Presented at the IEEE International Conference on Machine Learning Applications, ICMLA 05, Los Angeles, Dec. 2005.

Integrating Knowledge-Driven and Data-Driven Approaches for the Derivation of Clinical Prediction Rules

M. Kwiatkowska
Department of Computing Science
Thompson Rivers University
mkwiatkowska@tru.ca

M. S. Atkins
Department of Computing Science
Simon Fraser University
stella@cs.sfu.ca

N. T. Ayas
Faculty of Medicine
University of British Columbia
nayas@vanhosp.bc.ca

C. F. Ryan
Faculty of Medicine
University of British Columbia
fryan@interchange.ubc.ca

Abstract

Clinical prediction rules are created by medical researchers and practitioners based on their knowledge and clinical experience. Such expert-generated rules are then evaluated and refined in clinical tests. Once verified, these knowledge-driven rules are used to expedite diagnosis and treatment for the serious cases and to limit unnecessary tests for low-probability cases. Alternatively, machine learning techniques can be used for automated induction of comprehensible data-driven rules from vast amount of existing clinical data. This paper investigates how the rules generated by the clinical experts compare with the data-driven rules. The paper describes three outcomes: rule confirmation, contradiction, and expansion. The study concentrates on prediction rules for the diagnosis of obstructive sleep apnea using three clinical data sets with 1,318 records. The prototype system, Hypnos, includes both a framework for rule definition, and also a mechanism for rule induction.

1. Introduction

The development process for clinical prediction rules involves derivation, validation, and evaluation in clinical settings. The derivation is a demanding task requiring several refinements and clinical tests using standard statistical methods. This study demonstrates that machine learning techniques can support the derivation of the rules. This paper describes a framework for a unified rule definition and a mechanism for two-way rule generation: (1) from hypotheses to data and (2) from data to hypotheses. The former, leading from human generated hypotheses to tests on data, is based on the clinical experience of medical experts. The second approach is based on machine learning techniques, generating hypotheses from the data sets. The machine-generated rules are interpreted and compared with the human generated hypotheses. This interactive process has an exploratory and confirmatory purpose: it allows for the discovery of new patterns from data and provides confirmation or contradiction of hypothetical rules.

This paper focuses on the application of clinical prediction rules in the diagnosis of obstructive sleep apnea (OSA). Section 2 provides a brief introduction to OSA and its diagnostic criteria. Section 3 discusses the semiotic framework for rule representation. Section 4 describes the data sets. Section 5 presents the methods and experimental results. The last section provides the conclusion and the directions for future work.

2. Diagnosis of Obstructive Sleep Apnea

Obstructive sleep apnea is a common, serious respiratory disorder afflicting approximately 2-4% of the population. OSA is caused by collapse of the soft tissues in the throat as the result of the natural relaxation of muscles during sleep. The soft tissue blocks the air passage and the sleeping person literally stops breathing (apnea event) or experiences a partial obstruction (hypopnea event). Apnea occurs only during sleep and is, therefore, a condition that might go unnoticed for years. The gold standard for the diagnosis of OSA is an overnight in-laboratory polysomnography (PSG) study involving continuous recordings of EEG, ECG, EOG, EMG, airflow, breathing effort, and oxygen saturation. OSA is associated with hypertension, congestive heart failure, stroke, and coronary artery disease. Although the diagnosis of OSA using PSG is relatively straightforward, and treatment is readily available, a large segment of the population is not diagnosed because of time factors, costs, and limited access to the overnight in-clinic PSG. Therefore, patients suffering from OSA might spend several months waiting for diagnosis. However, we believe that by using a combination of predictive rules and home studies, early treatment can be initiated in appropriate patients before formal diagnosis by PSG.

The diagnosis of OSA uses two approaches: (1) a score of apnea/hypopnea events and (2) a combination of scoring and symptoms. Both approaches use an apnea-hypopnea index (AHI), calculated as a number of apnea and hypopnea events per hour of sleep [1]. An apnea is defined as a complete cessation of airflow for at least 10 seconds. A hypopnea is defined using various criteria consisting of one or more of the following three factors: partial reduction of airflow, oxyhemoglobin desaturation, and brief arousals from sleep. In the diagnosis based solely on the AHI index, apnea is classified as mild for AHI between 5 and 14.9, moderate for AHI between 15 and 29.9, and severe for AHI ≥ 30. However, the use of diverse scoring criteria for AHI calculations can result in significant differences in apnea diagnoses, especially for patients with low AHI scores [2,3]. Furthermore, the difficulty with the scoring of AHI is compounded by (1) natural night-to-night variations and (2) differences in diagnostic equipment.

