*engineering proceedings*

MDPI

*Proceeding Paper*

# Exploring Sleep Apnea Risk Factors with Contrast Set Mining: Findings from the Sleep Heart Health Study †

**Nhung H. Hoang * and Zilu Liang**

Graduate School of Engineering, Kyoto University of Advanced Science, Kyoto, Japan; zilu.liang@kuas.ac.jp

* Correspondence: 2023md05@kuas.ac.jp

† Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

**Abstract:** Sleep apnea is a common sleep disorder with potentially serious health consequences. Identifying risk factors for sleep apnea is crucial for early detection and effective management. Traditionally, this has been achieved through statistical methods such as Pearson's and Spearman's correlation analysis, which examine relationships between individual variables and sleep apnea. However, these methods often miss complex, nonlinear patterns and interactions among multiple factors. In this study, we applied contrast set mining to identify patterns in attribute-value pair combinations (contrast sets) in the Sleep Heart Health Study database that differentiate between groups with varying levels of sleep apnea severity. Our findings reveal that males and individuals aged 60 to 80 exhibit a higher risk of sleep apnea, with a confidence exceeding 75%. Moreover, male patients diagnosed with second-degree obesity, defined as a body mass index (BMI) between 35 and 39.9 kg/m2, show an elevated risk of severe apnea, with a lift of over 2.23, support over 16%, and confidence around 80%. In contrast, female patients with a BMI within the normal range (18–25 kg/m2) demonstrate a lower risk of sleep apnea, with a lift of 2.36, support of 17%, and confidence exceeding 90%. Contrast set mining helps uncover meaningful rules within subgroups that traditional methods, such as Pearson's or Spearman's correlation analysis, might overlook. Future research will focus on developing sleep apnea screening models based on these identified contrast set rules.

**Keywords:** contrast set mining; data mining; descriptive statistics; sleep health

**Citation:** Hoang, N.H.; Liang, Z. Exploring Sleep Apnea Risk Factors with Contrast Set Mining: Findings from the Sleep Heart Health Study. *Eng. Proc.* **2024**, *5*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Published: 26 November 2024



## 1. Introduction

This paper presents an approach for contrast set mining technique that might be meaningful in identifying sleep apnea patients. Contrast set mining can be defined as finding population subgroups that are statistically interesting (as large as possible and have the most unusual distributional characteristics) with respect to the property of interest [1]. Instead of defining an optimal measure for automated subgroup search and selection, the goal is to support the expert in performing flexible and effective search of a broad range of optimal solutions. Contrast set mining has been applied to uncover distinguishing characteristics of two groups of brain stoke patients [2], or to early detect atherosclerotic coronary heart disease [1]. In this study, contrast set mining is first used to explore sleep apnea risk factors.

The problem of sleep apnea detection is one of the hot topics for sleep health research due to it high prevalence and is still on the rise [3]. Previous studies on sleep apnea have primarily focused on epidemiological data using statistical tests, prevalence rates, and summary figures [4–7]. Few have explored feature combinations for detection. Machine learning solved the problem by enabling fast, accurate prediction using combination of multiple demographic factors [8,9]. However, improving model interpretability remains a challenge. Traditional methods, such as feature ranking and correlation analysis (e.g., Pearson, Spearman), highlight individual feature importance but fail to show how

features interact when combined together. In this study, contrast set mining is employed to reveal how demographic features combine for sleep apnea detection, supported by statistical evidence.

The contribution of this study is as follow:

- We designed a data mining pipeline based on the contrast set mining algorithm instead of traditional statistical methods. Different from correlation analysis or linear regression, our method allows the identification of associations that only manifest when the metrics are within certain value ranges.
- The data mining pipeline generates interesting rules and hypotheses that may help inform the design of future studies to deepen our understanding of the relationships between sleep and risk factors.
- The extracted rules are highly interpretable and can effectively guide data visualization efforts.

