Research on the Automatic Classification of Sleep Disorders in Sleep Health and Lifestyle Datasets Based on Large Language Model Prompting

**Abstract**

Sleep disorders are an urgent global public issue, affecting approximately 30% of the global population. This study utilizes the powerful semantic understanding and knowledge reasoning capabilities of LLMs to classify sleep disorders in a dataset covering sleep, lifestyle, and related health factors. Three novel prompting paradigms are employed to guide LLMs to automatically complete the design, training, and evaluation of classifiers. The experimental results show that the support vector machine classifier automatically found using decomposed prompting achieves a classification accuracy as high as 91.9% (F1 - score: 0.919), significantly outperforming traditional zero - shot prompting and few - shot prompting methods in terms of accuracy.This study uniquely integrates the semantic understanding and knowledge reasoning capabilities of large language models with automated machine learning seamlessly, providing a new paradigm for the classification of sleep disorders in health informatics.

Keywords: Sleep Disorder Classification; Large Language Model; Prompt Engineering; Health Informatics

1.Introduction

1.1 The Gravity of Sleep Disorder Issues

Sleep disorders are a significant problem, affecting a large portion of the population and potentially having a substantial impact on overall health and quality of life. According to statistics from the World Health Organization (WHO), approximately 10% of the global population is troubled by sleep disorders, and this proportion is on the rise against the backdrop of accelerated urbanization. Sleep disorders can not only lead to daytime fatigue and decreased attention but also trigger chronic diseases such as cardiovascular diseases and diabetes. Therefore, accurately identifying and classifying sleep disorders is of great significance for improving public health.

1.2 Dilemmas of Traditional Sleep Disorder Classification Methods

In the past, the classification of sleep disorders mainly relied on traditional machine - learning algorithms, such as decision tree algorithms [1], support vector machines (SVM) [2], random forest algorithms [3], etc. These algorithms have played a certain role in sleep disorder research. Researchers used decision tree algorithms to analyze sleep monitoring data, attempting to identify the characteristic patterns of different sleep disorders through a series of conditional judgments and branch decisions. Support vector machines distinguish normal sleep data from sleep disorder data by finding an optimal classification hyperplane. The random forest algorithm improves the accuracy and stability of classification by constructing multiple decision trees and synthesizing their prediction results

However, traditional machine learning algorithms have many difficulties in classifying sleep disorders. Every step of these algorithms requires a large amount of manual operation. In the data pre - processing stage, researchers need to manually handle missing values and outliers in the data, and perform operations such as standardization and normalization to ensure the quality and usability of the data. In the model training and tuning process, researchers need to manually select appropriate algorithms, set model parameters, and optimize the model performance through repeated trials.。

This manual operation method not only consumes a large amount of human and time costs but is also easily affected by human factors, resulting in certain limitations in the accuracy and reliability of the results. Moreover, the application of traditional machine - learning algorithms highly depends on expert experience. Experts need to select appropriate algorithms, determine the methods of feature engineering, and adjust model parameters based on their professional knowledge and experience. For complex sleep disorder classification problems, the experiences and judgments of different experts may vary, leading to inconsistencies in classification results.。

These problems of traditional machine - learning algorithms in sleep disorder classification limit the development and application of sleep disorder research, and a new technology and method are needed to break through these dilemmas.。

1.3 New Opportunities Brought by Large Language Models

Large language models (LLMs), as a cutting - edge technology in the field of natural language processing, have made remarkable progress in recent years. Based on the Transformer [14] architecture, they have learned rich language knowledge through unsupervised pre - training on massive text data. They have demonstrated great potential in numerous fields, providing new ideas and methods for solving complex problems.。

In the field of medical research [4], [6], [7], [8], [9], the application of large language models has brought new opportunities for sleep disorder classification. Large language models can understand and process natural language, enabling them to directly analyze and interpret the text information in sleep health and lifestyle datasets. They can extract key information from text data such as patients' sleep logs and descriptions of living habits.。

Large language models also possess powerful knowledge reasoning capabilities [5]. They can comprehensively analyze and judge the extracted information by combining existing medical knowledge and the diagnostic criteria of sleep disorders, thus achieving the automatic classification of sleep disorders. When faced with complex sleep disorder symptoms, large language models can accurately identify different types of sleep disorders, such as insomnia and sleep apnea, through reasoning and judgment.。

Moreover, large language models can also discover the potential relationships between sleep disorders and other factors through learning from a large amount of data, providing new bases for the diagnosis and treatment of sleep disorders. 。

1.4 Research Contributions

In the field of sleep disorder classification research, traditional methods were highly dependent on manual operations and expert experience, which was not only inefficient but also difficult to cope with the increasingly complex sleep health data, with obvious limitations。

Prompting , as a key technology in the application of large language models, plays a decisive role in guiding the model to generate the expected output. The performance of different prompting strategies in various tasks varies. In particular, how to skillfully use the prompting strategy to enable the large language model to accurately extract key information from sleep health and lifestyle data based on text information alone and achieve automatic sleep disorder classification is a difficult problem that urgently needs to be solved. Under this background, this study makes important and unique contributions in many aspects：

* **Innovative Model Application**： Pioneeringly introduce large language models into the field of automatic sleep disorder classification. Traditional sleep disorder classification methods have increasingly prominent problems such as low accuracy and poor adaptability when facing massive, complex, and changeable sleep health and lifestyle data. Large language models, with their powerful language understanding and generation capabilities, have broken through the numerous limitations of traditional methods. This study constructs a new set of automatic sleep disorder classification methods based on large language models. The core advantage of this method is that as long as text information is provided, the model can automatically perform machine learning (auto ML) to achieve high - precision prediction of sleep disorders, opening up a new and efficient way for the accurate diagnosis of sleep disorders。

1. **Deep Exploration of Prompting Strategies to Facilitate Automatic Classification：** This study deeply explores the application of zero - shot prompting, few - shot prompting, and decomposed prompting techniques in large language models. This study systematically explores the new potential of these prompting strategies in the automatic sleep disorder classification task, clearly clarifying their advantages and disadvantages in text - based sleep health classification.These in - depth analyses provide a clear direction for the subsequent optimization of prompting strategies and help to continuously improve the performance of large language models in automatic sleep disorder classification based on text information alone.。

* **Leading Cross - disciplinary Integration and Automation Innovation：** This study, through prompt - driven large language models, has achieved the innovative application of automatic machine - learning classification based on text information alone, breaking the shackles of traditional machine - learning algorithms and truly realizing the automation and intelligence of sleep disorder classification. This innovative application not only effectively solves practical problems in the field of sleep medicine but, more importantly, accumulates valuable experience for the wider application of large language models in the field of healthcare, greatly promoting the deep integration and development of cross - disciplinary technologies and opening up a new path full of infinite possibilities for the cross - research of sleep medicine and artificial intelligence.。
* **Expanding Automation Application Scenarios:：**in practical applications, large language models have demonstrated powerful automation expansion capabilities. They can be combined with a variety of technologies to further expand their automation application value in sleep disorder classification and related fields (such as sensor technology and wearable devices).

