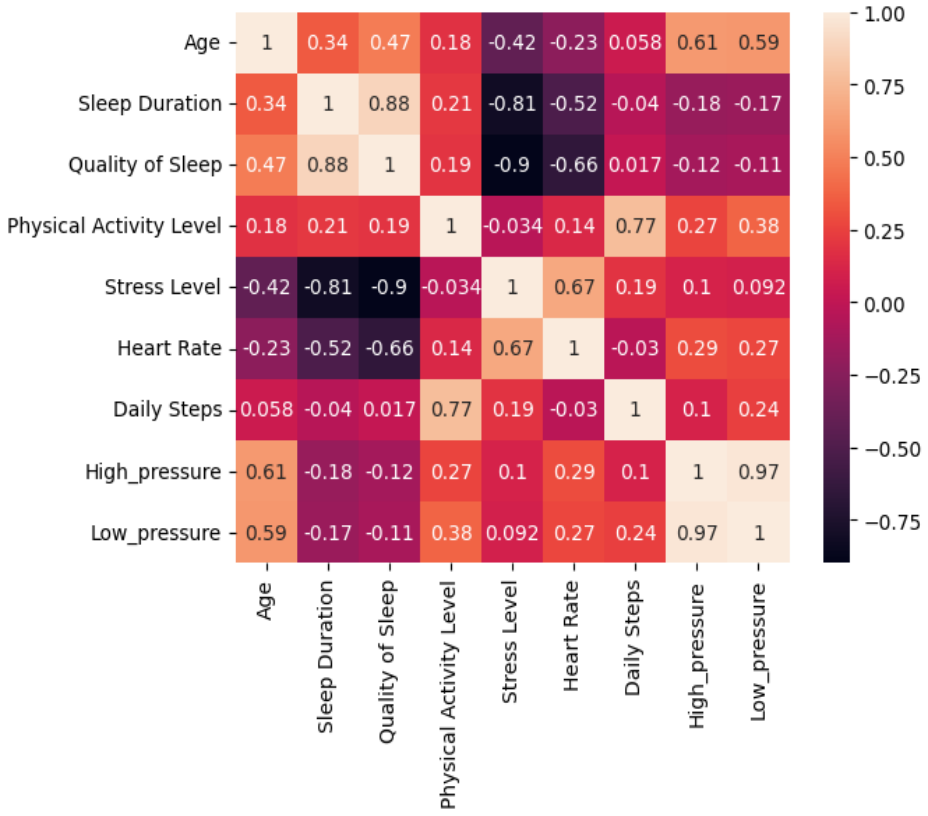

Figure 6. IEVA based correlation matrix.

4.2 Performance Evaluation

Table 3 compares the performance of various SDC methods, such as SVM [13], DLNN [19], RFC [15], and Proposed SDC-RDT. For accuracy, the proposed SDC-RDT (99.817%) shows a 2.93% improvement over SVM (96.976%) [13], reflecting a substantial gain in prediction accuracy. Compared to DLNN (97.781%) [19], the improvement is 2.08%, highlighting the SDC-RDT method's ability to handle more complex patterns effectively. Against RFC (98.397%) [15], SDC-RDT still improves by 1.44%, indicating its efficiency over even the high-performing RFC method. For precision, the proposed SDC-RDT (99.557%) outperforms SVM (96.334%) [13] with a 3.34% improvement, indicating that SDC-RDT has higher precision in correctly identifying positive cases. The DLNN (97.739%) [19] shows a 1.86% lower precision, demonstrating the superior feature extraction and classification process of SDC-RDT. For RFC (98.370%) [15], the improvement is 1.21%, which still indicates an edge in terms of accuracy in prediction.

For recall, the proposed SDC-RDT (99.068%) has a 2.35% improvement over SVM (96.798%) [13], showcasing its strength in identifying true positive cases more effectively. The SDC-RDT improves by 1.78% over DLNN (97.333%) [19], reflecting better generalization to unseen data. Compared to RFC (98.733%) [15], the gain is smaller at 0.34% but still indicates an advantage in recall. For F1-Score, the proposed SDC-RDT (99.504%) shows a 2.66% improvement over SVM (96.927%) [13], a substantial improvement indicating better overall performance. The DLNN (97.657%) [19] is outperformed by 1.89%, indicating a more balanced performance between precision and recall. Against RFC (98.553%) [15], the proposed SDC-RDT offers a 0.97% improvement, still leading in the overall effectiveness of the classification.