## 2. Materials and Methods

### *2.1. Dataset*

The Sleep Heart Health Study (SHHS) dataset [10,11] is a key resource in sleep research, aimed at investigating the relationship between sleep-disordered breathing and cardiovascular disease. Collected from over 6000 participants across the U.S., it includes comprehensive polysomnographic and demographic data. The dataset is stored in the shhs-harmonized-dataset-0.21.0.csv file, encompassing both demographic and sleep apnea severity data.

### *2.2. Data Preparation*

In this study, risk factors including age, body mass index, arousal index, gender, race, ethnicity, smoking status (current and past), systolic and diastolic blood pressure were utilized for contrast set mining. Most of the data used in this study is demographic and can be easily collected during a patient's clinic visit. Additionally, the arousal index, which measures the average number of arousals during sleep, is included. This data can be conveniently gathered using modern smart devices such as smartphones or smartwatches, making it accessible in both clinical and home settings.

Discrete variables such as gender and smoking status were maintained in their original categorical form. Continuous variables were discretized into subgroups, either following established clinical guidelines (e.g., BMI, systolic and diastolic blood pressure) or using user-defined thresholds (e.g., age, arousal index), to meet the requirements of contrast set mining, which operates on distinct subpopulations for analysis and comparison. Table 1 provides a summary of the data utilized in this study from both SHHS1 and SHHS2 datasets after preparation steps.

**Table 1.** Contrast set rules associated with normal people (Apnea-hypopnea index < 5 e/h).

| Contrast Set Rules | | | Sleep Apnea Severity | Lift | Support (%) | Confidence (%) |
|---|---|---|---|---|---|---|
| Arousal index=>[0, 20) | BMI=>normal | Age=>(40, 60] | Normal | 3.05 | 14.13 | 90.29 |
| Arousal index=>[0, 20) | BMI=>normal | BP_categories=>Normal | Normal | 2.88 | 11.74 | 88.33 |
| Arousal index=>[0, 20) | BMI=>normal | Ever_smoker=>no | Normal | 2.51 | 12.64 | 88.89 |
| Arousal index=>[0, 20) | BMI=>normal | Gender=>female | Normal | 2.36 | 17.00 | 95.40 |
| Arousal index=>[0, 20) | BMI=>normal | | Normal | 2.27 | 14.15 | 89.74 |
| Arousal index=>[0, 20) | Age=>(40, 60] | BP_categories=>Normal | Normal | 2.21 | 14.29 | 90.30 |
| Arousal index=>[0, 20) | Age=>(40, 60] | BMI=>overweight | Normal | 2.09 | 11.65 | 88.05 |
| Gender=>female | BMI=>normal | Ever_smoker=>no | Normal | 2.13 | 13.29 | 79.29 |
| BMI=>normal | Ever_smoker=>yes | Gender=>female | Normal | 2.13 | 13.29 | 79.29 |

The dataset is predominantly composed of individuals identified as "White" and "Not Hispanic or Latino" (at 85.57% and 95.39% respectively). To minimize bias during contrast set discovery, we opted to retain only the majority group and excluded minority groups from the analysis. This approach helps to reduce skewed results due to unequal group representation.

*2.3. Contrast Set Mining*

The goal of contrast set mining is to identify rules that highlight meaningful differences between groups, defined by metrics such as support, confidence, and lift. The STUCCO algorithm (Search and Testing for Understandable Consistent Contrasts) [12], introduced in the foundational contrast set mining study, facilitates the interpretation of large datasets by uncovering statistically significant contrasts across subpopulations.

Contrast set mining generates logical rules that consist of antecedents (input conditions) on the left-hand side and a consequent (outcome) on the right-hand side. For a given dataset D, let $Y = \{Y_1, Y_2, \ldots, Y_n\}$ represent a set of consequents, and $X = \{X_1, X_2, \ldots, X_n\}$ represent a set of antecedents, which may contain one or more attributes. The quality of these rules is evaluated using several metrics:

- Support: This is the proportion of instances in D that contain both X1∪Y1. It reflects how frequently the rule occurs in the dataset.
- Confidence: This is the proportion of instances with X1 that also exhibit Y1, indicating the predictive power of the rule.
- Lift: Lift measures the strength of the association, with values greater than 1 indicating a positive association between antecedents and consequents.
- The output is expected to have the following format: $(X_1\ AND\ X_2) \rightarrow Y_1$

To ensure interpretability and avoid overfitting, long rules with numerous antecedents are often discouraged. In this study, rule length was limited to four features, as longer rules may become difficult to interpret and can lead to redundancy. A contrast set is considered valid if it meets the following criteria: support ≥ 10%, confidence ≥ 75%, and lift ≥ 2.