The results of this study not only help to significantly improve the diagnosis and treatment level of sleep disorders and improve the quality of life of patients but also, with the continuous development and improvement of technology, the application prospects of large language models in the automatic classification of sleep medicine based on text information will be broader, injecting strong impetus into the automation and intelligence development of the entire sleep medicine industry.。

## 2.Sleep Health and Lifestyle Dataset

### 2.1 Source and Composition of the Dataset

The sleep health and lifestyle dataset used in this study is sourced from the Kaggle website [10]. The sleep health and lifestyle dataset used this time contains 374 rows and 13 columns of data. Details are as follows：

1. **Personal basic information:**Person ID, as the unique identifier for each respondent, helps to accurately identify and track individual data during the data processing and analysis. Gender information can be used to study the differences in the incidence of sleep disorders and sleep patterns between different genders. Age is one of the important factors affecting sleep. As age increases, sleep quality tends to decline, and the incidence of sleep disorders also increases. Occupation reflects the potential impact of work nature, working hours, and work pressure on sleep.。
2. **Sleep - related characteristics**： Sleep Duration directly reflects an individual's sleep time. Sufficient sleep duration is crucial for maintaining physical health and normal physiological functions. Quality of Sleep is a subjective evaluation index, measured on a scale from 1 - 10, which reflects aspects such as the depth, continuity, and recovery effect of sleep. 。
3. **Lifestyle factors**： Physical Activity Level reflects an individual's daily exercise volume. Stress Level is one of the important factors affecting sleep. BMI Category is an indicator to measure whether an individual's weight status is healthy and is closely related to sleep disorders. Daily Steps is a simple and intuitive indicator to measure the physical activity level, through which the daily activity volume of an individual can be understood, and then its impact on sleep can be analyzed. 。
4. **Health indicators**： Blood Pressure is essential for maintaining the normal functions of various organs in the body. Heart Rate reflects the functional state of the heart。

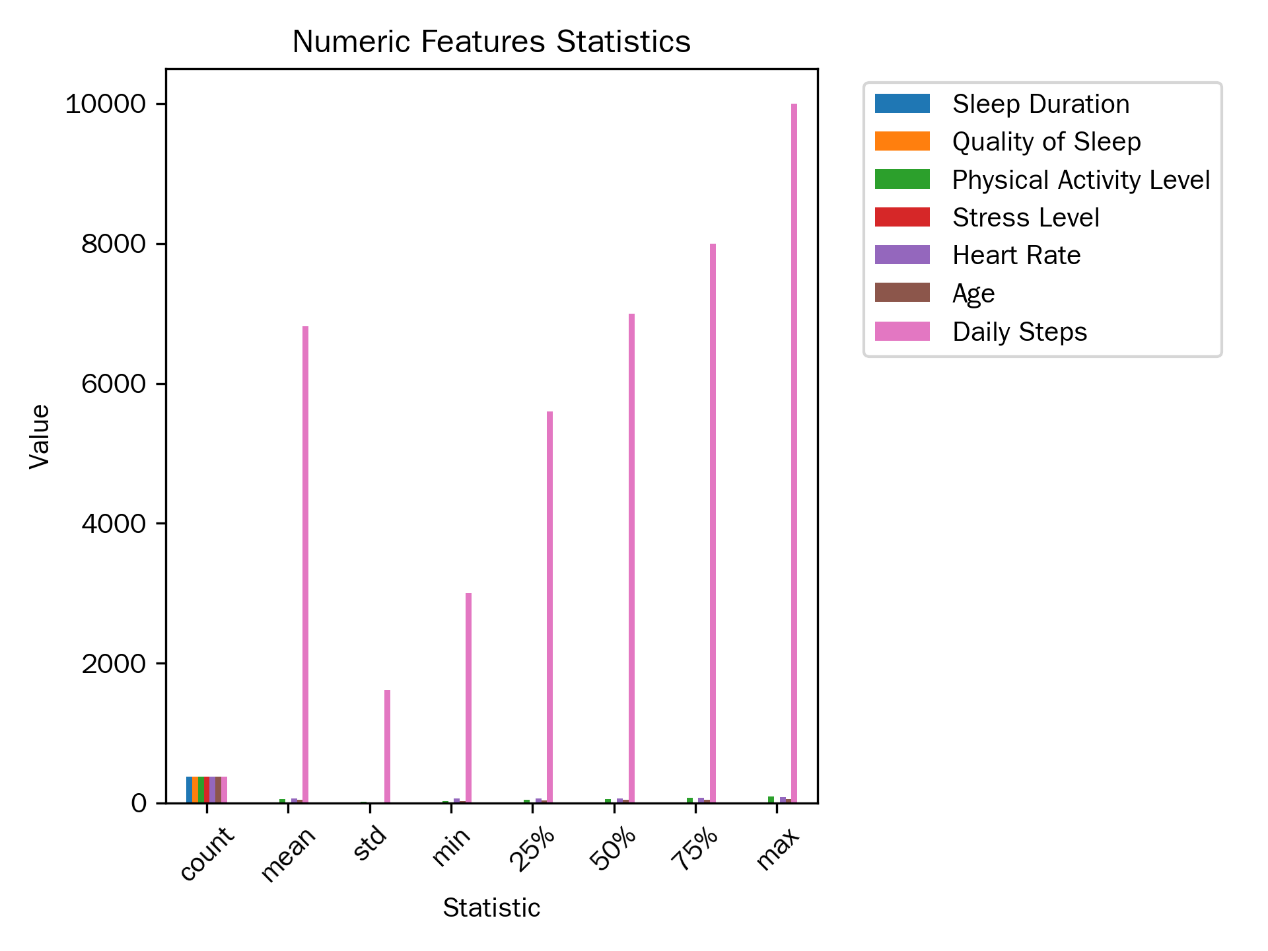
These variables are interrelated and jointly reflect the sleep health status and lifestyle characteristics of the respondents. Through in - depth analysis of these variables, the potential relationship between Sleep Disorder and lifestyle can be revealed, providing a scientific basis for the diagnosis, treatment, and prevention of sleep disorders.。

### 2.2 Data Feature Analysis and Visualization

**(1)Statistical information of numerical features (Figure 1)**

The following inferences can be made from the statistical data of these numerical features：

* **Sleep Duration**： The average sleep duration is approximately 7.13 hours, with a relatively small standard deviation, indicating that the overall distribution is relatively concentrated. Most people's sleep duration is between 6.4 - 7.8 hours, which is within the normal sleep duration range for adults. However, there may still be some individuals with insufficient or excessive sleep duration, which may be related to factors such as living habits and work pressure.。
* **Quality of Sleep score:** The average score is 7.31 points, with a standard deviation of 1.20 points, indicating some differences among individuals. The 25% - 75% quantiles show that most people's scores are between 6 - 8 points, and the overall sleep quality is above - average. However, a considerable number of respondents have poor sleep quality, with scores below 5 points, which may be comprehensively affected by various factors such as psychological stress and living environment.。
* **Physical Activity Level**： The average level is 59.17, with a relatively large standard deviation of 20.83, indicating significant differences in physical activity levels among different individuals. This is closely related to factors such as personal exercise habits and occupational characteristics. For example, there may be a large difference in physical activity levels between people in physically - laborious occupations and those who sit in the office for long hours.。
* **Stress Level**： The average stress level is 5.39, with a standard deviation of 1.77, indicating a certain degree of dispersion in the distribution of stress levels. Different individuals face different levels of stress, which may be related to factors such as work nature and life events. Higher stress levels may have a negative impact on health indicators such as sleep quality。
* **Heart Rate**： The average heart rate is 70.17 beats per minute, with a standard deviation of 4.14. The fluctuation range of the heart rate is relatively small, and most people's heart rates are between 68 - 72 beats per minute. This reflects that the heart rate in this dataset is generally stable, but the heart rates of some individuals may deviate from the normal range, which may be related to physical health conditions, exercise, etc。
* **Age**： The average age is 40.06 years old, with a standard deviation of 13.34, indicating a certain degree of dispersion in the age distribution. People of different age groups may have differences in sleep - related characteristics. For example, older people may have shorter sleep duration or poorer sleep quality.。
* **Daily Steps:** The average number of daily steps is 7070.26 steps, with a standard deviation of 3344.52. The differences in the number of steps reflect the different exercise habits and activity levels of different individuals. The number of steps may be related to physical activity levels, sleep quality, etc. Those with more steps may have higher physical activity levels and better sleep quality.。



