Exploring the Effectiveness of PCA in Predictive Analysis of Sleep Disorders

Reema Abdallah[1], Sarah Alshumayri[2], Leen Sharab[3], Prof. Passent M. ElKafrawy[4]

[1,2,3,4]Computer Science, Effat University, Jeddah, 22230, Makkah, Saudi Arabia.

Contributing authors: reoabdallah@effat.edu.sa; Samalshumayri@effat.edu.sa; leosharab@effat.edu.sa; pelkafrawy@effatuniversity.edu.sa;

Abstract

This study explores the application of machine learning and feature reduction techniques, with a focus on the predictive analysis of sleep disorders using high-dimensional data. By employing Principal Component Analysis (PCA), the research aims to overcome challenges such as overfitting and computational inefficiency, which are prevalent in complex datasets. Utilizing the "Sleep Health and Lifestyle" dataset[1], the research follows a structured methodology that includes data preprocessing, visualization, PCA implementation, machine learning modeling, and evaluation to effectively predict sleep disorders. The implementation of PCA, in particular, is shown to significantly enhance the accuracy of models such as RandomForestClassifier, Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) illustrating the technique's efficacy in simplifying datasets while retaining essential information. Furthermore, the study provides a comparative analysis of various feature selection methods, highlighting the impact of dimensionality reduction on the interpretability and efficiency of machine learning models. This approach not only demonstrates the potential of PCA in improving model performance but also emphasizes the broader implications for healthcare analytics. The findings suggest that dimensionality reduction, especially through PCA, can lead to more targeted, efficient, and interpretable applications in machine learning, offering valuable insights for future research in the field.

Keywords: Feature Reduction, Sleep Disorders, Machine Learning, PCA.

1 Introduction

Sleep disturbances encompass a wide range of mental and physical difficulties that cause distress throughout the day, disruptions in sleep-wake cycles, anxiety, and a variety of other problems[2]. As a result, the primary goal of this study was to efficiently use machine learning to predict the existence of sleep disorders.

In the world of machine learning and data processing, dealing with high-dimensional data is a complex challenge. This is because complex datasets can lead to problems like overfitting and high computational costs, making it difficult to create efficient and understandable machine-learning models. To address these challenges, we explored methods like unsupervised feature selection, recursive feature elimination, t-distributed Stochastic Neighbor Embedding (t-SNE), and Principal Component Analysis (PCA) to simplify complex data features while retaining important information, with a primary focus on PCA.

The scope of our research is defined by a thorough evaluation of relevant literature, which serves as a framework for framing our work within a broader discussion on feature reduction. By using the "Sleep Health and Lifestyle" dataset [1], the main focus was on the Principal Component Analysis (PCA) technique for a realistic scenario, predicting sleep disorders.

The goal of this project is to develop a predictive model that can accurately classify individuals into having a sleep disorder or not based on their health and lifestyle attributes. We will experiment with several machine learning models to determine which provides the best performance and accuracy. The challenge lies in effectively processing and utilizing high-dimensional data, which we intend to manage using dimensionality reduction techniques such as PCA.

This study utilized Python and its data science libraries for data preprocessing, manipulation, visualization, and modeling, essential for implementing PCA for dimensionality reduction and preparing the data for machine learning analysis. This systematic process, following a structured five-phase approach, aimed to refine the dataset for analysis, visualize the data for better understanding, train machine learning models to predict sleep disorders, and evaluate the model's performance for further improvements. The experiment concludes with the implementation of PCA, notable for its capability to simplify complex datasets without losing important information, emphasizing its effectiveness in enhancing the accuracy of machine learning models. This methodological approach enables us to demonstrate the applicability of these strategies in a real-world situation while simultaneously evaluating their efficacy in improving model performance.

The structured implementation of the PCA algorithm involved transforming training features into principal components, visualizing their relationships, selecting top eigenvectors for dimensionality reduction, and evaluating the results using variance explained and feature loadings. The impact of PCA on model accuracy was assessed using RandomForestClassifier, Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) chosen for their diversity and wide application in classification tasks. The accuracy of these models was assessed before and after the application of PCA, providing insight into the impact of dimensionality reduction on model performance.

This paper will delve into the following research questions:

- How do different feature selection techniques compare in terms of their effectiveness for reducing dimensionality in machine learning models?
- What is the impact of feature selection methods on the interpretability and efficiency of machine learning models, particularly in the context of sleep disorder prediction?
- Can the integration of Principal Component Analysis (PCA) into machine learning models for predicting sleep disorders enhance the model's accuracy and efficiency?

These research questions will guide our investigation into the effectiveness and implications of various feature selection techniques, with a focus on PCA, in enhancing the accuracy and interpretability of machine learning models for predicting sleep disorders.

