DOI: 10.15439/2019F258
ISSN 2300-5963 ACSIS, Vol. 18

Medical data exploration based on the heterogeneous data sources aggregation system

Andrzej Opaliński*, Krzysztof Regulski*, Barbara Mrzygłód*, Mirosław Głowacki*†,
Aleksander Kania‡, Paweł Nastałek‡, Natalia Celejewska-Wójcik‡, Grażyna Bochenek‡ and Krzysztof Sładek‡
*AGH University of Science and Technology,
al.A.Mickiewicza 30, 30-059, Krakow, Poland
†The Jan Kochanowski University, ul.Zeromskiego 5, 25-001, Kielce, Poland
‡II Chair of Internal Medicine, Faculty of Medicine, Jagiellonian University Medical College,
Skawinska 8, 31-066, Krakow, Poland

Abstract—**The paper presents the implementation and use of the IT system implemented in the Department of Pulmonology of The University Hospital in Cracow. The system integrates data from heterogeneous sources of therapy, diagnosis and medical test results of patients with Obstructive Sleep Apnea (OSA). The article presents the main architectural assumptions of the system, as well as an example of data mining analyzes based on the data served by the system. The example of the research aims to present the possibilities offered by the integration of clinical data in telemedicine and the diagnosis of patients with sleep disordered breathing that may lead to certain comorbidities and premature death.**

I. Introduction

OBSTRUCTIVE Sleep Apnea (OSA) is a widespread sleep disorder. It is estimated that the syndrome is present in approximately 5% of the general human population [1]. It is characterized by obstruction of the upper airway despite ongoing breathing efforts that lead to intermittent hypoxia and awakenings. The typical symptoms of OSA are loud snoring with pauses in breathing and daytime sleepiness. If untreated, OSA can lead to a number of severe medical conditions, mainly cardiovascular complications [2].

Polysomnography is the gold standard in the diagnosis of OSA [1]. Continuous positive airway pressure (CPAP) is the gold standard in OSA treatment [14], [9], [20]. Diagnostic process as well as the therapy requires access to information from many sources. Both patient history and clinical examination as well as polysomnography (PSG) results are considered. Obtained data has a very diverse form and is generated from many sources and by various devices (physician, PSG result, therapy devices etc.). So far, the data was collected in various places — paper documentation, patient registration system, PSG service system and SD cards of CPAP devices. In the present paper we demonstrate unique solution among another polysomnography software, assembling various data in the one system.

The implemented system allows for detail research and data analysis leading to improvement of diagnostic quality and shortening its time by gathering all required data in one system [15]. Clinical data in connection with PSG results and CPAP recordings are input to the analysis of multidimensional and multicriteria links between individual indicators and other tests (i.e. blood lipids, arterial blood gases, glucose, creatinine etc.). It is expected that the development of research based on these indicators will enable the creation of a metamodel of mechanisms operating in this area in the future. At the moment, the aim of the research was to identify in which situations the basic diagnostic criteria (individual indicators mentioned above) fail.

For the initial cardiovascular risk assessment, the SCORE cardiovascular risk algorithm (Systematic COonary Risk Evaluation) is frequently used. The main risk factors for cardiovascular complications in the studied OSA cohort collected in the database include: (1) BMI (body mass index), (2) blood cholesterol, (3) systolic blood pressure, (4) package years — a factor calculated as a combination of years of smoking of cigarettes and the number of pieces smoked per day, (5) gender, (6) age.

Selected PSG parameters include: (1) AHI (Apnea Hypopnea Index) — number of apneas and hyponeas per hour of sleep), the most important OSA indicator used to determine its severity, (2) ODI (Oxygen Desaturation Index) — number of hemoglobin oxygen saturation falls per hour of sleep) demonstrate the level of sleep hypoxia.

AHI is the basic and the most common objectively used index to stage OSA severity. Judging the severity of disease we should take into account other tools including subjective doctor's opinion. This personal impression may have important value and can be measured by Clinical Global Impression Severity Scale (CGISS).

The above list shows a certain space of possible inaccuracies. Based on the medical history and physical examination, physician evaluating patient using this subjective scale decides of the urgency for PSG examination.