Once the contrast sets are generated, rules that share similar attributes are grouped together to enhance interpretability. This grouping enables a clearer understanding of how different combinations of features contribute to the detection or classification process, as showed in Tables 1 and 2.

**Table 2.** Contrast set rules associated with severe apnea patients (Apnea-hypopnea index > 30 e/h).

| Contrast Set Rules | | | | Sleep Apnea Severity | Lift | Support (%) | Confidence (%) |
|---|---|---|---|---|---|---|---|
| Current_smoker=>no | Gender => male | BMI => class 2 obesity | Age=>(60, 70] | Severe apnea | 2.94 | 16.51 | 77.78 |
| Current_smoker=>no | Gender => male | BMI => class 2 obesity | | Severe apnea | 2.23 | 19.25 | 80.72 |
| Current_smoker=>no | Gender => male | BMI => class 1 obesity | Age=>(70, 80] | Severe apnea | 2.24 | 16.28 | 82.35 |

## 3. Results and Discussion

In contrast set mining, an increase in the number of rule components typically results in more rules. However, complex rules with numerous elements are harder to interpret, so this study limits rule length to four components or fewer. Most rules focus on distinguishing severe and normal sleep apnea against other groups.

Regarding the normal group, among the rule components in Table 1, the arousal index is the most frequent, which aligns with prior research. Sleep apnea episodes disrupt sleep continuity, causing frequent arousals, particularly in severe cases, highlighting the

importance of the arousal index in sleep apnea detection. The contrast sets reveal that when the arousal index is below 20 events/h and BMI is within the normal range, it is typically associated with a healthy population. Adding the age factor (within 40–60 years) increases confidence to over 90%, with support of 14.1% and a lift of 3.05. Combining arousal index, BMI, and blood pressure slightly reduces these indicators, suggesting age has more influence than blood pressure. Additionally, the female gender contributes significantly to the normal population, with higher support and confidence but a lower lift due to the larger size of the female group compared to the 40–60 age group.

In the severe sleep apnea group, factors such as male gender, age over 60, and a BMI above 30 (indicative of obesity) were key determinants. Most rules associating male sex or class 2 obesity with severe apnea demonstrated confidence levels exceeding 75% which consisten with perivous studies [3,13,14]. While smoking status appeared frequently in the rules presented in Table 2, this reflects the dataset's non-smoker majority rather than a causal relationship. Smoking is not a well-established risk factor for obstructive sleep apnea (OSA) and its presence in the rules should not be misinterpreted as influencing OSA risk [6].

The advantage of contrast set mining is that it not only clarifies causal relationships but also simplifies data visualization. By identifying small subgroups with statistically significant differences, contrast set mining allows for focused analysis of key groups without extensive effort. This method enables quick identification of variations within the dataset, which can be effectively illustrated with simple visual aids, as demonstrated in Figure 1. Using contrast set rules as a guide enhances both the interpretation and visual representation of complex data patterns.