Table 4 presents an ablation study of the proposed SDC-RDT, comparing its performance when certain components are removed. The results indicate that preprocessing plays a critical role in the proposed SDC-RDT's

performance. When preprocessing is omitted, the accuracy drops to 97.214%, a significant decrease from the full methodology's 99.817%. Similarly, precision without preprocessing is 97.369%, compared to 99.557% with the complete methodology. Recall also decreases to 97.563% without preprocessing, compared to 99.068% with preprocessing. The F1-Score falls to 97.002% without preprocessing, highlighting the importance of this step, as the full methodology achieves 99.504%. These reductions underscore that preprocessing is crucial for enhancing the model's performance by ensuring clean and standardized data. The absence of IEVA feature extraction also impacts performance, though not as drastically as the absence of preprocessing. Without IEVA, the accuracy is 98.340%, which is still lower than the complete methodology's 99.817%. Precision decreases to 98.653% without IEVA, compared to 99.557% with it. Recall drops slightly to 98.600% from 99.068%, and the F1-Score is 98.495% without IEVA, versus 99.504% with it. These results demonstrate that IEVA is essential for reducing dimensionality while preserving the data's variance, thus improving the model's efficiency and predictive power.

Table 3. Performance comparison of various SDC classification methods.

Metric	SVM [13]	DLNN [19]	RFC [15]	Proposed SDC-RDT
Accuracy	96.976	97.781	98.397	99.817
Precision	96.334	97.739	98.370	99.557
Recall	96.798	97.333	98.733	99.068
F1-Score	96.927	97.657	98.553	99.504

Table 4. Ablation study of proposed SDC-RDT.

Metric	Proposed SDC-RDT without preprocessing	Proposed SDC-RDT without IEVA feature extraction	Proposed SDC-RDT
Accuracy	97.214	98.340	99.817
Precision	97.369	98.653	99.557
Recall	97.563	98.600	99.068
F1-Score	97.002	98.495	99.504

5. Conclusion

In conclusion, the proposed SDC framework, which integrates IPCA with RDT, offers a significant advancement in the classification of sleep disorders like insomnia and sleep apnea. By addressing the limitations of traditional methods, such as inefficient feature reduction and suboptimal classification accuracy, this approach demonstrates enhanced performance in handling complex, high-dimensional datasets. The dynamic feature extraction through IPCA, combined with the adaptability of RDT, ensures more accurate diagnoses and better generalization across varied patient profiles. Future scope includes extending this framework to other sleep-related conditions, refining the model with larger and more diverse datasets, and exploring hybrid models that combine IPCA-RDT with deep learning techniques to further enhance predictive accuracy and scalability in clinical applications. Additionally, real-time monitoring and diagnosis through wearable devices can be integrated into this model to facilitate continuous patient evaluation and timely interventions.

References

[1] Sarber, K. M., & Patil, R. D. (2024). Comorbid Insomnia and Sleep Apnea: challenges and treatments. *Otolaryngologic Clinics of North America*, *57*(3), 385-393.

[2] Khazaie, H., Aghazadeh, M., Zakiei, A., Maazinezhad, S., Tavallaie, A., Moghbel, B., ... & Sharafkhaneh, A. (2024). Co-morbid Insomnia and Sleep Apnea (COMISA) in a large sample of Iranian: prevalence and associations in a sleep clinic population. *Sleep and Breathing*, 1-8.

[3] Ma, Yan, Janet M. Mullington, Peter M. Wayne, and Gloria Y. Yeh. "Heart rate variability during sleep onset in patients with insomnia with or without comorbid sleep apnea." *Sleep Medicine* 122 (2024): 92-98.

[4] Hein, Matthieu, Benjamin Wacquier, Matteo Conenna, Jean-Pol Lanquart, and Camille Point. "Risk of Comorbid Insomnia Disorder Associated with Major Depression in Apneic Patients: A Cross-Sectional Study." *Clocks & Sleep* 6, no. 3 (2024): 389-401.

[5] San, Luis, and Belén Arranz. "The night and day challenge of sleep disorders and insomnia: A narrative review." *Actas espanolas de psiquiatria* 52, no. 1 (2024): 45.

[6] Park, Jung-A., Jee-Eun Yoon, Xiaoyue Liu, Yoonhee Chang, Giuseppe Maiolino, Martino F. Pengo, Genmin Lin, and Younghoon Kwon. "Cardiovascular Implications of Sleep Disorders Beyond Sleep Apnea." *Current Sleep Medicine Reports* (2024): 1-9.

[7] Ko, Jisu, Jae Hyeok Lim, Dan Bi Kim, Min Jeong Joo, Yun Seo Jang, Eun-Cheol Park, and Jaeyong Shin. "Association between alcohol use disorder and risk of obstructive sleep apnea." *Journal of Sleep Research* 33, no. 4 (2024): e14128.

[8] Wang, Will Ke, Jiamu Yang, Leeor Hershkovich, Hayoung Jeong, Bill Chen, Karnika Singh, Ali R. Roghanizad, Md Mobashir Hasan Shandhi, Andrew R. Spector, and Jessilyn Dunn. "Addressing wearable sleep tracking inequity: a new dataset and novel methods for a population with sleep disorders." *Proceedings of Machine Learning Research* 248 (2024): 380-396.