2 Literature Review

The art of characterizing the world of machine learning and data processing is the core idea of feature reduction and it deserves our thorough memorization. In the midst of constant change, the importance of reducing data dimensions becomes clear in two ways: an efficient way to make the model smarter, and simultaneously, its interpretability and efficiency would also improve. Therefore, the joint objective of uncovering feature relevance, as well as making data visualization and modeling computationally tractable, is the approach chosen for addressing the complex challenges of high-dimensional data, such as model overfitting and heavy computational cost. This comprehensive and detailed literature review is centered around five feature (input variable) reduction techniques: Unsupervised Feature Selection (UFS), Recursive Feature Elimination (RFE), t - Distributed Stochastic Neighbor Embedding (t-SNE), Principal Component Analysis (PCA), Feature Importance and Artificial Bee Colony (ABC). A set of questions becomes more apparent, like their abilities and limitations, enabling possibilities of new research and discovery. Taking into account the above-mentioned methods, conclude that the effectiveness of the data reduction will be achieved by the simplicity of information, which will help to create clear and open machine learning technologies.

2.1 Unsupervised Feature Selection (UFS)

The key question in functioning research on the UFS is the conduct of a detailed investigation of the UFS algorithm. The authors in Solorio-Fernández et al. classify unsupervised feature selection methods to filter, wrapper, and hybrid models according to the strategy employed in choosing features. Filter algorithms use intrinsic data properties as their basis for feature selection, thus delivering speed and scalability benefits [3]. In contrast, wrappers use outputs of some specific clustering algorithms in order to evaluate feature subsets expecting to improve clustering results which unfortunately makes computational requirements higher and require specific algorithms. Hybrid methods attempt to strike a balance, leveraging the strengths of both filter and wrapper methods to achieve both computational efficiency and effective feature selection[3].

Chen et al. suggest a new approach that classifies as a hybrid type, with a focus on the immediate generation of clustering outcomes without the requirement for a k-means clustering step after processing. We propose a novel UFSELM method, which is an unsupervised feature selection-based extreme learning machine (UFSELM). It attaches feature selection to the optimization process by applying an L2,1 norm regularization to the output weights. This method not only clarifies the validity of the clustering approach but also pushes progress by removing unnecessary features via an innovative method that adopts filter and wrapper properties simultaneously[4].

The work by Solorio-Fernández et al. and Chen et al. in general indicates progress in unsupervised feature selection that underscores the need for fast and robust approaches to complex data analysis [3, 4].

2.2 Recursive Feature Elimination (RFE)

The Recursive Feature Elimination (RFE) technique has been utilized in conjunction with machine learning algorithms to create systems aimed at the detection of chronic kidney diseases. This method involves the use of an integrated model that employs RFE to pinpoint the most significant representative features of the task at hand. Notably, the study by Sharma Yadav et al. employs four different machine learning algorithms – SVM, KNN, Decision Tree, and Random Forest – to diagnose chronic kidney diseases, resulting in promising levels of accuracy. The deployment of such proficient machine learning techniques, in collaboration with expert physicians, holds potential for widespread application in medical diagnostics [5].

2.3 t- Distributed Stochastic Neighbor Embedding (t-SNE)

According to [6], in 2008, Van der Maaten and Hinton introduced a nonlinear algorithm, t-Distributed Stochastic Neighbor Embedding, or t-SNE which became an improvement over the earlier SNE algorithm of Hinton and Roweis (2002). T-SNE is based on non-convex optimization and has become the standard for visualization in a wide range of applications as Arora et al. said. Arabnia et al. [7] introduced t-distributed Stochastic Neighbor Embedding (t-SNE) in their book as a solution for dimensionality reduction that is particularly suited for visualizing high-dimensional data. They described t-SNE as a probabilistic technique that does not rely on linear assumptions. It operates by looking at the original high-dimensional data and finding a representation in two or three dimensions that preserves the distribution of distances between points as much as possible. In this reduced dimensional space, similar data points are modeled to be close to each other, while dissimilar points are modeled to be far apart.

Arora et al. [6] stated that for t-SNE to be successful, two conditions need to be met. First, the data within each cluster should be tightly packed to some degree, and controlled by a parameter. Specifically, the distance between any two points within the same cluster should not vary wildly and should be concentrated around a certain value. This condition is called "y-spherical". The second condition states that the minimum distance between points in different clusters should be significantly larger than the maximum distance between points within the same cluster. This separation ensures

that t-SNE can effectively distinguish between clusters in the lower-dimensional space. When these conditions are met, t-SNE is likely to provide a meaningful visualization that accurately represents the cluster structure of the data. Even when the conditions are not perfectly met, t-SNE can still provide partial visualizations that may reveal some of the structure in the data.

2.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a foundational approach in the field of feature reduction, demonstrating its adaptability across several disciplines by efficiently reducing complex datasets. Its exceptional value was demonstrated in a study on solar power forecasting, where PCA's strategic reduction of data dimensionality to six principle components accounted for more than 91.95 percent of the data variance. This efficiency not only expedited data processing operations but also significantly improved the prediction accuracy of machine learning models targeting electricity production forecasts[8]. PCA's role in neuroimaging expands its application by demonstrating its substantial influence on medical research. PCA helps to simplify the breakdown of complicated, high-dimensional facts into understandable insights, which is critical for understanding the underlying patterns of brain illnesses. This capacity is important, clearing the path for advances in customized medicine and improving diagnostic and prognostic models, emphasizing PCA's critical role in both technical progress and healthcare improvement[9].