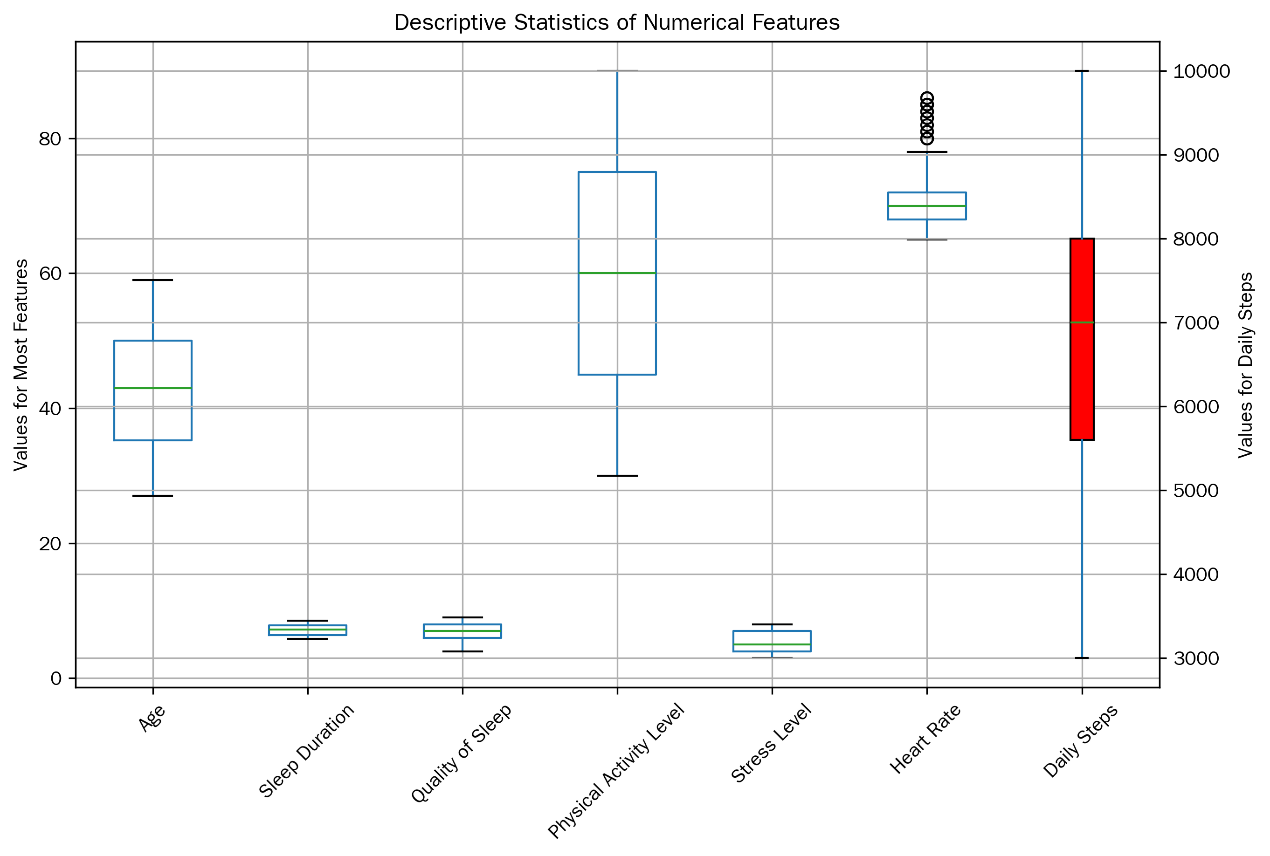


Figure 1: Statistical chart of numerical features

**(2)Distribution information of categorical features (Figure 2)**

From the distribution of categorical features, we can observe the following：

* **Gender**： There are 212 males and 162 females. The number of males is slightly more than that of females. People of different genders may have differences in sleep - related characteristics. For example, due to physiological factors, females may perform differently in terms of sleep quality and stress level.。
* **Occupation**： The number of people in each occupation is the same, all 46, reflecting a certain balance in the sample in terms of occupation. Different occupations may show different characteristics in terms of sleep duration, quality, and stress level due to factors such as work pressure and working hours. For example, nurses may have their sleep duration and quality affected and a higher stress level due to the shift - work nature of their jobs。
* **BMI Category**： There are 156 respondents with normal weight, 121 who are overweight, and 97 who are obese. The proportion of respondents with normal weight and overweight is relatively high. People in different BMI categories may have different incidences of sleep disorders. Obese people may be more prone to problems such as sleep apnea.。
* **Sleep Disorder status**： There are 236 respondents with normal sleep, 85 with sleep apnea, and 53 with insomnia. Understanding the distribution of different types of sleep disorders helps to conduct targeted research on the causes and preventive measures of sleep disorders. For example, patients with sleep apnea may be related to factors such as obesity and age。