The work was realized as a part of fundamental research financed by the Ministry of Science and Higher Education, grant no. 16.16.110.663.
The research was co-financed with National Scientific Leading Center project founds.
This research was supported in part by PLGrid Infrastructure

II. Related Works

While the problem of the diagnosis and treatment of diseases related to OSA is a well known subject, the number of IT solutions supporting doctors in this field is relatively small. One of the works in this field mentioned in the literature is the system described by Passali et al. [16], which concerns the database of OSA patients undergoing upper airway surgery. Anthropometric data, results of scales diagnosing OSA occurrence, data from PSG tests, laryngological tests and laboratory tests were stored in the database. The collected data concerned the condition of patients before and after surgery, allowing their use as a source for methods supporting the automatic diagnosis of patients.

A separate extensive system that collects data on patients with OSA is the ESADA database [5], which integrates data from 22 medical units from all over Europe. The system stores data regarding the treatment of patients from the moment of diagnosis through the entire treatment process. Such a diverse range of patient groups allowed for a series of studies related to the detection of previously unknown dependencies, the causes of disease, including environmental and epidemiological conditions [19]. Based on data collected in the ESADA system, the relationship between OSA and problems related to hypertension [21], [22], kidney diseases [12] and diabetes [10] were also determined. These data also allowed to indicate the relationship between the use of different scales of diagnosis of OSA on the effectiveness and accuracy of the diagnosis process itself [2]. Due to the large number of data stored in the ESADA system, we can expect in the near future further publications of research results, developed on their basis.

When it comes to OSA data integration systems in individual countries, two such solutions have emerged in recent years. One of them was the Turkish TURKAPNE system (The Turkish Sleep Apnea Database), which began operation in 2017 and is to collect information about patients treated with OSA within the next 10 years. Another is the Danish NDOSA patients database [7], [8], which is assumed to collect data on the treatment of OSA-related conditions, in order to improve the quality of treatment in this field. Based on the publications data available in the literature, systems supporting treatment in the OSA field begin to appear in the medical market to facilitate and improve the process of diagnosis and treatment of patients. However, these are mostly databases themselves, without advanced diagnostic algorithms based on more extensive methods of data analysis

III. Solution concept and system architecture

In order to integrate all heterogeneous data sources, and ensure their consistency and security, an IT system was developed. It's main elements are presented in Fig. 1 and it consists of:

- A virtual central server maintained in the infrastructure of ACK Cyfronet ;

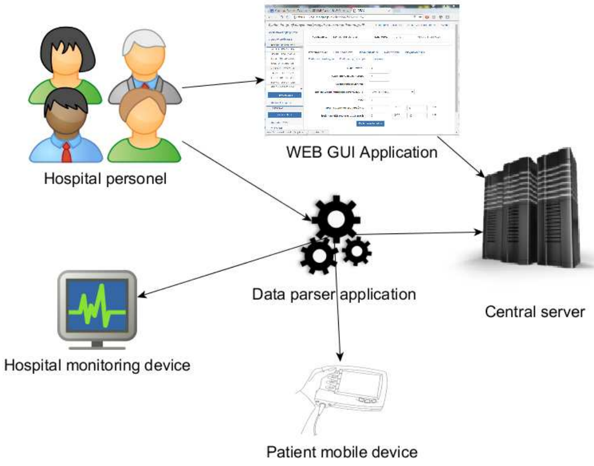


Fig. 1. Main system components

- A graphical user interface that provides access to the system for the hospital's medical staff (for entering data in non electronic format — diagnoses, lab results, etc.);
- Application that allows data to be exported to the system from two types of devices that monitor patients' sleep:
  - PSG device — advanced monitoring of patient's sleep during hospital stay;
  - CPAP devices — mobile patient devices that monitor patient's sleep during therapy outside the hospital.

The main application operating on the central server is based on the MVC (model-view-controller) architectural model developed with PlayFramework programming environment. The graphical user interface (GUI) of the application, which allows hospital staff to enter data and manage the system is presented in Fig. 2.

Datasets stored in the system are based on following:

- Clinical data — interview, physical examination, diagnosis, drugs;
- Diagnostic tests — PSG, CPAP, laboratory tests, spirometry;
- Medical recommendation — previous and planned treatment.

At the model layer, data is stored in a relational MySQL database, the structure of which is presented in Fig. 3. The business logic of the application is implemented in JAVA language on a virtualized CentOS operating system. The presentation layer is based on script templates in SCALA with HTML output code.