2.5 Feature Importance

The paper presents a unique solution to deep learning's Feature Importance Ranking (FIR) using a dual-net architecture, solving the task's fundamental combinatorial optimization issue. The strategy improves the ability to explain AI models by simultaneously identifying an ideal feature subset of fixed size and ranking the importance of features within this subset. This is especially important in domains where knowing the contribution of individual aspects is critical, such as medical diagnostics, since it improves the interpretability of model decision-making processes. Through comprehensive assessment across synthetic, benchmark, and actual datasets, the proposed technique outperforms existing FIR and supervised feature selection methods, highlighting its potential to significantly enhance model performance and dependability in practical application[10]. In the field of feature importance, an establishing paradigm uses entropy metrics to assess the value of features in complicated datasets. This strategy is notably useful for improving data preparation and model interpretability, allowing for a better understanding of the complex interactions between attributes and outcomes. By selecting features based on their informative value, this strategy dramatically improves model performance and assists in the deletion of duplicate or unnecessary data, speeding the analytical process and improving the insights gained from machine learning models. When combined with the document's investigation of feature significance ranking using a dual-net architecture, it becomes evident that such approaches are critical in improving the transparency and efficacy of AI systems across a wide range of applications[11].

2.6 Artificial Bee Colony Algorithm (ABC)

Schiezaro and Pedrini defined the Artificial Bee Colony (ABC) algorithm as an optimization method inspired by honey bees' foraging behavior that is particularly useful for feature selection, where it represents candidate solutions to the feature selection problem using bit vectors [12]. They explained that the algorithm can be divided into several phases, each serving a specific purpose. In the initialization phase, the algorithm generates a set of food sources (solutions) randomly. These food sources represent potential subsets of features. During the employed bee phase, the bees actively explore the search space in an attempt to find new solutions. They assess the food sources' quality and share their findings with other bees. The bees choose their food sources during the observer bee phase according to their level of fitness. Bees can select food sources that show higher potential for enhancing the algorithm's overall performance. In the scout bee phase, new food sources are found to replace exhausted ones. By doing this, the algorithm is guaranteed to keep searching and to stay out of local optima. The main goal of the ABC algorithm is finding the optimal subset of characteristics that minimizes the number of features used while maximizing classification accuracy. The algorithm provides a practical method for resolving feature selection problems by utilizing honey bee behavior found in the natural world[12].

In conclusion, the investigation of feature reduction techniques in this literature review highlights the diverse strategies and algorithms available to address the challenges of complex data in machine learning and data processing. From unsupervised feature selection and recursive feature elimination to advanced methods such as t-SNE, PCA, Feature Importance, and the Artificial Bee Colony algorithm, each approach provides unique advantages and addresses specific aspects of the feature reduction objective. The comparative analysis highlights the importance of selecting the appropriate method based on the specific needs of the data and the goals of the analysis. Ultimately, the effective application of these techniques improves model performance and efficiency while also opening the way for clearer, more interpretable machine learning technologies. Moving forward, our research will focus on the PCA technique. We aim to carefully investigate PCA's potential for managing complex datasets, focusing on its implementation and utility. This specific strategy is aimed at improving our understanding of PCA and relying on its capabilities in feature reduction.

3 Data

This study utilizes the "Sleep Health and Lifestyle" dataset[1], a comprehensive collection of data aimed at exploring the correlation between various lifestyle factors and sleep health. This dataset comprises a wide range of variables, including demographic information, sleep patterns, lifestyle habits, and health-related factors, collected from a diverse sample population. The primary goal of employing this dataset is to apply and evaluate the effectiveness of feature reduction techniques, particularly Principal Component Analysis (PCA), in the context of machine learning models focused on sleep disorder prediction.

Table 1 Descriptive Statistics of Numerical Variables

Variable	Mean	Standard Deviation	Range
Age	42.0	8.00	27 to 59
Sleep Duration (hrs)	7.1	0.79	5.8 to 8.5
Quality of Sleep (1-10)	7.0	1.00	4 to 9
Physical Activity Level (min/day)	59.0	20.00	30 to 90
Stress Level (1-10)	5.0	1.00	3 to 8
Daily Steps	6816.0	1617.00	3000 to 10000
Blood Pressure (systolic/diastolic)	128/84	7/6	115/75 to 142/95
Heart Rate (bpm)	70.0	4.00	65 to 86