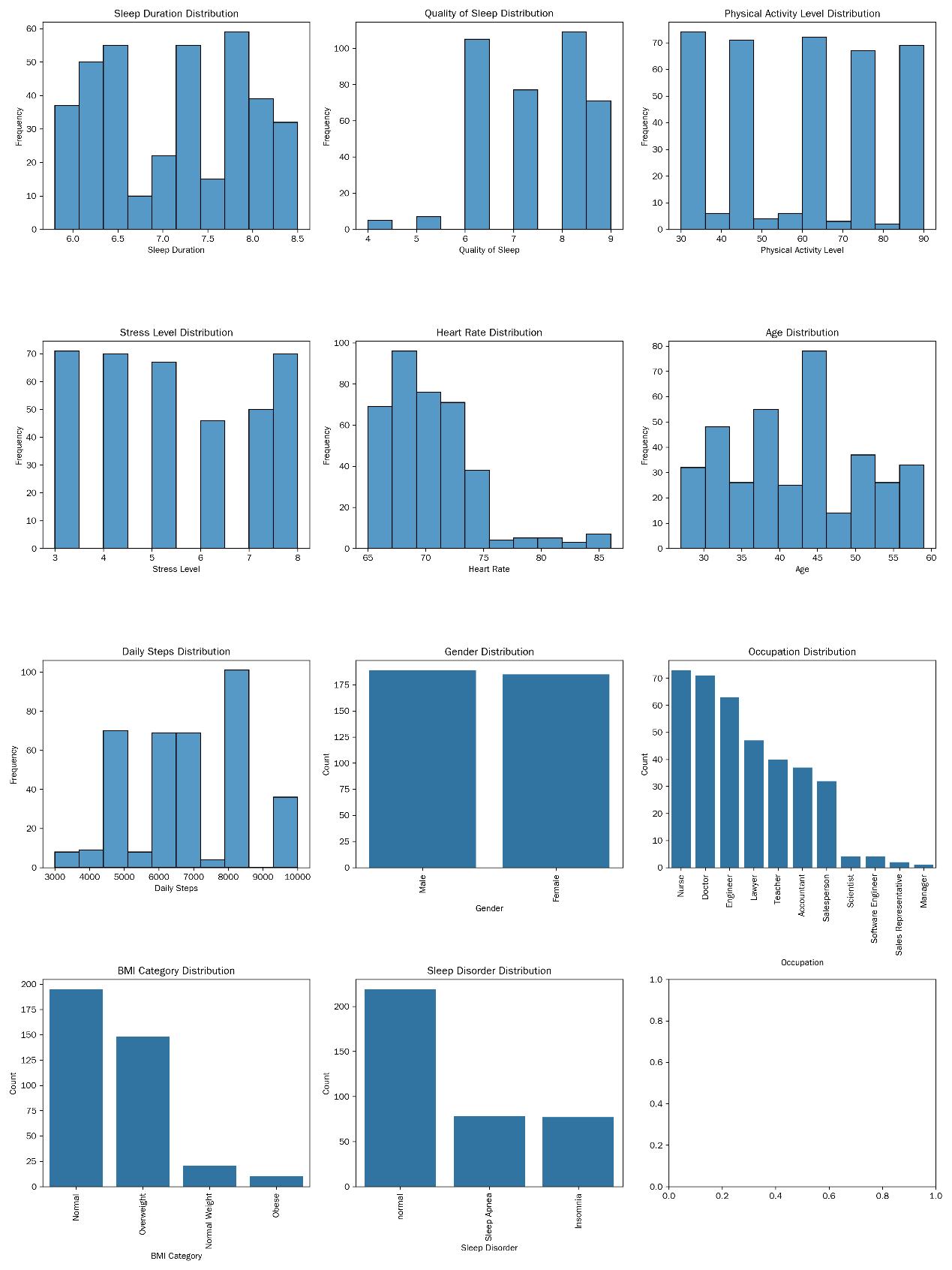


Figure 2: Distribution chart of categorical features

**(3) Correlation analysis among features**

The Pearson correlation coefficients among various features were calculated (for categorical features, appropriate methods were used to analyze their relationships with other features, such as using the chi - square test to analyze the relationship between Gender and Sleep Disorder, etc.). The correlation coefficient matrix was presented by drawing a heatmap (Figure 3)。

The following conclusions can be drawn from the correlation coefficient matrix：

1. **Sleep Duration and Quality of Sleep**： The correlation coefficient between the two is 0.883213, showing a strong positive correlation. This is in line with the general understanding that the longer the sleep duration, the more fully the body and brain can recover, thus improving sleep quality. This result suggests that in measures to improve sleep quality, adjusting sleep duration can be considered.。
2. **Physical Activity Level and other features**： The correlation coefficients between Physical Activity Level and Sleep Duration, Quality of Sleep are 0.212360 and 0.192896 respectively, showing a weak positive correlation. This indicates that moderate physical activity has a certain positive impact on sleep, but this impact is relatively limited. The correlation coefficient between Physical Activity Level and Stress Level is almost 0, indicating that the linear relationship between the two is not obvious. The correlation coefficient with Heart Rate is 0.136971, showing a weak positive correlation, which may be that a higher physical activity level can cause a certain increase in heart rate.。
3. **Stress Level and other features**： Stress Level shows a strong negative correlation with Sleep Duration and Quality of Sleep, with correlation coefficients of - 0.811023 and - 0.898752 respectively. This indicates that stress is an important factor affecting sleep. When people are under high stress, the stress hormones secreted by the body will interfere with falling asleep and sleep quality. The correlation coefficient between Stress Level and Heart Rate is 0.670026, showing a positive correlation, indicating that the greater the stress, the higher the heart rate may be, reflecting the impact of stress on physical physiological indicators。
4. **Heart Rate and other features**： Heart Rate shows a negative correlation with Sleep Duration and Quality of Sleep, meaning that when the heart rate is high, the sleep duration and quality may be poor. This may be because an abnormal heart rate reflects a certain uncomfortable state of the body, which in turn affects sleep.。
5. **Age and other features**： Age shows a negative correlation with Sleep Duration and Quality of Sleep, indicating that as age increases, sleep duration may decrease and sleep quality may deteriorate. Age shows a positive correlation with Stress Level, and it is possible that older people face relatively greater life pressure.。
6. **Daily Steps and other features**： Daily Steps show a positive correlation with Sleep Duration and Quality of Sleep, indicating that more steps may help to extend sleep duration and improve sleep quality. Daily Steps show a strong positive correlation with Physical Activity Level, which is as expected, the more steps, the higher the physical activity level.

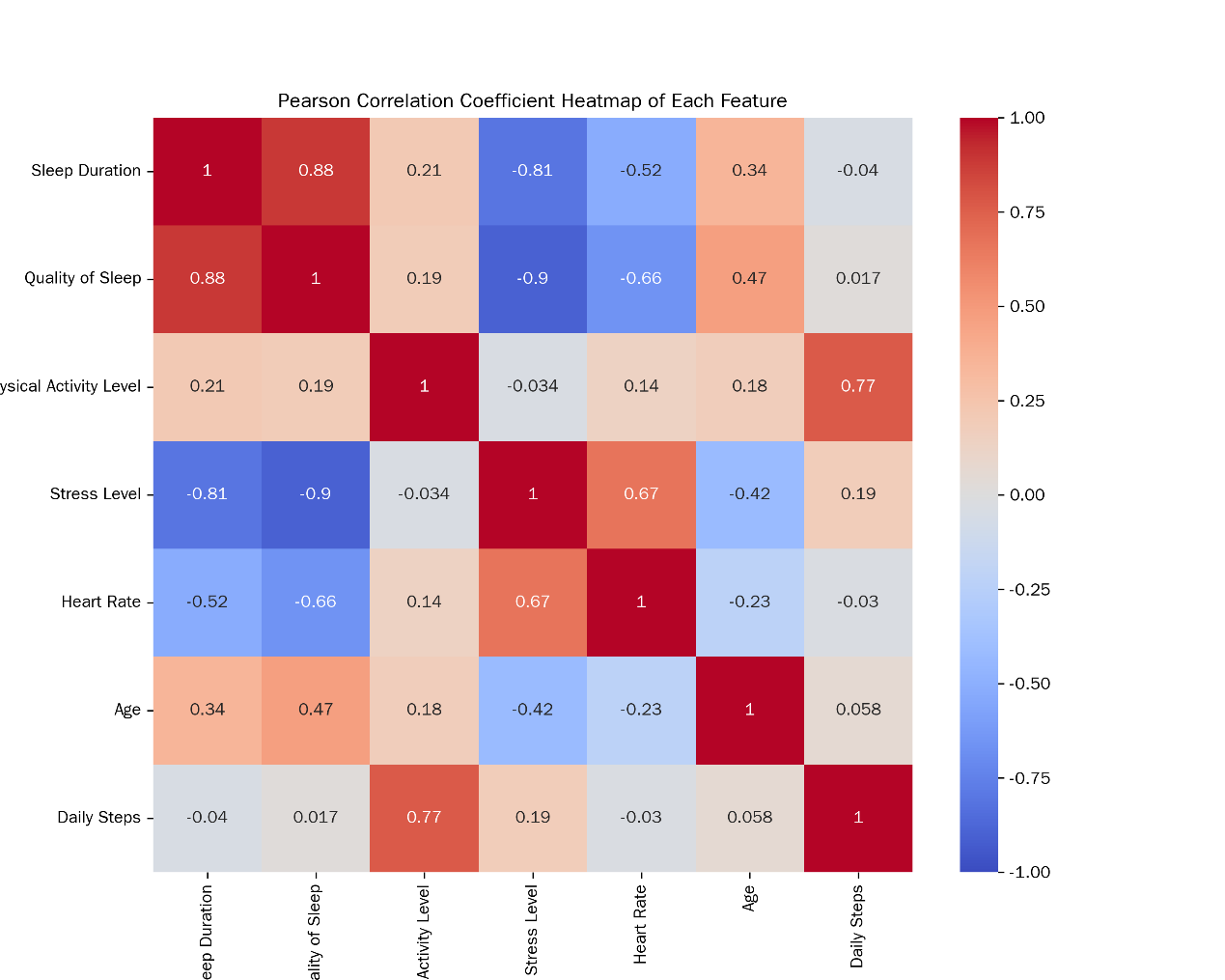


Figure 3: Heatmap of Pearson correlation coefficients among features

By conducting statistical analysis and visualizing the distribution of each feature in the dataset, we can gain a deeper understanding of the characteristics and patterns of the data, providing strong data support for the subsequent use of large language models in sleep disorder classification. In subsequent research, we can select appropriate features and models based on these analysis results to improve the accuracy and reliability of sleep disorder classification 。