3. Framework for rule representation

A knowledge representation framework defines two essential diagnostic concepts: prediction rules and predictors. The concepts are defined at three levels: syntactic, semantic, and pragmatic.

3.1. Prediction rules

The clinical prediction rule (CPR) is specified by an IF-THEN statement, a certainty factor, and usability. We define CPR as a triplet: < RS, CF, U >. The rule statement, RS, represents the rule's syntax, the rule certainty factor, CF, is a part of the rule's semantics, and the usability, U, determines the rule's pragmatic value. The rule pragmatic value is an important criterion introduced by us to describe how the rule can be used in a clinical setting.

3.1.1. Rule syntax. The rule is comprised of two parts: a premise and a consequent. The premise of the rule uses predefined predictors, for example, age, gender, neck circumference, or hypertension. A proposition is a logical expression composed of a predictor variable, the relational operator (<, ≤, >, ≥, =), and a value; for example, age > 65, hypertension = yes. The rules are in the conjunctive propositional form, for example, age > 65 AND gender = female. The conclusion of the rule includes the class label. The rule statement is defined in extended BNF grammar, as follows:

```
<Rule statement> ::=  IF <Rule premise> THEN <Rule
   consequent>
<Rule premise> ::=  <Relational expression> {AND
   <Relational expression>}
<Relational expression> ::= <Predictor variable>
   <Relational operator> <Value>
<Relational operator> ::= < | ≥ | > | ≤ | =
<Value> ::=  numerical value | categorical value
<Rule consequent> ::= class label
```

3.1.2. Rule semantics. The clinical prediction rule is a hypothetical statement with two functions: descriptive and predictive. In the descriptive sense, rules characterize the subpopulations of patients with higher or lower risks for the disease. In the predictive sense, rules assess the probability of a new patient belonging to one of the classes. The hypothetical quality of the rule is defined by the certainty factor (CF), a degree of belief ranging from -1.0 (absolute disbelief) to +1.0 (absolute belief), assigned to the rule by medical experts based on their clinical experience.

3.1.3. Rule pragmatics. The rule's pragmatic value is determined by three criteria: internal validity, external validity, and clinical usability. Internal validity is based on specificity and sensitivity. The external validity is based on rule generality: transferability to a different data set. The clinical usability comprises interpretability, simplicity, and practicality. The rule interpretability and practicality are qualitatively determined by the medical experts. The rule simplicity is measured, for example, by the length of the rule.

3.2. Predictors

A predictor is an established or suspected symptom, sign, correlate, or co-morbid condition. In general, OSA predictors are divided into six categories: (1) anatomical signs: obesity, large neck circumference, and high Mallampati score, (2) nocturnal symptoms: snoring, breathing pauses, and choking, (3) diurnal symptoms:

excessive daytime sleepiness, (4) demographic factors: gender, age, and familial aggregation, (5) coexisting medical conditions: hypertension and coronary artery disease; and (6) lifestyle factors: smoking and alcohol use [4,5].

Predictors are described at three levels: semantic (conceptualization), syntactic (operationalization), and pragmatic (utilization of measurements). For example, in our study, hypertension is defined as blood pressure BP ≥ 140/90 mmHg, or current treatment with antihypertensive medications. Thus, the concept of hypertension can be represented syntactically by (1) categorical binary values: yes/no, (2) continuous values for systolic and diastolic blood pressure combined with the indicator of the current treatment, or (3) ordinal values: *low, normal*, *high normal*, *high*, and *severe high*.

This study investigates six predictors: age, gender, hypertension (HTN), body mass index (BMI) in kg/m^2, neck circumference (NC) in cm, and Mallampati score (MP). The Mallampati score is determined based on visual inspection of the wide open patient's mouth. The scale ranges from I to IV: I – entire uvula visible, II – majority of uvula visible, III – only soft palate visible, IV – only hard palate visible. Clinical studies [6] show correlation between the score and obstructive sleep apnea (OSA).

4. Clinical data

Three data sets A (N=795), B (N=233), and C (N=290) were obtained from the Sleep Disorders Program, Vancouver Acute Hospitals. All patients were diagnosed based on standard in-clinic overnight PSG.

Set A (795) has four attributes: gender, age, BMI, and hypertension; 539 males and 256 females; mean age of 50.4 years (STD=12.4), mean BMI of 31.9 (STD=7.6), and 241 instances of hypertension (HTN=yes).