Based on the graphical user interface, hospital personnel provide the data in the system during the patient's visit to the hospital. In addition to standard questionnaires and medical diagnosis, data on further treatment as well as laboratory tests and lung tests — stored so far in external IT systems — are also provided.

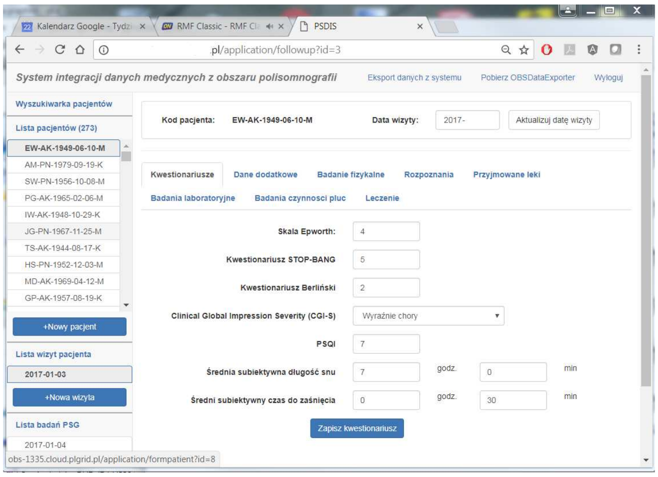
Another key element of the system is the application that allows to extract data from devices that monitor the patient's

http://www.cyfronet.krakow.pl/en/4421,main.html sleep. The current version of the system supports two types of devices. The first of these is the PSG device, which monitors the patient's sleep during his several-day visit to the hospital. The Sleepware3D application from Phillips Respironics is used to operate this device, and the data obtained from it is saved in RTF format. The second type of devices that monitors the patient's sleep during his stay outside the hospital are mobile versions of devices — CPAP, which generate reports in PDF format. In order to import selected reports from CPAP and PSG devices into the system, hospital staff run manually a dedicated application. The application developed within the system allows (using dedicated parsers algorithms) to extract data from device reports and export them to a central server that integrates them with other data related to a specific patient. PDFParser and RTFEditoKit JAVA libraries were used to extract data from documents in PDF and RTF formats as well as dedicated templates for extracting relevant data from these documents. Due to the fact that data import is performed for an individual patient, there is no problem of overloading the system during the import process.

Such integrated data, acquired from many heterogeneous sources, previously stored within various information systems and in paper form, are integrated into one universal data model (Fig. 3) within the presented system and made available for processing for advanced data processing and analysis methods.

IV. Data exploration results

In the Introduction section, a research problem was initially drawn up. Data analysis presented in this research was performed to show the huge potential for integrated data collection in the IT system for their use in clinical practice. All computations was carried out on an integrated patient database, with the use of STATISTICA software. At first, the relationships between some patients' clinical data and their characteristics were investigated. The analysis showed that some dependencies exist. It should be stated here that particular variables could have different scales. Some of them were qualitative (e.g. sex), other ordinal (e.g. measuring scales) and other quantitative (e.g. index of disorders). A series of tests on the dependence of parameters was performed — correlation test (Pearsons's coefficient), t tests as well as chi square tests. Some of the dependencies can be presented in the charts Fig. 4.



EW-AK-1949-06-10-M
AM-PN-1979-09-19-K
SW-PN-1956-10-08-M
PG-AK-1965-02-06-M
IW-AK-1948-10-29-K
JG-PN-1967-11-25-M
TS-AK-1944-08-17-K
HS-PN-1952-12-03-M
MD-AK-1969-04-12-M
GP-AK-1957-08-19-K

2017-01-03

2017-01-04

Kod pacjenta: EW-AK-1949-06-10-M
Data wizyty: 2017-

Skala Epworth: 4
Kwestionariusz STOP-BANG 5
Kwestionariusz Berliński 2
Clinical Global Impression Severity (CGI-S) Wyraźnie chory ▾
PSQI 7
Średnia subiektywna długość snu 7 godz. 0 min
Średni subiektywny czas do zaśnięcia 0 godz. 30 min

Fig. 2. WEB interface for system management

From the charts of means in groups (Fig. 4) we can find the relationship between gender and the number of package-years and AHI. The AHI variable includes three classes that express the severity of the disease, where AHI = 3 means a severe OSA. We can observe a pattern that men with severe OSA smoke more. The question is that AHI depends on the amount of cigarettes smoked. The third graph shows that smoking correlates with AHI. While it cannot be demonstrated that smoking causes OSA, it can be assumed that perhaps smoking aggravates the severity of the disease. Such conclusions could be drawn from the tested sample it is certainly a pattern that is worth further research. Does severe OSA predispose to heavy nicotine dependence?