## 3.Research Methods

### 3.1 Selection of Large Language Mode

In the research field of automatic sleep disorder classification, the selection of a large language model is a key factor affecting the accuracy and efficiency of research results. After comprehensive consideration, this study finally selects the Doubao model, mainly based on the following in - depth analysis.。

* The Doubao model is built on the Transformer [14] architecture. Through unsupervised pre - training on massive text data, it has extensively and deeply learned the knowledge systems of language in semantics, grammar, pragmatics, and other aspects. Its unique multi - head attention mechanism can concurrently focus on different parts of the input text, thus capturing rich semantic associations and contextual information in the text. The multi - layer neural network structure further deeply processes and extracts features from this information, enabling the Doubao model to deeply analyze text content from multiple dimensions such as vocabulary, sentences, and passages
* In natural language processing tasks, the Doubao model can fully understand professional medical terms in the field of sleep health, such as "apnea - hypopnea index" and "periodic limb movement disorder". Moreover, with good generalization performance, the Doubao model can quickly adapt to the complex and variable text data in the field of sleep health, including medical records, log data from sleep monitoring devices, etc. These data come from a wide range of sources, with significant differences in format and language style, but the Doubao model can handle them effectively.
* In addition, the Doubao model has some unique advantages that greatly facilitate the realization of automated sleep disorder classification. At the interaction level, some versions (such as the PC version) of the Doubao model have a user - friendly interface that can directly accept and analyze common data formats such as CSV files. This enables researchers to conveniently import the sleep health and lifestyle dataset into the model for processing without spending a lot of time on pre - processing tasks such as data format conversion. At the technical implementation level, the Doubao model has the ability to automatically generate and execute code. In the automated sleep disorder classification process, this ability can quickly generate the code logic for implementing the classification algorithm based on the text analysis results and directly execute it. This not only greatly improves the classification efficiency but also can correct potential syntax errors and logical loopholes that may occur during manual code writing, providing strong support for the realization of efficient and accurate automated sleep disorder classification

In summary, with its powerful natural language processing capabilities, high adaptability to data in the field of sleep health, and unique interaction and technical implementation advantages, the Doubao model becomes the top choice for automatic sleep disorder classification in this study。

### 3.2 Exquisite Prompt Design Strategy

As shown in Table 1, details of prompts, we designed three prompting strategies：

* + Zero - shot Prompting: Rule - guided Classification Exploration: The core concept is to guide the model to generate corresponding answers or complete specific tasks based only on a clear description of the task and the model's own pre - training knowledge without providing any specific examples. The principle of this strategy is based on the pre - training of large language models on large - scale data, enabling them to have extensive language understanding and knowledge reserve capabilities. They can extract relevant content from the learned knowledge and conduct reasoning and judgment according to the semantic information in the task description.
* Few - shot Prompting( Data - pattern - driven Classification Optimization):Its design idea is based on the in - depth exploration and utilization of data patterns. It aims to guide the large language model to learn the corresponding relationship between features and classification results in the data by providing a certain amount of sample data, so as to achieve more accurate classification. The advantage of few - shot prompting is that it can utilize the data patterns in the training set to provide more abundant information for the large language model, thereby improving the accuracy of classification
* Decomposed Prompting(Classification Advancement through Task Decomposition):Its core lies in carefully decomposing complex tasks into a series of more manageable sub - tasks, thus guiding the large language model to complete tasks more efficiently. In the field of natural language processing, this strategy can give full play to the advantages of large language models and improve the quality and efficiency of task completion.

Each strategy contains two sub - tasks. Task 1 is to perform multi - class classification on the data in the specified CSV file and generate a new CSV file. Task 2 (the same for the three prompts) is to evaluate he classification results using specific evaluation indicators and draw relevant charts.