Set B (233) has five attributes: gender, age, BMI, neck circumference, and Mallampati score; 193 males and 40 females; mean age of 49.2 (STD=12.1) and mean BMI of 29.2 (STD=5.75).

Set C (290) has four attributes: gender, age, BMI, and hypertension; 210 males and 80 females; mean age of 49.2 (STD=12.1), mean BMI of 31.2 (STD=6.6), and 55 instances of hypertension (HTN=yes).

Since the data sets include solely clinical records, they are biased towards the positive instances of OSA. However, the prevalence of OSA in sets A and B depends strongly on the AHI cut-off values. The changes in prevalence are illustrated by table 1. In our study, we use AHI ≥ 15 to define OSA, since this value typically indicates clinically important OSA requiring treatment. The records with AHI < 15 are classified as non-OSA.

Table 1. OSA prevalence

	Prevalence of OSA based on AHI cut-off values		
	AHI ≥ 5	AHI ≥ 10	AHI ≥ 15
Set A (795)			
OSA = yes	91.9% (731)	78.6% (625)	65.8% (523)
OSA = no	8.1% (64)	21.4% (170)	34.2% (272)
Set B (233)			
OSA = yes	83.7% (195)	69.5% (162)	58.8% (137)
OSA = no	16.3% (38)	30.5% (171)	41.2% (96)
Set C (290)			
OSA = yes	83.1% (241)	62.4% (181)	47.9% (139)
OSA = no	16.9% (49)	37.69 (109)	52.1% (151)

5. Methods and results

The two-way rule generation involves (1) the knowledge-driven method, based on the hypothetical rules created by medical experts, (2) the data-driven method, based on machine-generated rules, and (3) integration of both methods. The preliminary results show that the rules extracted through machine learning algorithms can confirm, contradict, or expand the rules created by medical experts.

This study applies two hypothetical rules from the knowledge-driven method, *ER1* and *ER2*, which exemplify (1) a high-risk group: older male patients with morbid obesity (BMI > 40), and (2) a low-risk group: young female patients with normal weight (BMI < 25):

ER1 = IF BMI>40 AND age>65 AND gender=male
THEN OSA=yes,
ER2 = IF BMI<25 AND age<25 AND gender=female
THEN OSA=no.

For the machine-generated rules, we use a decision tree classifier C4.5 [7] to induce small, interpretable, yet sufficiently specific decision trees and to generate comprehensible rules from trees. The experimental results were produced by the prototype system, Hypnos, based on the Weka decision tree learner J48 [8]. The classifiers were trained and tested on sets A, B, and C using the stratified 10-fold cross-validation.

5.1. Experimental results

Three classifiers were induced: (1) Model 1 (Tree 1d) based on age, gender, and BMI; (2) Model 2 based on age, gender, BMI, and HTN; and (3) Model 3 based on age, gender, BMI, and MP. In all figures, the nodes represent the predictors, the branches correspond to predictor values, and the leaves correspond to the outcome classes. The two numbers in the leaves represent instances covered by the rule premise and exceptions from the rule.

5.1.1 Model 1. Model 1 (Tree 1d) was constructed in two steps: (1) induction of separate classifiers from data sets A, B, and C; and (2) induction of Tree 1d using redistribution of the instances among sets A, B, C to balance the female to male ratio. Figures 1, 2, and 3 represent decision trees 1a, 1b, and 1c induced separately from data sets A, B, and C. The difference between the predictors at the root level (BMI, Gender) is the result of a relatively small number of female records. This low ratio of female to male patients is typical in clinical practice [9]. In set B, females constitute 21.46% (40/233) of all records and only 9.5 % (13/137) of OSA patients (AHI ≥15). Similarly, in set C, females constitute 27.59 % (80/290) of all patients and only 17.99 % (25/139) of OSA patients. Figure 4 shows the decision tree induced from combined sets A and B. Tree 1d is chosen as Model 1, since it has a better accuracy than tree induced from combined sets A, B and C, and tree induced from combined sets A and C. The rules are generated directly from the decision Tree 1d. Each leaf results in one independent conjunctive rule. For example, the right sub-tree is converted into three rules: (1) “IF BMI > 28.03 AND GENDER= female AND AGE > 33 THEN OSA=yes” (coverage = 18.39%, accuracy = 60.85 %); (2) “IF BMI > 28.03 AND GENDER= female AND AGE ≤ 33 THEN OSA=no” and (3) “IF BMI > 28.03 AND GENDER= male THEN OSA=yes”.