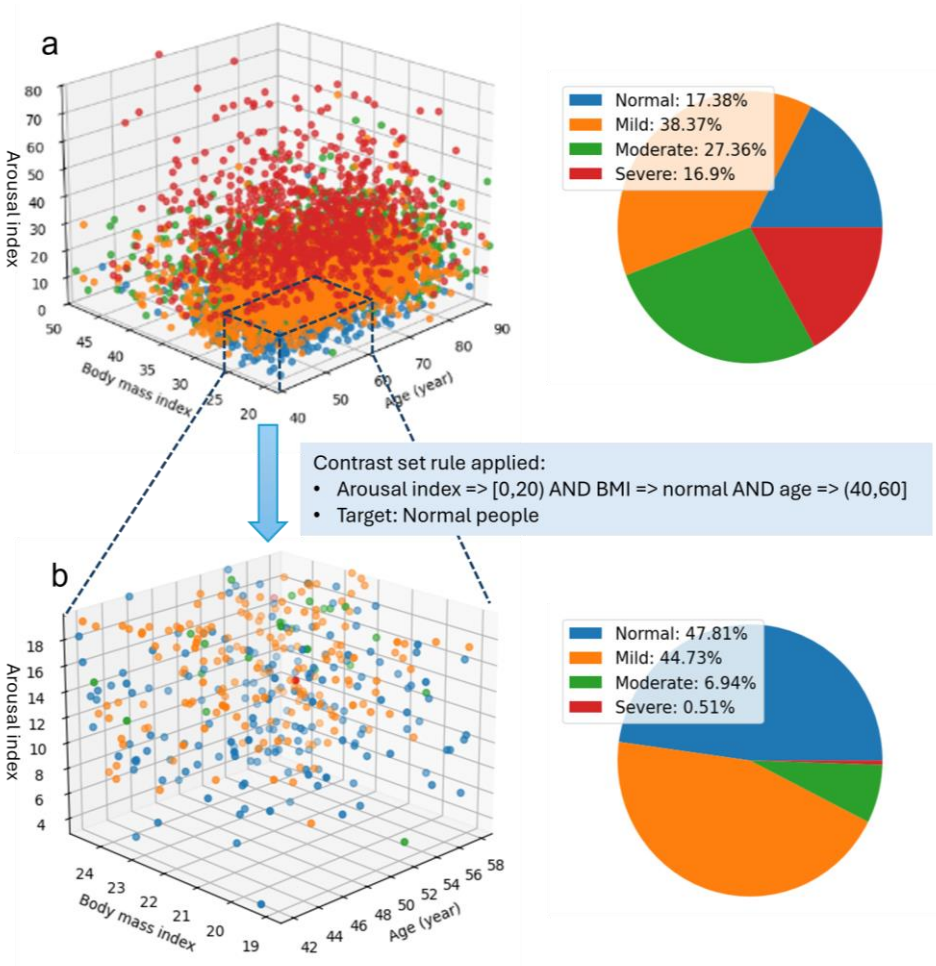

**Figure 1.** illustration of the discovery of a smaller group within the dataset with unusual distributional characteristics compared to the whole group distribution.

Figure 1 illustrates the visualization of the first contrast set from Table 1. In Figure 1a, the overall distribution of more than 5000 subjects based on arousal index, BMI, and age shows overlapping severity levels, making it difficult to discern specific patterns. The pie chart shows a relatively balanced distribution of sleep apnea severity levels, with the *Mild* group comprising 38.37% and the *Severe* group the smallest at 16.9%. However, when analyzing specific components within the contrast set, a significant shift in group percentages is observed. In this context, the *Severe* group represents a noticeable low percentage (under 1%), with over 93% of subjects classified as either normal or mild. Furthermore, individuals aged 40–60, presenting with normal BMI and low arousal, exhibit a notably reduced risk of developing moderate or severe sleep apnea.

## 4. Conclusions

In this study, we demonstrated the efficacy of contrast set mining in identifying associations between sleep apnea and its associated risk factors. The derived rules align with previous findings, providing partial validation of the method's reliability. A key advantage of contrast set mining is its capacity to elucidate the relationships between sleep apnea and multiple risk factors simultaneously, as opposed to relying on pairwise comparisons such as Pearson or Spearman correlation coefficients. However, this advantage comes with the caveat that computational complexity may increase significantly when considering a large number of risk factors. Additionally, utilizing contrast set rules facilitates enhanced interpretation and visual representation of intricate data patterns.