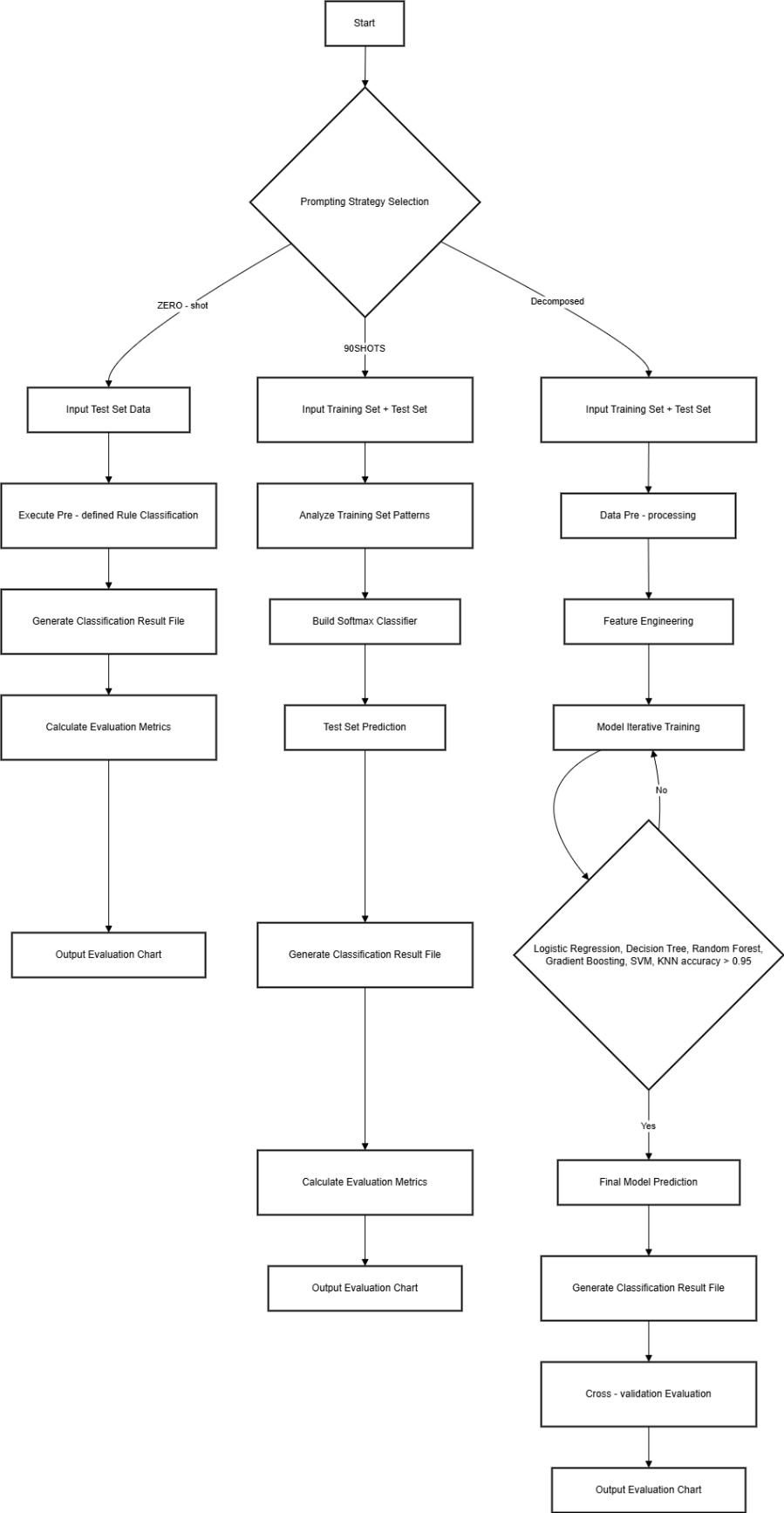


Figure 4: Flowchart of Prompting Strategies

## 4.Experimental Design and Implementation

### 4.1 Construction of the Experimental Environment

The construction of the experimental environment is the foundation to ensure the smooth progress of the research. Its hardware and software configurations have a significant impact on the accuracy and reliability of the experimental results. In this sleep disorder classification research, the selected processor for the experiment is Intel (R) Core (TM) i5 - 6400T CPU @ 2.20GHz. This processor has a main frequency of 2201 Mhz, with 4 cores and 4 logical processors. The operating system used is Microsoft Windows 10 Home Chinese Edition, with the version number 10.0.19045 (build 19045). The large language model selected is the Doubao PC version 1.41.6.。

### 4.2 Experimental Procedure

1. Division of the dataset and selection of samples:：
   * Randomly select 30 samples from each of the three categories (normal, sleep apnea, and insomnia) in the original dataset, resulting in a total of 90 samples as "prompts 90 examples". This random selection method ensures the representativeness and randomness of the samples, which can reflect the characteristics of different sleep disorder types in the dataset to a certain extent. Save these 90 samples as Sleep\_health\_and\_lifestyle\_dataset\_selected\_90.csv as the training set.
   * Delete the above 90 samples from the original dataset file and save it as Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv as the ground truth. The ground truth is an important basis for evaluating the classification accuracy of the model. By comparing the model's prediction results with the ground truth, the performance of the model can be accurately evaluated.
   * Make a copy of Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90.csv, delete the last column ("Sleep Disorder"), and save it as Sleep\_health\_and\_lifestyle\_dataset\_remaining\_90\_without\_last\_column.csv as the test set. The test set is used to evaluate the performance of the model on unseen data and can test the generalization ability and classification accuracy of the model.
2. Manually upload the above three CSV files to Doubao. This step ensures that the large language model can obtain the required data, providing a data foundation for the subsequent design, training, and evaluation of the classifier.
3. Design, train, and evaluate the classifier according to different prompting strategies in Figure 4 (details can be found in Annex 1).

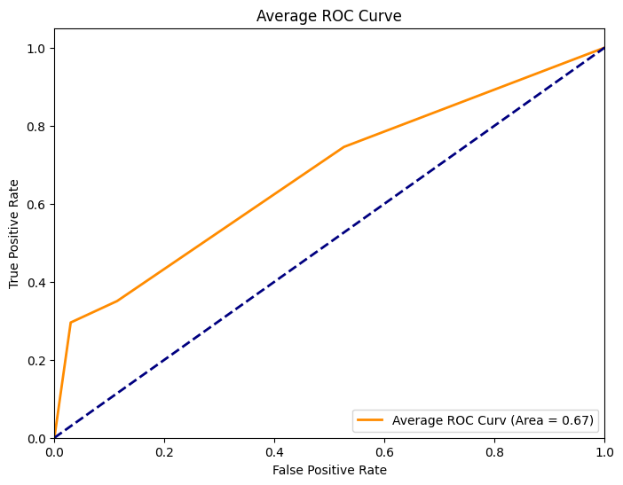
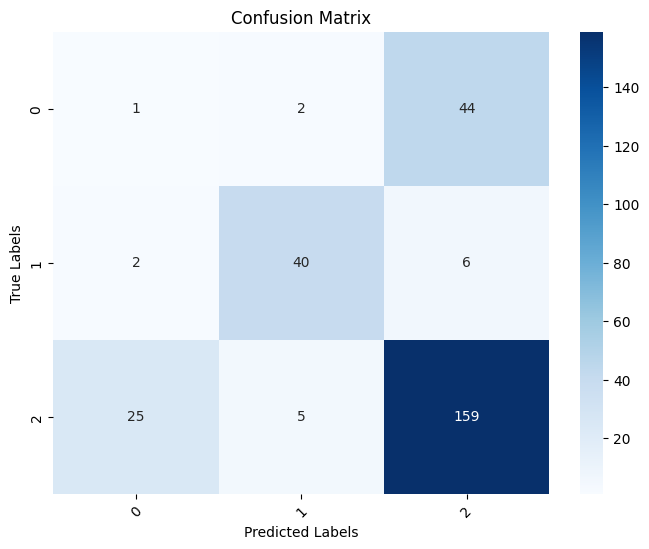
。

### 4.3 Experimental Results

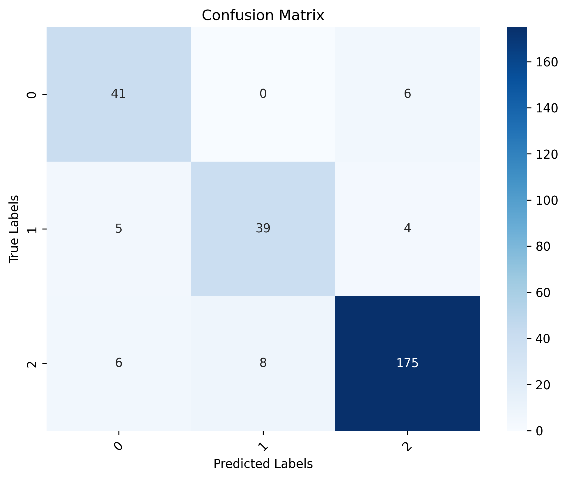
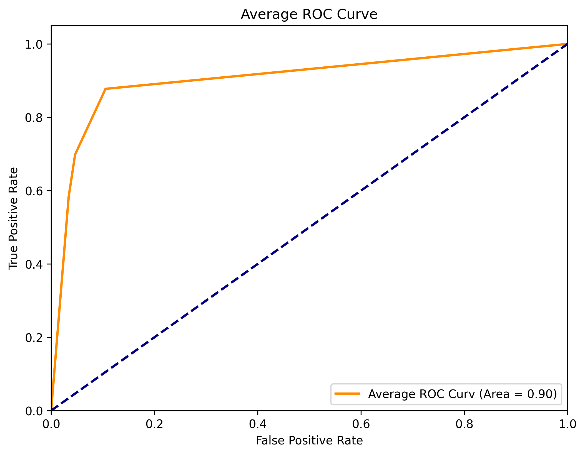
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Prompting Strategy | Accuracy | Precision | Recall | F1 Score | AUC Value |
| Zero-shot Prompting | 0.704225352112676 | 0.6560363530055422 | 0.704225352112676 | 0.6784663247048188 | 0.6709964855486846 |
| 90-sample Prompting | 0.897887323943662 | 0.9002509272025304 | 0.897887323943662 | 0.8986319612644122 | 0.9041985643947756 |
| Decomposed Prompting | 0.9190140845070423 | 0.9191754537248555 | 0.9190140845070423 | 0.9188775418205605 | 0.9163083064019824 |

To more intuitively display the classification performance of the model under different prompting strategies, we drew confusion matrix diagrams and ROC curves. (The ROC curve evaluates the classification performance of the model by showing the relationship between the true positive rate and the false positive rate of the model at different thresholds. The closer the curve is to the upper - left corner, the better the classification performance of the model.) The confusion matrix diagram shows the relationship between the model's prediction results and the ground truth in the form of a matrix, which can clearly reflect the classification accuracy and misclassification situations of the model for different categories.。