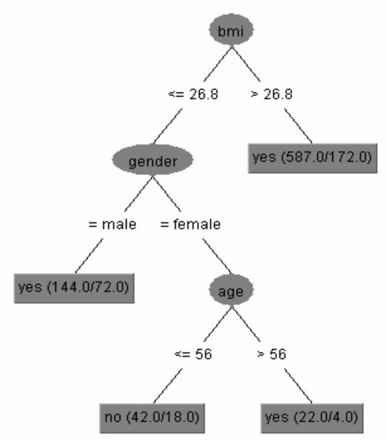

Figure 1. Tree 1a induced from data set A

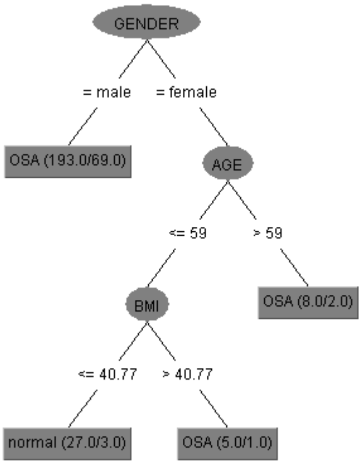

Figure 2. Tree 1b induced from data set B

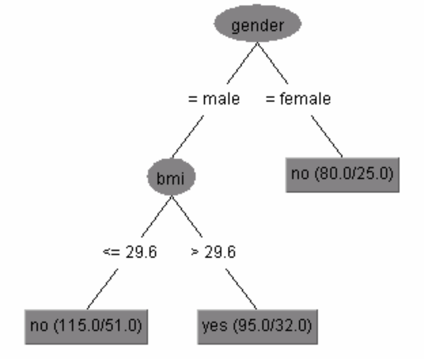

Figure 3. Tree 1c induced from data set C

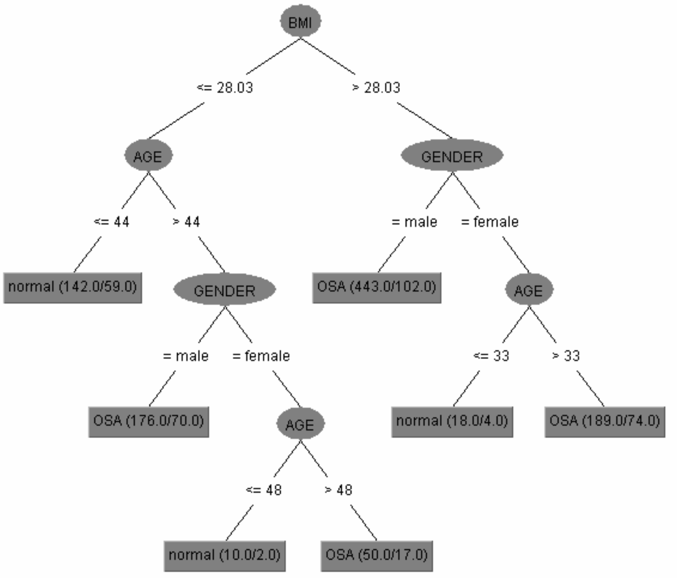

Figure 4. Tree 1d induced from data set A+B

5.1.2 Model 2. Figure 5 shows the model based on gender, age, BMI, and hypertension (htn).

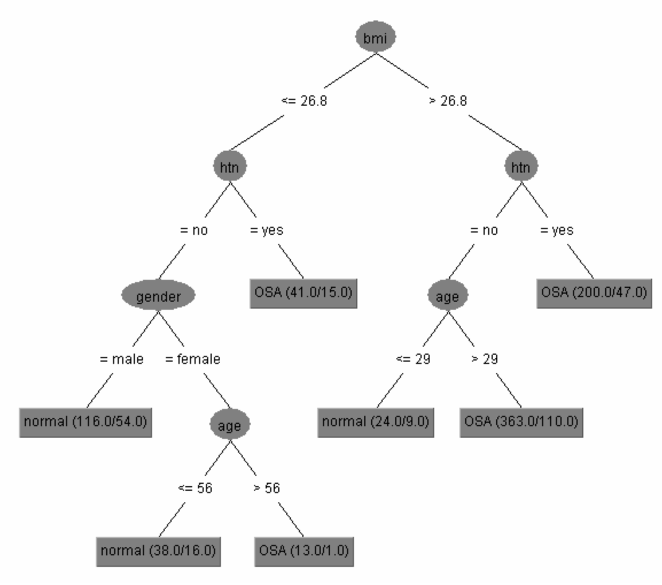

Figure 5. Model 2 induced from data set A

5.1.3 Model 3. Figures 6-7 show the model based on gender, age, BMI, and Mallampati score (MP). The root node splits the data by gender: the left subtree relates to males only; while the right subtree relates to females. Figure 6 illustrates the subtree for male records. The Mallampati score > 2 is indicated as a good predictor of OSA [6]. The leaf for MP=1 (OSA) results from one exceptional clinical record. Figure 7 shows the right subtree for female records. The structure of the tree indicates that MP predictor is weaker in case of females. However, this finding is based only on the particular distribution of 40 female records and should be further studied on larger sets.

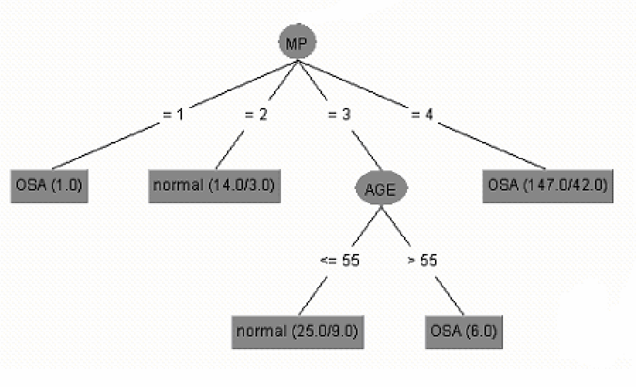

Figure 6. Left subtree induced from data set B

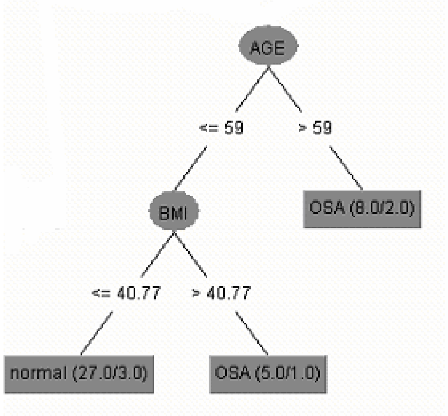

Figure 7. Right subtree induced from data set B

5.1.4 Comparison of data-driven models. The accuracies of the machine-generated classifiers are shown in Table 2.

Table 2: Comparison of classifiers

	Model 1	Model 2	Model 3
Sensitivity	92.28 %	80.69 %	82.50 %
Specificity	10.59 %	26.47 %	50.00 %
Accuracy	63.42 %	62.14 %	69.81 %