We can study the effect of risk factors on various OSA indices to find differences in their diagnostic strength. (Fig. 5). It can be seen that while age and BMI clearly affect the SCORE value, their impact on AHI and ODI (direct OSA indices) is small. This may negatively affect SCORE diagnostic capabilities, as evidenced by subsequent analyzes.

A. *False Negatives Recognition*

Looking for naturally occurring data structures, groups of patients with similar indicators and clinical characteristics, the method of clustering — unsupervised learning — k-means was used. Clustering is the most extensive group of machine learning methods called "unsupervised". Among the many known algorithms (EM algorithm, fuzzy c-means, Kohonens' Neural Networks, etc.) one of the oldest and most popular tools for the development of other tools is the k-means algorithm [11]. Clustering deals with searching for a structure in a set of unidentified data. It is the process of organizing objects into groups whose elements are in some way similar to each other [4]. In terms of computations, the algorithm is reduced to two-criteria optimization, where the distance between cluster objects is minimized, and the distance between clusters is maximized [6], [18].

The results are presented in Table I. The tests were successful, we managed to determine the concentration of patients with similar characteristics with a small error of validation. The distance between clusters was calculated by the Euclidian metrics, while the means and the most frequent values for descriptive variables are presented in Table I.

The analysis shows that 5 clusters can be distinguished from the patients (the number of clusters was set with cross-validation method): cluster 1, 3 and 4 these are cases of elevated AHI — which means that they group patients suffering from severe OSA. In contrast to clusters 2 and 5, which focus patients with mild form of the disease. It can be demonstrated using Table I that cluster 1 and 4 are: women with severe OSA and men with severe OSA — their CGIS (subjective

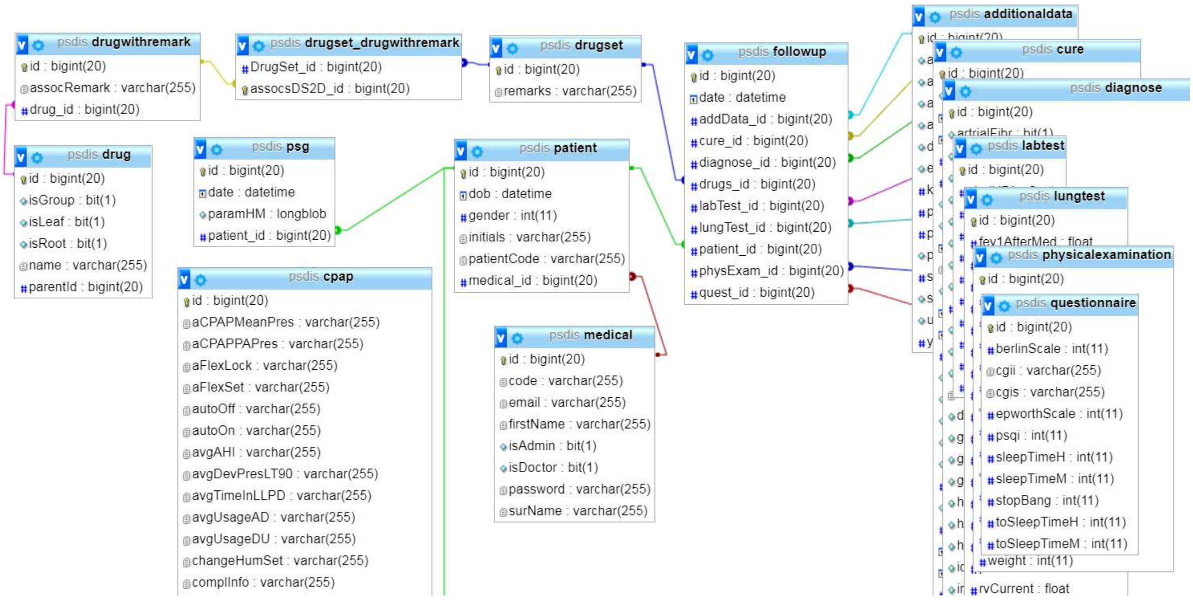


Fig. 3. Database structure

assessment of the doctor) and SCORE are high. Similarly, cluster 2 and 5 are the men and women with the lowest AHI — there CGIS and SCORE are low.