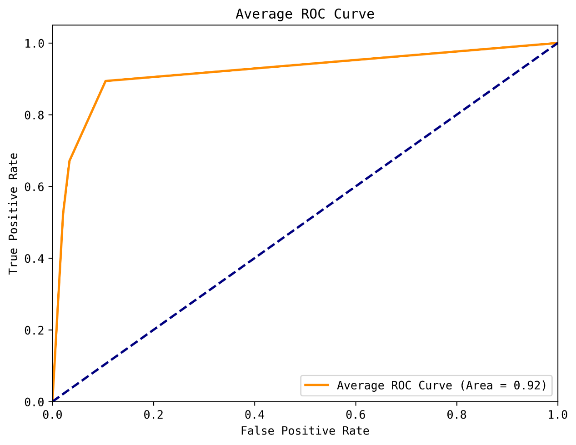
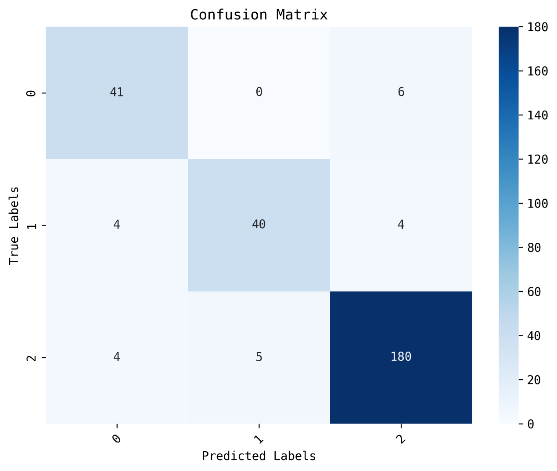
In the confusion matrix of zero - shot prompting, we can see that when the model distinguishes between the three types of sleep disorders: normal, sleep apnea, and insomnia, there are relatively many misclassification situations. A large number of normal samples are misjudged as sleep apnea or insomnia, and there is also a certain degree of misjudgment between sleep apnea and insomnia samples。



In the confusion matrix of 90 - sample prompting, the misclassification situation has been significantly improved. The number of correctly classified normal samples and sleep apnea samples has increased, and the misjudgment situation has decreased. However, there are still some misjudgment situations in the classification of insomnia samples.。

The confusion matrix of decomposed prompting shows that the classification accuracy of the model for each category has been improved. The number of correctly classified normal samples, sleep apnea samples, and insomnia samples has reached a relatively high level, with the least misclassification situations. This reflects the effectiveness and superiority of the decomposed prompting strategy in the sleep disorder classification task.。



From the drawn ROC curves, it can be seen that the curve under the decomposed prompting strategy is closest to the upper - left corner, with the largest AUC value, indicating the best classification performance. The ROC curve of the 90 - sample prompting strategy is the second - best, and the ROC curve of the zero - shot prompting strategy is the farthest from the upper - left corner, with the smallest AUC value and the worst classification performance 。

### 4.4 Analysis of Experimental Results

* The zero - shot prompting performs poorly in the sleep disorder classification task, mainly due to its limitations in data utilization and insufficient model learning ability. Zero - shot prompting classifies only according to simple predefined rules and lacks in - depth learning of the training data. In the sleep disorder classification task, the determination of sleep disorders is comprehensively affected by various factors. These factors are intertwined to form complex non - linear relationships. Zero - shot prompting cannot learn the relationships between these complex features and classification results from the training data and has difficulty capturing the potential patterns in the data, resulting in poor classification performance.
* Compared with zero - shot prompting, the 90 - sample prompting has a significantly improved classification performance. The large language model can learn the potential relationships between different feature combinations and sleep disorder types by referring to the data patterns, and the correspondence between features and classification results in the training set. For example, by analyzing the relationships between factors such as gender, age, and occupation in the training set and sleep disorder types, it can better understand the roles of these factors in classification. Moreover, when using the logistic regression model for classification, pre - processing such as encoding categorical features and standardizing numerical features can improve data quality and optimize the learning effect, making the judgment more accurate when facing common sleep disorder types. However, the logistic regression model is linear, and its classification ability is limited when dealing with data with complex non - linear relationships between features. When dealing with the complex non - linear relationships between sleep disorders and various factors, it may not be able to accurately capture these relationships, resulting in classification errors.
* The decomposed prompting strategy has achieved the best results in the sleep disorder classification task, mainly due to its effective task decomposition and optimized selection of multiple classifiers. The decomposed prompting decomposes the task, tries multiple classifiers such as logistic regression, decision tree, random forest, gradient boosting, support vector machine, and K - nearest neighbor, and performs parameter tuning. Through the comparison and screening of different classifiers, the support vector machine (SVM) is determined to be the most suitable model for this dataset. The support vector machine can find the optimal classification hyperplane in high - dimensional space and effectively handle complex non - linear relationships. In the sleep disorder classification problem, there are complex non - linear relationships between many features, such as sleep duration and sleep quality, stress level and daily activity volume. SVM can accurately capture these relationships and thus obtain the best classification performance. When judging the relationships between insomnia and factors such as stress level and living habits, SVM can accurately distinguish insomnia samples from other samples by finding the optimal classification hyperplane. The task decomposition of decomposed prompting enables the model to more deeply understand the task requirements and gradually complete the links such as data processing, model training, and evaluation, improving the classification accuracy and reliability of the model.

By analyzing and comparing the experimental results of different prompting strategies, we can see that in the sleep disorder classification task, making full use of training data, enhancing the model's learning ability, and reasonably decomposing tasks and selecting classifiers are the key factors to improve the model's performance. The decomposed prompting strategy has obvious advantages in dealing with complex sleep disorder classification problems, providing more effective methods and ideas for sleep disorder classification。

## 5.Research Conclusions and Prospects

### 5.1 Research Conclusions

In this sleep disorder classification research, we deeply explored the application of large language models on the sleep health and lifestyle dataset. Through a carefully designed experimental plan and a rigorous analysis process, we have achieved a series of results. Large language models have demonstrated significant potential in the automatic sleep disorder classification task. Different prompting strategies have a crucial impact on the performance of large language models：

1. As a simple and direct prompting strategy, zero - shot prompting performs relatively weakly in the sleep disorder classification task. This indicates that without specific examples and in - depth learning, large language models can hardly accurately cope with the complexity of sleep disorder classification relying only on pre - training knowledge and simple rules .

2. The 90 - sample prompting strategy provides the large - language model with more abundant learning information by introducing a certain amount of sample data. Compared with the zero - shot prompting, its classification performance has been significantly improved. This fully demonstrates that leveraging the patterns and corresponding relationships in the training data can enhance the large - language model's understanding and execution ability for the sleep disorder classification task. However, the 90 - sample prompting strategy still has certain limitations. Due to the linear nature of the logistic regression model it uses, it is difficult to accurately capture and handle the relationships between features when there are complex non - linear relationships in the data, thus limiting the model's classification ability.

3. The decomposed prompting strategy has achieved the most outstanding results in the sleep disorder classification task. A comprehensive attempt has been made on various classifiers, along with in - depth parameter tuning. Through evaluation and comparison, the reasonable decomposition of tasks and the selection of appropriate classifiers indicate that it can give full play to the advantages of the large - language model, improving the accuracy and reliability of sleep disorder classification.