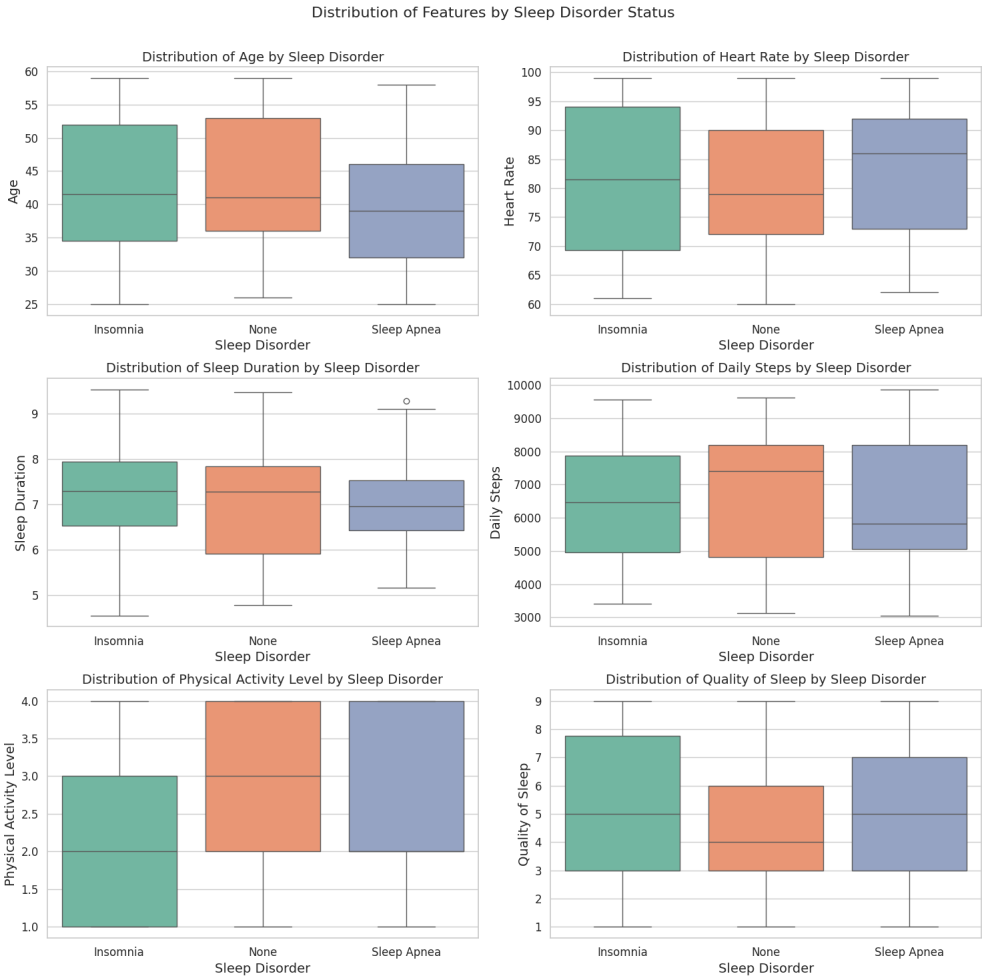

Fig. 1 Distribution of Features by Sleep Disorder Status

Figure 1 presents a series of boxplots that illustrate the distribution of various health-related metrics segmented by sleep disorder status: None, Sleep Apnea, and Insomnia. The top row displays the distribution of Age and Heart Rate, indicating potential trends in how these physiological parameters vary with sleep disorders. The second row reveals variances in Sleep Duration and Daily Steps, providing insights into lifestyle differences across the sleep disorder spectrum. The third row compares Physical Activity Level and Quality of Sleep, offering a visual summary of how these factors correspond to the presence of sleep disorders. The data suggests that there are observable differences among individuals with different sleep disorder statuses, particularly in terms of Sleep Duration and Physical Activity Level, where those with Sleep Apnea exhibit distinctly different patterns compared to the other groups. These visualizations underscore the importance of considering a multifaceted approach when studying the impact of lifestyle factors on sleep health.

4 Methodology

4.1 Methods/Tools Explored

Python's versatility and extensive ecosystem of data science libraries made it an excellent choice for our research project. Using these capabilities, we used Python to preprocess the "Sleep Health and Lifestyle" dataset [1], preparing it for further analysis and modeling.

The implementation was done using several Python libraries, reflecting a broad toolkit for data manipulation, visualization, and machine learning:

- **Pandas:** Employed for data handling, manipulation, and preliminary analysis.
- **Seaborn and Matplotlib:** Employed for data visualization, enabling the inspection of data distributions and relationships between features.
- **NumPy:** Employed for heavy mathematical operations to perform eigenvalue and eigenvector computations.
- **Scikit-learn:** Provided the core PCA functionality, alongside tools for data preprocessing (such as StandardScaler for normalization) and model evaluation. It was specifically used to perform PCA, demonstrating the extraction of principal components and the subsequent reduction in dataset dimensionality.

4.2 Detailed Analysis Outline

The model illustrated in figure 2 outlines a five-phase approach to data analysis and machine learning, emphasizing the role of Principal Component Analysis (PCA) within the process. Below is a concise overview of each phase:

1. **Data Pre-processing and Cleaning:** Refining and organizing the sleep health and lifestyle data which includes removing duplicate records, ensuring uniform data formatting, dividing the data into sections for training and testing, and normalizing the features to be prepared for analysis.