However, the analysis has identified yet another cluster — number 3. It is characteristic due to the high class of AHI (this class means that in the PSG examination the patient had the Apnea-Hypopnea Index >30), and at the same time low SCORE and CGIS. This means that they are possibly false negatives cases — patients with risk of having undiagnosed OSA (if the decision on referral for PSG examination should be made on the basis of machine learning methods).

Statisticians call this situation False Negatives. In statistical hypothesis testing false negatives are type II errors, where a negative result corresponds to not rejecting the null hypothesis. We can call it an underdiagnoses error. The conducted research has shown that it is possible to select risk groups — patients whose diagnosis may be subject to the error false negatives, especially if the diagnosis would be carried out without PSG examination.

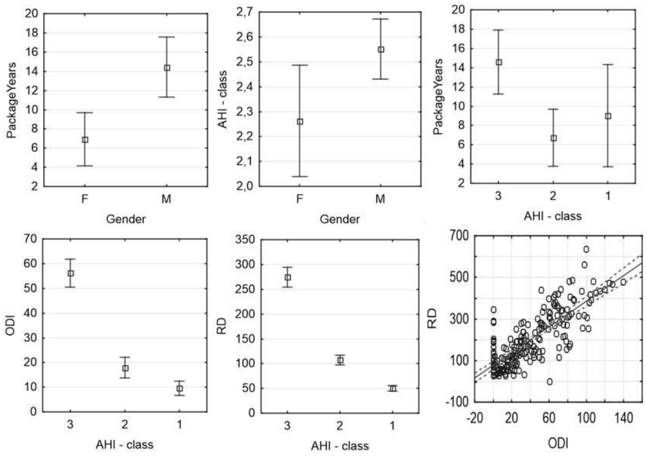


Fig. 4. Selected relationships between patients’ characteristics

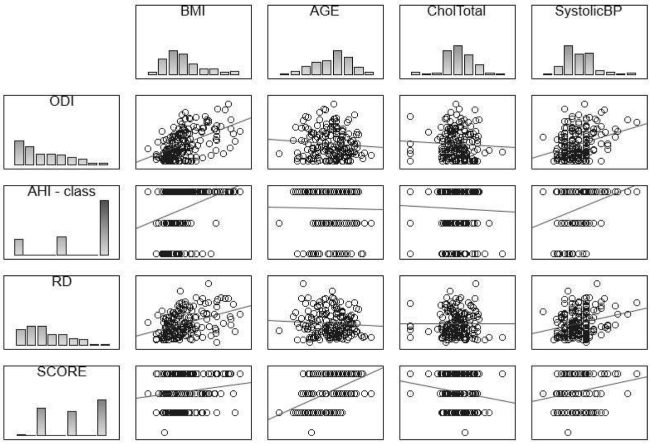


Fig. 5. The impact of risk factors on various OSA indicators

Is it possible to use this knowledge in future diagnoses? How machines can predict that a patient may belong to the FalseNegatives group and protect him against a mistaken underdiagnoses?

Cluster 3 is men who smoke less than other patients, who is also the youngest group among the respondents. They have high BMI, elevated blood pressure and cholesterol at the same time.

B. False Negatives Prediction

In order to create a classification model, the algorithm for creating CART classification trees was used. Decision trees

TABLE I
K-MEANS CLUSTER CHARACTERISTICS

Cluster	Gender	Age	CGIS	Package Years	BMI	Systolic BP	SCORE	Chol Total	AHI dominant	No of cases	(%)
1	F	58,4	3,6	8,5	40,0	140,5	3,15	4,29	3	32	17,0
2	F	59,3	3,0	4,2	30,1	132,2	2,80	5,27	1	25	13,2
3	M	43,1	3,0	9,2	34,3	141,9	2,20	4,90	3	29	15,4
4	M	62,8	3,7	19,5	34,8	142,4	3,78	4,41	3	64	34,0
5	M	56,7	2,8	10,3	28,4	133,2	2,71	4,46	1	38	20,2

are very popular among data mining tools [3]. Repeatedly described in the literature, they have found countless examples of applications [17], [21]. Their usefulness has been determined by the following characteristics: (1) easy to interpret by a human; (2) simple representation of complex relationships occurring in data sets; (3) no assumptions on the variability of input and output parameters, and (4) no assumptions on the probability distribution of variables; (5) the ability to operate on incomplete and noisy sets [13].