Despite its advantages, contrast set mining has certain limitations. Firstly, it is a method of descriptive rule induction that identifies specific characteristics of the dataset under investigation. The statistical indices employed to derive these rules may not adequately represent a broader population. Secondly, as the length of the rules increases, the quantity of generated rules can also rise, leading to potential duplication. Therefore, a post-processing method is essential to eliminate redundant rules and retain only the most significant and meaningful ones. Thirdly, the algorithm necessitates that data be segmented into specific groups. This segmentation is straightforward for categorical data or data with established standards, such as BMI and blood pressure. However, for data lacking standardized measures, such as the arousal index, segmentation relies on the subjective judgment of the user. Consequently, the output can vary significantly based on the chosen segmentation method. Thus, there is a critical need to develop algorithms that optimize data segmentation in such contexts.

**Author Contributions:** N.H.H. contributed to study design, analysis and interpretation of results. N.H.H. and Z.L. contributed to study conception, manuscript drafting and revision. All authors have reviewed the results and approved the final version of the manuscript.

**Funding:** This study was sponsored by a Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Early-Career Scientists (grant 21K17670). The funder had no role in the study design, data collection and analysis, or manuscript preparation.

**Institutional Review Board Statement:**

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:**

**Acknowledgments:** The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve

University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gamberger, D.; Lavrac, N. Expert-guided subgroup discovery: Methodology and application. *J. Artif. Intell. Res.* **2002**, *17*, 501–527.
2. Novak, P.K.; Lavrač, N.; Gamberger, D.; Krstačić, A. CSM-SD: Methodology for contrast set mining through subgroup discovery. *J. Biomed. Inform.* **2009**, *42*, 113–122.
3. Franklin, K.A.; Lindberg, E. Obstructive sleep apnea is a common disorder in the population—A review on the epidemiology of sleep apnea. *J. Thorac. Dis.* **2015**, *7*, 1311.
4. Young, T.; Peppard, P.E.; Taheri, S. Excess weight and sleep-disordered breathing. *J. Appl. Physiol.* **2005**, *99*, 1592–1599.
5. Bixler, E.O.; Vgontzas, A.N.; Ten Have, T.; Tyson, K.; Kales, A. Effects of age on sleep apnea in men: I. Prevalence and severity. *Am. J. Respir. Crit. Care Med.* **1998**, *157*, 144–148.
6. Wetter, D.W.; Young, T.B.; Bidwell, T.R.; Badr, M.S.; Palta, M. Smoking as a risk factor for sleep-disordered breathing. *Arch. Intern. Med.* **1994**, *154*, 2219–2224.
7. Newman, A.B.; Foster, G.; Givelber, R.; Nieto, F.J.; Redline, S.; Young, T. Progression and regression of sleep-disordered breathing with changes in weight: The Sleep Heart Health Study. *Arch. Intern. Med.* **2005**, *165*, 2408–2413.
8. Rodrigues Jr, J.F.; Pepin, J.-L.; Goeuriot, L.; Amer-Yahia, S. An extensive investigation of machine learning techniques for sleep apnea screening. In Proceedings of the Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual, 19–23 October 2020; pp. 2709–2716.
9. Shi, Y.; Zhang, Y.; Cao, Z.; Ma, L.; Yuan, Y.; Niu, X.; Su, Y.; Xie, Y.; Chen, X.; Xing, L. Application and interpretation of machine learning models in predicting the risk of severe obstructive sleep apnea in adults. *BMC Med. Inform. Decis. Mak.* **2023**, *23*, 230.
10. Zhang, G.-Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1351–1358.
11. Quan, S.F.; Howard, B.V.; Iber, C.; Kiley, J.P.; Nieto, F.J.; O'Connor, G.T.; Rapoport, D.M.; Redline, S.; Robbins, J.; Samet, J.M. The sleep heart health study: Design, rationale, and methods. *Sleep* **1997**, *20*, 1077–1085.
12. Bay, S.D.; Pazzani, M.J. Detecting change in categorical data: Mining contrast sets. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 15–18 August 1999; pp. 302–306.
13. Bearpark, H.; Elliott, L.; Grunstein, R.; Cullen, S.; Schneider, H.; Althaus, W.; Sullivan, C. Snoring and sleep apnea. A population study in Australian men. *Am. J. Respir. Crit. Care Med.* **1995**, *151*, 1459–1465.
14. Strohl, K.P.; Redline, S. Recognition of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.* **1996**, *154*, 279–289.