5.2. Integration of knowledge-driven and data-driven methods

The integration of expert-generated models and machine-generated models is based on three criteria: (1) the equivalency of the predictors, (2) internal and external validity, and (3) clinical usability: interpretability, simplicity, and practicality. All models use the common three predictors: BMI, age, and gender. Model 2 uses additionally HTN and Model 3 uses additionally MP.

The interpretation of computer-generated rules might (1) confirm the hypothetical rules, (2) provide contradictory examples, or (3) identify new insights. The following examples illustrate the three outcomes:

5.2.1. Confirmation. Expert-generated rule *ER1* specifying high OSA risks for morbidly obese older male patients: "IF BMI > 40 and age > 65 AND gender =male THEN OSA=yes" is confirmed by the following rules: (1) The rule derived from Model 1: IF BMI > 28.03 AND GENDER=male THEN OSA=yes; and (2) Two rules obtained from Model 2: "IF BMI > 26.8 AND HTN = yes THEN OSA=yes", "IF BMI > 26.8 AND HTN= no AND age > 29 THEN OSA = yes."

5.2.2. Contradiction. Expert-generated rule *ER1* specifying high OSA risks for morbidly obese older male patients: "IF BMI > 40 and age > 65 AND gender =male THEN OSA=yes" is contradicted by rule extracted from Model 3: "IF GENDER=male AND MP=2 THEN OSA=no."

Expert-generated rule *ER2* specifying low OSA risks for young female patients with normal weight: "IF BMI < 25 AND age < 25 AND gender = female THEN OSA = no" is contradicted by the rule generated from Model 2, which additionally includes hypertension (HTN): "IF BMI<=26.8 AND HTN=yes THEN OSA=yes."