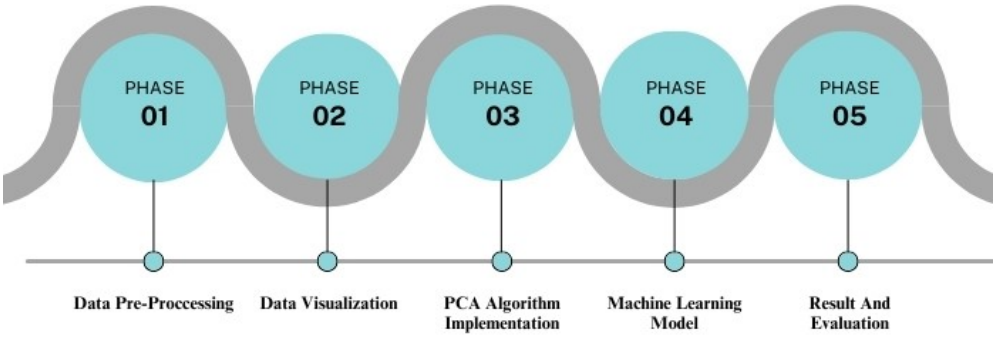

Fig. 2 Sequential Flowchart of the Analytical Process

2. **Data Visualization:** Facilitating a comprehensive understanding of the distributions of variables and the interconnections between them using a range of visualization tools—including histograms, bar charts, and scatter plots.
3. **PCA Algorithm Implementation:** Implementing the PCA algorithm in code form to perform data processing on the "Sleep Health and Lifestyle" dataset.
4. **Machine Learning:** Building and training a machine learning model using the pre-processed and dimensionality-reduced data. The model is designed to learn from the data to make predictions or decisions.
5. **Result And Evaluation:** Performance of the machine learning model is evaluated. This could involve testing the model on a separate dataset to assess its accuracy and precision. Through thorough analysis of these results, the effectiveness of the model is assessed, facilitating any required adjustments or enhancements to optimize its performance.

4.3 Feature Reduction PCA

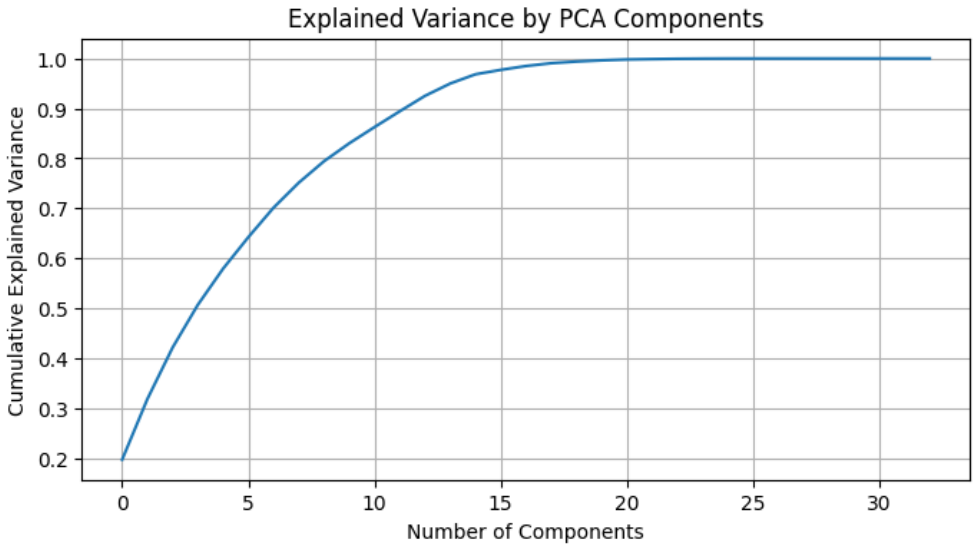

Fig. 3 cumulative variance ratio

The implementation of the PCA algorithm was carried out through a structured approach encompassing several key steps to effectively analyze the dataset and evaluate the dimensionality reduction process. Initially, the PCA Application to the scaled training data to determine the optimal number of components to capture at least 95% of the variance. Then the cumulative variance ratio to visually assess the number of components required as shown in Figure 3.

The plot shows the explained variance leveling off as it nears 26 components, indicating that additional components beyond this number contribute less to the variance explained by the model. Therefore, selecting 26 components strikes a balance between retaining most of the variance in the data and reducing dimensionality to simplify the model and potentially improve computational efficiency.

The original dataset consisted of 33 features, which have been reduced to 26 principal components. This reduction in dimensionality simplifies the data without sacrificing critical information.