[9] Burch, James B., Alexandria F. Delage, Hongmei Zhang, Alexander C. McLain, Meredith A. Ray, Austin Miller, Swann A. Adams, and James R. Hebert. "Sleep disorders and cancer incidence: examining duration and severity of diagnosis among veterans." *Frontiers in Oncology* 14 (2024): 1336487.

[10] Sharma, Manish, Harsh Lodhi, Rishita Yadav, and U. Rajendra Acharya. "Sleep disorder identification using wavelet scattering on ECG signals." *International Journal of Imaging Systems and Technology* 34, no. 1 (2024): e22980.

[11] Dritsas, E., & Trigka, M. (2024). Utilizing Multi-Class Classification Methods for Automated Sleep Disorder Prediction. *Information*, *15*(8), 426.

[12] Ingle, Manisha, Manish Sharma, Shresth Verma, Nishant Sharma, Ankit Bhurane, and U. Rajendra Acharya. "Automated explainable wavelet-based sleep scoring system for a population suspected with insomnia, apnea and periodic leg movement." *Medical Engineering & Physics* 130 (2024): 104208.

[13] T. S. Alshammari, "Applying Machine Learning Algorithms for the Classification of Sleep Disorders," in *IEEE Access*, vol. 12, pp. 36110-36121, 2024, doi: 10.1109/ACCESS.2024.3374408.

[14] Anisha, P. R., C. Kishor Kumar Reddy, Marlia M. Hanafiah, Bhamidipati Ramana Murthy, R. Madana Mohana, and Y. V. S. S. Pragathi. "An intelligent deep feature based metabolism syndrome prediction system for sleep disorder diseases." *Multimedia Tools and Applications* 83, no. 17 (2024): 51267-51290.

[15] Tareq, Wadhah Zeyad Tareq. "Sleep Disorders Detection and Classification Using Random Forests Algorithm." In *Decision Making in Healthcare Systems*, pp. 257-266. Cham: Springer International Publishing, 2024.

[16] Widyastuty, Wiwiek, and Mochamamd Abdul Azis. "Classification and Evaluation of Sleep Disorders Using Random Forest Algorithm in Health and Lifestyle Dataset." *Compiler* 13, no. 1 (2024): 11-18.

[17] Lee, Yeon-Hee, Seonggwang Jeon, Q-Schick Auh, and Eun-Jae Chung. "Automatic prediction of obstructive sleep apnea in patients with temporomandibular disorder based on multidata and machine learning." *Scientific Reports* 14, no. 1 (2024): 19362.

[18] Qin, S., Zheng, Z., Li, R., Wu, C., & Wang, W. (2024). Analyzing the Prevalence of Depression and Its Influencing Factors in Elderly Patients With Obstructive Sleep Apnea: A Machine Learning Approach. *Ear, Nose & Throat Journal*, 01455613241271632.

[19] Kim, Kyoungmin, Jeongho Park, Soonhyun Yook, Ho Sung Kim, and Eun Yeon Joo. "Prediction of Sleep Disorder From Actigraphy Data Using Deep Learning." Journal of Sleep Medicine 21, no. 2 (2024): 73-79.

[20] Choi, M.S., Han, D.H., Choi, J.W. and Kang, M.S., 2024. A Study on Improving Sleep Apnea Diagnoses Using Machine Learning Based on the STOP-BANG Questionnaire. *Applied Sciences*, *14*(7), p.3117.

[21] Bedoya, Oscar, Santiago Rodríguez, Jenny Patricia Muñoz, and Jared Agudelo. "Application of Machine Learning Techniques for the Diagnosis of Obstructive Sleep Apnea/Hypopnea Syndrome." *Life* 14, no. 5 (2024): 587.

[22] Şenol, A., Talan, T., & Aktürk, C. (2024). A new hybrid feature reduction method by using MCMSTClustering algorithm with various feature projection methods: a case study on sleep disorder diagnosis. *Signal, Image and Video Processing*, *18*(5), 4589-4603.

[23] Li, Enguang, Fangzhu Ai, and Chunguang Liang. "A machine learning model to predict the risk of depression in US adults with obstructive sleep apnea hypopnea syndrome: a cross-sectional study." *Frontiers in Public Health* 11 (2024): 1348803.

[24] Varshini, Ganti Venkata, V. Sakthivel, P. Prakash, and Jae Woo Lee. "Sleep Apnea and Rapid Eye Movement Detection using ResNet-50 and Gradient Boost." *International Journal of Advanced Computer Science & Applications* 15, no. 6 (2024).

[25] Linh, Tran Thanh Duy, Nguyen Thi Hoang Trang, Shang-Yang Lin, Dean Wu, Wen-Te Liu, and Chaur-Jong Hu. "Detection of preceding sleep apnea using ECG spectrogram during CPAP titration night: A novel machine-learning and bag-of-features framework." *Journal of Sleep Research* 33, no. 3 (2024): e13991.