5.2.3. Knowledge expansion. All models include specific sub-trees concerning female patients (GENDER=female), which classify the females into two groups based on the age predictor. Model 1 divides females into groups based on age ≤ 48 and age > 48 (for BMI ≤ 28.03) and age ≤ 33 and age > 33 (for BMI > 28.03). Model 2 divides into groups based on age ≤ 56 and age > 56 (for BMI ≤ 26.8 and HTN=no). Model 3 divides into groups based on age ≤ 59 and age > 59. The rule extracted from Model 3 classifies all females above 59 as having OSA: "IF GENDER=female AND AGE>59 THEN OSA=yes." This specific age-based division could be associated with an increased risk of OSA among postmenopausal women. However, menopause is also associated with increased central obesity [10]. This issue requires further studies on larger sets of female records.

6. Conclusion and future work

Medical researchers and clinical practitioners study various methods to improve the validity and reliability of clinical prediction rules. In this paper we describe how the machine learning techniques can be used to facilitate and refine the rule derivation process. The integration of the results from the knowledge-driven and data-driven approaches provides confirmation, contradiction, or expansion for the expert-generated prediction rules. Although our study is limited to few predictors, the results demonstrate that our approach is valid, and warrants future work involving additional predictors and further machine learning techniques.

In this study, we identified two problems: (1) diverse definitions of OSA diagnostic criteria based on AHI ≥ 5, ≥ 10, ≥ 15; and (2) a high prevalence of patients with OSA in our datasets. The first problem was addressed by restricting the definition of OSA to specific AHI threshold values. The second problem is present in many medical studies of OSA. Since the diagnostic criteria involve the gold standard of overnight PSG at a cost of at least $1000 dollars per study, the additional studies of healthy control groups are cost prohibitive.

We are planning to expand our work in three directions: (1) development of models based on all known and suspected OSA predictors, (2) application of other machine learning techniques and approaches, for example, fuzzy decision trees, (3) training and testing on larger and more diversified data sets. Furthermore, we are developing a telemedicine application, which will test the utility of the rules in clinical settings.

Acknowledgments: Dr. Ayas is supported by a New Investigator Award from CIHR/BC Lung Association, a Scholar Award from the Michael Smith Foundation, and a Departmental Scholar Award from the UBC. Funding for the database was supported in part by a Michael Smith Foundation Infrastructure Award and an operating grant from the BC Lung Association.

7. References

[1] N. J. Douglas, *Clinicians' Guide to Sleep Medicine,* Arnold, London, 2002.

[2] S. Redline, V. K. Kapur, M. H Sanders, S. F. Quan, D. J. Gottlieb, D. D. Rapoport, et al., "Effects of Varying Approaches for Identifying Respiratory Disturbances on Sleep Apnea Assessment." *Am. J. Respir. Crit. Care Med*., 161, 2000, pp. 369-374.

[3] W. H. Tsai, W. W. Flemons, W. A.. Whitelaw, and J. E. Remmers, "A Comparison of Apnea-Hypopnea Indices Derived from Different Definitions of Hypopnea." *Am. J. Respir. Crit. Care Med*., 159(1), 1999, pp. 43-48.

[4] C. R. F. Nieto, T. Young, B. K. Lind, E. Shahar, J. Samet, S. Redline, et al., "Association of Sleep-Disordered Breathing, Sleep Apnea, and Hypertension in a Large Community-Based Study", *JAMA*, 283(14), 2000, pp. 1829-1836.

[5] T.Young, , J. Skatrud,. and P. E., Peppard, "Risk Factors for Obstructive Sleep Apnea", *JAMA*, 291(16), 2004, pp. 2013-2016.

[6] B. Lam, M.S.M. Ip and C.F. Ryan, "Craniofacial profile in Asian and white subjects with obstructive sleep apnoea", *Thorax,* 60, 2005, pp.504-510.

[7] J.R. Quinlan, *C4.5.:Programs for machine learning*, Morgan Kaufmann, San Francisco, 1993.

[8] I. H. Witten and E. Frank,Data Mining: Practical Machine Learning Tools and Technologies with Java Implementations, Morgan Kaufmann, San Francisco, 2005.

[9] A.S. Jordan and R.D. McEvoy, "Gender differences in sleep apnea: epidemiology, clinical presentation and pathogenic mechanisms", *Sleep Medicine Reviews*, 7(5), 2003, pp. 377-389.

[10] R. Grunstein, "Endocrine Disorders" in *Principle and Practice of Sleep Medicine*, Third Edition, M.H. Kryger, T. Roth, W.C. Dement, Eds. W.B. Saunders Company, Philadelphia, 2000, pp.1103-1112.