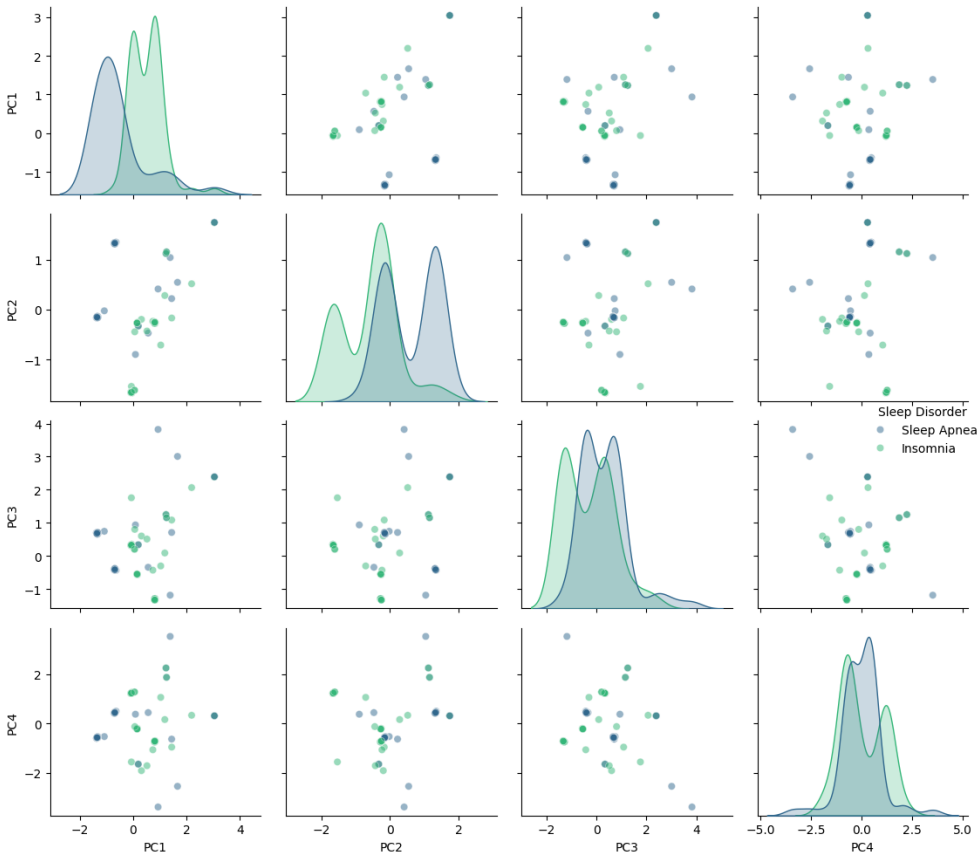

Fig. 4 Scatter Plots of PCA Components

The interrelationships among the first four principal components resulting from PCA on the sleep disorder dataset are depicted through a series of scatter plots. In this visualization, we examine each pairwise combination of principal components, presented in a grid format. The distributions of individual components are shown along the grid diagonal through density plots, while the off-diagonal scatter plots provide insight into the correlation patterns between component pairs. Each data point is color-coded to represent distinct categories of the target variable, "Sleep Disorder," specifically "Sleep Apnea" and "Insomnia." The plots offer a visual exploration of the dataset's structure, as transformed by PCA, highlighting how the observations cluster or disperse across the principal component axes.

Moreover, the results of PCA were evaluated by performing PCA on a dataset composed of random values. The variance explained by each principal component was calculated and plotted in a Scree Plot, along with the cumulative variance explained, aiding in determining the optimal number of components to retain for optimal data representation.

Finally, the output included feature reduction achieved through PCA, with the original features transformed into a new set of variables capturing most of the original dataset's information. Additionally, PCA component loadings were analyzed to quantify each original feature's contribution to the direction of each principal component. Lastly, the model's performance was assessed using RandomForestClassifier, Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) models on both the PCA-transformed features and the full feature set, evaluating the impact of PCA on model accuracy through performance metrics such as accuracy scores.

5 Machine Learning Models

This study evaluated the efficacy of Principal Component Analysis (PCA) in enhancing the accuracy of various machine-learning models within the context of predicting sleep disorders. The machine learning models employed are RandomForestClassifier, Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) chosen for their diversity and wide application in classification tasks. The accuracy of these models was assessed before and after the application of PCA, providing insight into the impact of dimensionality reduction on model performance.

- **RandomForestClassifier:** a robust ensemble learning method based on decision trees, is renowned for its accuracy and capability to handle large data sets with a higher dimensionality. It operates by constructing multiple decision trees during training time and outputting the class that is the mode of the classes of the individual trees[13].
- **Logistic Regression**: despite its name, is a linear classification model rather than regression. It is particularly useful for binary classification problems. This model estimates probabilities using a logistic function, which is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but not exactly at those limits[14].
- **Support Vector Classifier (SVC)**: is a powerful, non-linear classification technique that seeks to find the optimal hyperplane that maximizes the margin between two classes. It is well-suited for complex classification issues with a clear margin of separation[15].
- **K-Nearest Neighbors (KNN)**: is a classic non-parametric classification method that merely stores feature vectors and class labels of the training data. In testing, it assigns a label based on the most frequent among the k nearest training instances. Statistically, k-NN guarantees an error rate not more than twice the Bayes error rate as data size increases indefinitely [16].

PCA was applied to the dataset before model training to reduce its dimensionality. This method of feature reduction aims to transform the original variables into a smaller number of uncorrelated variables, known as principal components while retaining as much of the original variability as possible.

6 Result and Evaluation

6.1 Principal Component Analysis (PCA) Impact on Machine Learning Models

We applied PCA to a preprocessed sleep health and lifestyle dataset, selecting the top seven principal components to reduce dataset complexity while retaining critical information. This process was validated through scree plots and cumulative variance explained graphs, as shown in Fig. 4.