In the research of using large - language models for sleep disorder classification, the understanding of data and tasks has a significant impact on the model's performance, which is directly related to classification accuracy. Providing detailed examples, reference information, and reasonably decomposing tasks can help improve the classification accuracy of the model. Therefore, future research should focus on in - depth analysis of data and tasks, and optimize the prompting strategy to enhance the model's performance. At the same time, in the process of decomposed prompting, choosing an appropriate model according to the task characteristics and optimizing it is the core point for improving the classification performance. Since different classifiers are suitable for different scenarios, when facing complex sleep disorder classification tasks, it is necessary to comprehensively consider various factors, accurately select the best classifier, and carefully adjust the parameters to achieve the optimal classification effect.。

### 5.3 Research Limitations and Future Directions

Although this research has achieved certain results in sleep disorder classification, there are still limitations in aspects such as data, model, experimental design, and prompt engineering. For example, the data lacks diversity [15], the selection of the classifier model does not utilize the powerful feature - learning ability of the latest deep learning, the comparative experiments are insufficient, there is no external verification, and the optimization of the prompting strategy is inadequate. Future research can be carried out in multiple directions mentioned above to further improve the performance and application value of large - language models in the automatic sleep disorder classification task 。

* Expanding the dataset is an urgent task. It is necessary to widely collect samples of various sleep disorders, especially data on rare or special types of sleep disorders. At the same time, it is crucial to address the issue of unbalanced data distribution. Data resampling techniques such as oversampling (SMOTE algorithm) and undersampling [16] can be used to adjust the proportion of different types of samples in the dataset, enabling the model to learn the characteristics of various sleep disorders more evenly during the training process, thus improving the model's generalization ability for specific groups.
* In terms of improving the classifier model, it is essential to explore more advanced architectures and feature - engineering methods. Deep - learning models have powerful feature - learning capabilities [17], such as convolutional neural networks (CNN) [18], recurrent neural networks (RNN) [19], and their variants long - short - term memory networks (LSTM) [20], gated recurrent units (GRU) [21], etc. These models have advantages in processing time - series data and complex data structures. They can automatically learn the features and patterns in these data, and better explore the potential relationships between sleep disorders and difficult - to - quantify factors.
* In terms of improving the experimental design, more comprehensive comparative experiments and external verification should be carried out. Comprehensively comparing the impacts of different model architectures (including LLM and classification models), prompting strategies, and other factors on the classification performance can help us deeply understand the action mechanisms of various factors and find the most suitable methods and parameter settings for the sleep disorder classification task. By testing on multiple external datasets from different sources, the generalization ability of the model under different data distributions and scenarios can be evaluated, and the problem of overfitting can be avoided.
* In terms of optimizing the prompting strategy, more complex prompting methods can be explored, such as Chain of Thought Prompting [22], Contextual Prompting [23], etc., to guide the model to conduct more in - depth reasoning and analysis. Chain of Thought Prompting can enable the model to gradually display its reasoning process when answering questions, improving the interpretability of the model [24]. Contextual Prompting can provide the model with more background information and context according to the specific situation of the task, helping the model better understand the task requirements and thus more accurately classify sleep disorders.

6.Acknowledgments

During the process of completing this research, I have received help from many individuals and institutions. Here, I would like to express my sincere gratitude to them。

I would like to wholeheartedly thank my supervisor. Throughout the research process, with profound academic attainments and rich experience, my supervisor has pointed out the research direction for me. When I encountered confusion and difficulties in the research, my supervisor always listened to my questions patiently and provided me with valuable suggestions and guidance. During the analysis of the dataset and the design stage of the prompting strategy, my supervisor's professional insights gave me a deeper understanding of the research and helped me avoid many detours. My supervisor's rigorous academic attitude and persistent pursuit of scientific research have always inspired me to keep moving forward and served as a role model on my academic path.。

I also want to thank the data providers. They have provided us with the valuable sleep health and lifestyle dataset, which is the foundation of the entire research。

On the future academic path, I will continue to work hard, keep exploring, and contribute my efforts to the development of the field of sleep medicine。

7.References

[1] Ibomoiye Domor Mienye, N., Jere, N. (Year Unknown). Survey of Decision Trees: Concepts, Algorithms, and Applications [Literature Type]. In IEEE Xplore . <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10562290>  
[2] Kumari, A., Akhtar, M., Shah, R., et al. Support matrix machine: A review[EB/OL]. <https://arxiv.org/abs/2310.19717>.  
[3] Curth, A., Jeffares, A., van der Schaar, M. Why do Random Forests Work? Understanding Tree Ensembles as Self-Regularizing Adaptive Smoothers[EB/OL]. <https://arxiv.org/abs/2402.01502>.  
[4] Kim, Y., Xu, X., McDuff, D., et al. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data[EB/OL].[Submitted on 12 Jan 2024 (v1), last revised 27 Apr 2024 (this version, v2)].  
[5] Katz, D. M., Bommarito, M. J., Gao, S., et al. GPT-4 passes the bar exam[J]. Philosophical Transactions of the Royal Society A, 2024, 382(2270):20230254.  
[6] Nori, H., Lee, Y. T., Zhang, S., et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine[EB/OL]. arXiv preprint arXiv:2311.16452, 2023.  
[7] Saab, K., Tu, T., Weng, W.-H., et al. Capabilities of Gemini models in medicine[EB/OL]. arXiv preprint arXiv:2404.18416, 2024.  
[8] Singhal, K., Tu, T., Gottweis, J., et al. Towards expert-level medical question answering with large language models[EB/OL]. arXiv preprint arXiv:2305.09617, 2023.  
[9] McDuff, D., Schaekermann, M., Tu, T., et al. Towards accurate differential diagnosis with large language models[EB/OL]. arXiv preprint arXiv:2312.00164, 2023.  
[10] Sleep Health and Lifestyle Dataset[EB/OL]. <https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset>.  
[11] Wang, G., Zhao, W., Han, J., et al. MedFound: The First Medical Large Language Model Passing the Physician Qualification Examination [J]. Journal of Artificial Intelligence, 2024, 5(1): 1 - 12.  
[12] McDuff, D., Xu, X., Kim, Y., et al. Personal Health Large Language Model (PH-LLM): leveraging large language models for personalized health insights [EB/OL]. arXiv preprint arXiv:2311.17133, 2023.  
[13] Zhang, Y., Maziarka, P., Klicpera, J., et al. DiffSBDD: Equivariant diffusion for structure-based drug design [EB/OL]. arXiv preprint arXiv:2403.14338, 2024.

[4]Vaswani, A., Shazeer, N., Parmar, N., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017).

[15] Liu, X., Uchiyama, M., Okawa, M., et al. Prevalence and correlates of insomnia in the Japanese general population: Results from the Japan epidemiological sleep study[J]. Sleep, 2000, 23(4):497 - 506.  
[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over - sampling Technique[J]. J Artif Intell Res, 2002, 16:321 - 357.  
[17] LeCun, Y., Bengio, Y., Hinton, G. Deep Learning[J]. Nature, 2015, 521(7553):436 - 444.  
[18] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Gradient - Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278 - 2324.  
[19] Graves, A., Mohamed, A. - R., Hinton, G. Speech Recognition with Deep Recurrent Neural Networks[C]. In Proceedings of the 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645 - 6649), 2013.  
[20] Hochreiter, S., Schmidhuber, J. Long Short - Term Memory[J]. Neural Comput, 1997, 9(8):1735 - 1780.  
[21] Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches[EB/OL]. arXiv preprint arXiv:1409.1259, 2014.  
[22] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichtertz, R., Joshi, S., Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[EB/OL]. arXiv preprint arXiv:2201.11903, 2022.  
[23] Zhou, X., Huang, M., Wang, H., Zhang, Z. Contextual Prompting for Few - Shot Text Classification[C]. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 9312 - 9327), 2022.  
[24] Ribeiro, M. T., Singh, S., Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier[C]. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1135 - 1144), 2016.