6.2 Comparative Analysis of Machine Learning Models

Table 2 Model Performance Before and After PCA

Model	Accuracy Before PCA	Accuracy After PCA	Time Before PCA	Time After PCA
RandomForest	83.87%	93.54%	0.177907	0.168384
LogisticRegression	87.10%	90.32%	0.005161	0.002880
SVC	90.32%	90.32%	0.001254	0.001681
KNeighbors	83.87%	90.32%	0.000565	0.00670

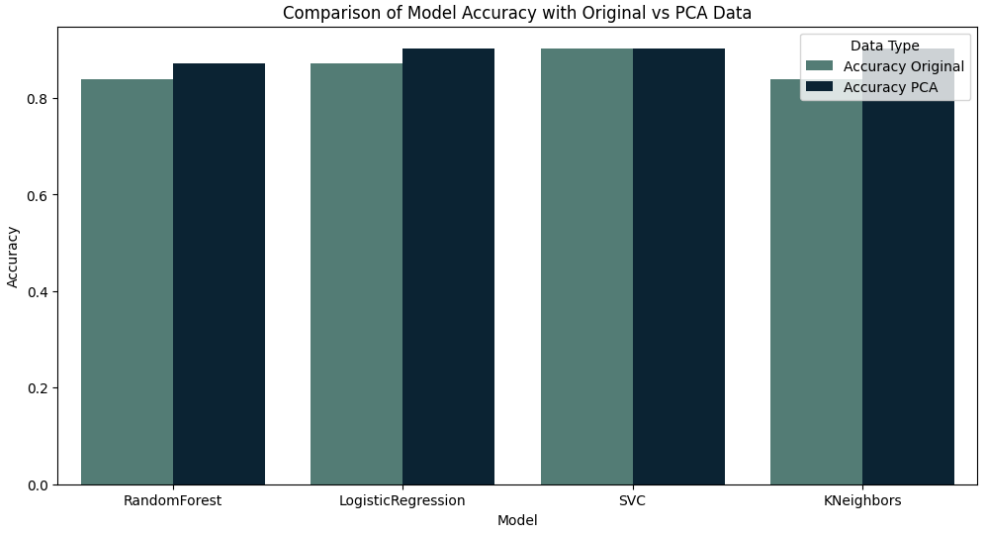

Fig. 5 Accuracy Tracking Across Models for Original and PCA-Transformed Data

1. **RandomForestClassifier**
   - **Accuracy:** There was an improvement post-PCA, with accuracy increasing from 83.87% to 93.54%.
   - **Training Time:** A slight increase in training time was observed.
   - **Evaluation:** PCA has contributed to better model accuracy with a negligible increase in training time, suggesting a favorable trade-off.
2. **LogisticRegression**
   - **Accuracy:** Accuracy improved from 87.10% to 90.32% after applying PCA.
   - **Training Time:** There was a decrease in training time.
   - **Evaluation:** The increase in accuracy coupled with reduced training time post-PCA indicates that Logistic Regression benefits from dimensionality reduction in both performance and efficiency.
3. **SVC**
   - **Accuracy:** Remained constant at 90.32% after PCA.
   - **Training Time:** Training time decreased slightly.
   - **Evaluation:** The SVC model's performance is unaffected by PCA, showcasing its robustness to changes in feature space dimensionality.
4. **KNeighbors**
   - **Accuracy:** Marked increase in accuracy from 83.87% to 90.32% post-PCA.
   - **Training Time:** Training time remained relatively the same, with a minor decrease.
   - **Evaluation:** The k-Nearest Neighbors classifier shows a significant improvement in accuracy with PCA, which could be attributed to noise reduction in the feature set.

6.3 Comparative Analysis with Published Solution

We compared our model performances with those reported by Alshumayri et al. (2024)[17], who also used machine learning to predict sleep disorders and used the exact dataset[1].

Table 3 Performance Comparison Table

Model	Our Model Accuracy	Reported Accuracy	Difference
RandomForest	93.54%	92.15%	+1.39%
Support Vector Machine	90.32%	90.02%	+0.30%
k-Nearest Neighbors	90.32%	86.91%	+3.41%

The reported accuracy values are the mean accuracies from the paper[17].

6.4 Discussion and Justification of Results

The application of PCA as shown in Table 2 has yielded mixed effects on the performance of the evaluated models. For RandomForest, there was an accuracy increase from 83.87% to 93.54%, alongside a minor uptick in training time. LogisticRegression experienced an accuracy boost from 87.10% to 90.32%, with the added benefit of reduced training time, which underscores the efficiency gains attributable to PCA. The SVC model maintained its accuracy consistently at 90.32%, although its training time increased slightly. In the case of KNeighbors, PCA remarkably enhanced accuracy from 83.87% to 90.32%, with a negligible rise in training time. These outcomes underscore that PCA's influence is not uniform across different algorithms; it can significantly bolster both efficiency and accuracy for certain models, while for others, the principal advantage may lie in computational time-saving rather than predictive performance.

As delineated in Table 3, the RandomForest model shows strong performance, slightly outperforming the accuracy reported in the study by Alshumayri et al.[17] with an increase of 1.39%. Conversely, our Support Vector Machine model shows a modest improvement over the reported figure, with a 0.30% increase in accuracy, suggesting a stable performance regardless of dimensionality reduction through PCA. The k-Nearest Neighbors model, post-PCA, shows a significant boost, outpacing the reported accuracy by 3.41%. This considerable improvement indicates that PCA may have effectively simplified the feature space, thereby enhancing the model's ability to make accurate predictions. These observations underscore the varied impact of PCA on different machine learning models, highlighting the necessity of bespoke preprocessing strategies to maximize model accuracy.

7 Conclusion

This research explored the utilization of Principal Component Analysis (PCA) alongside other dimensionality reduction techniques to predict sleep disorders effectively. The strategic application of PCA has shown differential impacts on the performance of various machine learning models used to predict sleep disorders. Post-PCA, RandomForestClassifier and Logistic Regression models exhibited increased accuracy, with Logistic Regression also benefiting from reduced training time. The SVC model's accuracy remained unchanged, indicating its robustness to dimensionality reduction, while the K-Nearest Neighbors model saw a significant improvement in accuracy. These findings affirm PCA's role in improving predictive analytics in certain contexts, highlighting its potential as a valuable tool in enhancing the interpretability and efficiency of machine learning models within the domain of healthcare analytics.

This variance in model performances before and after PCA application underscores the nuanced nature of PCA's impact. It suggests that while PCA is generally effective in streamlining complex datasets and expediting training processes, the extent of its benefits is model-dependent. The study's findings advocate for a contextual application of PCA, tailored to the unique demands of each model, to harness the full potential of dimensionality reduction techniques. As predictive modeling becomes increasingly vital in healthcare analytics, the insights from this study could guide practitioners

in employing PCA judiciously, optimizing the balance between model complexity and interpretability to advance the field of medical diagnostics.

References

[1] uom190346a: Sleep Health and Lifestyle Dataset. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset. Accessed: 2023-09-24 (2023)

[2] Alazaidah, R., Samara, G., Aljaidi, M., Qasem, M.H., Alsarhan, A., Alshammari, M.S.: Potential of machine learning for predicting sleep disorders: A comprehensive analysis of regression and classification models. Diagnostics (2023) https://doi.org/10.3390/diagnostics14010027

[3] Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A review of unsupervised feature selection methods. Artificial Intelligence Review **53**(2), 907–948 (2020)

[4] Chen, J., Zeng, Y., Li, Y., Huang, G.-B.: Unsupervised feature selection based extreme learning machine for clustering. Neurocomputing **386**, 198–207 (2020)

[5] Sharma, N.V., Yadav, N.S.: An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers. Microprocessors and Microsystems **85**, 104293 (2021)

[6] Arora, S., Hu, W., Kothari, P.K.: An analysis of the t-sne algorithm for data visualization. In: Bubeck, S., Perchet, V., Rigollet, P. (eds.) Proceedings of the 31st Conference On Learning Theory. Proceedings of Machine Learning Research, vol. 75, pp. 1455–1462. PMLR, ??? (2018). https://proceedings.mlr.press/v75/arora18a.html

[7] Soni, J., Prabakar, N., Upadhyay, H.: Visualizing High-Dimensional Data Using t-Distributed Stochastic Neighbor Embedding Algorithm, pp. 189–206 (2020). https://doi.org/10.1007/978-3-030-43981-1_9

[8] Chahboun, S., Maaroufi, M.: Principal Component Analysis and Machine Learning Approaches for Photovoltaic Power Prediction: A Comparative Study. https://www.mdpi.com/2076-3417/11/17/7943 (2023)

[9] Kherif, F., Latypova, A.: Chapter 12 - principal component analysis. In: Mechelli, A., Vieira, S. (eds.) Machine Learning, pp. 209–225. Academic Press, ??? (2020). https://doi.org/10.1016/B978-0-12-815739-8.00012-2 . https://www.sciencedirect.com/science/article/pii/B9780128157398000122

[10] Wojtas, M., Chen, K.: Feature importance ranking for deep learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 5105–5114. Curran

Associates, Inc., ??? (2020)

[11] Adler, A.I., Painsky, A.: Feature importance in gradient boosting trees with cross-validation feature selection. Entropy **24**(5) (2022) https://doi.org/10.3390/e24050687

[12] Schiezaro, M., Pedrini, H.: Data feature selection based on artificial bee colony algorithm. EURASIP Journal on Image and Video Processing **2013**(1) (2013) https://doi.org/10.1186/1687-5281-2013-47

[13] Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)

[14] Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. John Wiley & Sons, ??? (2013)

[15] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)

[16] Shi, Z.: Improving k-nearest neighbors algorithm for imbalanced data classification. IOP Conference Series: Materials Science and Engineering **719** (2020) https://doi.org/10.1088/1757-899X/719/1/012072

[17] Alshumayri, S., Abdallah, R., Asseri, Y., Balfagih, Z.: Sleep disorder analysis: Unveiling the interplay between lifestyle health and sleep quality. In: 21st Learning and Technology Conference (L&T), pp. 149–154. IEEE, Jeddah, Saudi Arabia (2024). https://doi.org/10.1109/LT60077.2024.10468720