The algorithm evaluates the discriminative power of variables and chooses for division successively those that provide better separation of objects between classes. Then the split point is selected. The lowest Gini Index is chosen as the best dividing point. This process is repeated until a satisfactory tree is obtained (based on the leaf size or total classification error).

Fig. 6 presents the matrix of errors for the inducted CART classification tree. Using the tree we can predict that the patient belongs to the group of patients at risk of underdiagnoses mistake in 82.7% of cases. We reduce the risk of misdiagnosis by as much as a percentage if we consider the indications from the analysis.

Because false negatives cases mainly concern men, in graph in Fig. 7 only a fragment of the tree about men is presented. Only in this area it is possible to predict the cluster 3.

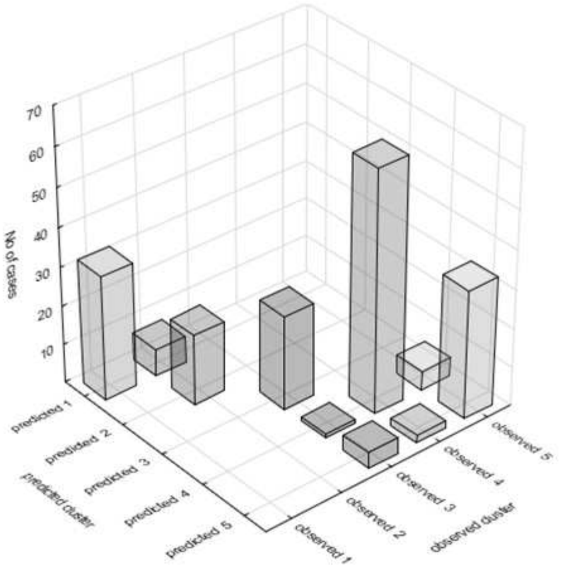


Fig. 6. The misclassification matrix for the False Negatives classification tree

Reading the rules induced by the tree we can notice: If someone is a man under 54 years of age, with a SCORE index below 3, it belongs to cluster 3 (FN) with 65% of certainty. If we additionally take into account the results of spirometry, including Tiffeneau index after bronchodilator below 86, the probability increases to 73.6%. Taking into account laboratory tests, it is worth paying attention to the content of C-reactive protein. If it is above 2 we are sure that it is FN (cluster 3), otherwise we check if the patient is younger than 43 years, then it certainly belongs to cluster 3. The remaining patients belong to cluster 5 — means true positives — patients with mild OSA.

When interpreting the results, it should be taken into account that the obtained results are susceptible to an error resulting from a small number of data stored so far in the system, and thus a relatively small number of variables that can be used in the analyzes. In the future, successive algorithms (decision trees and clustering) should operate on disjoint sets of attributes to avoid problem of endogeneity.

V. SUMMARY

This paper presents the design, implementation and preliminary results of a prototype system that enables to improve the diagnosis and possible decision-making about treatment of patients with sleep disorders. The main advantages of the system are aggregation of data from various sources (diagnostics, lab-tests, medical history) and its integration with devices for OSA diagnosis (PSG) and treatment (CPAP). It can significantly improve the work of the hospital staff and facilitate their access to previously distributed patient data.

Besides, the examples of data analyzes that allow searching for dependencies between clinical tests and diagnosis have been shown. Data mining analyzes allowed to find the characteristics of a group of patients for whom the risk of underdiagnoses was the highest. However, from a medical point of view, all presented results and obtained dependencies should be treated with extreme caution, because a small number of randomly selected parameters does not necessarily reflect reality. At this stage, the paper has only statistical and IT value, and its purpose was to show the huge potential for integrated data collection in the IT system for their use in clinical practice. The intention of the presented system is to improve the quality of diagnosis and treatment of patients affected by OSA and it seems that this goal is achieved after implementation of the system into daily clinical practice.

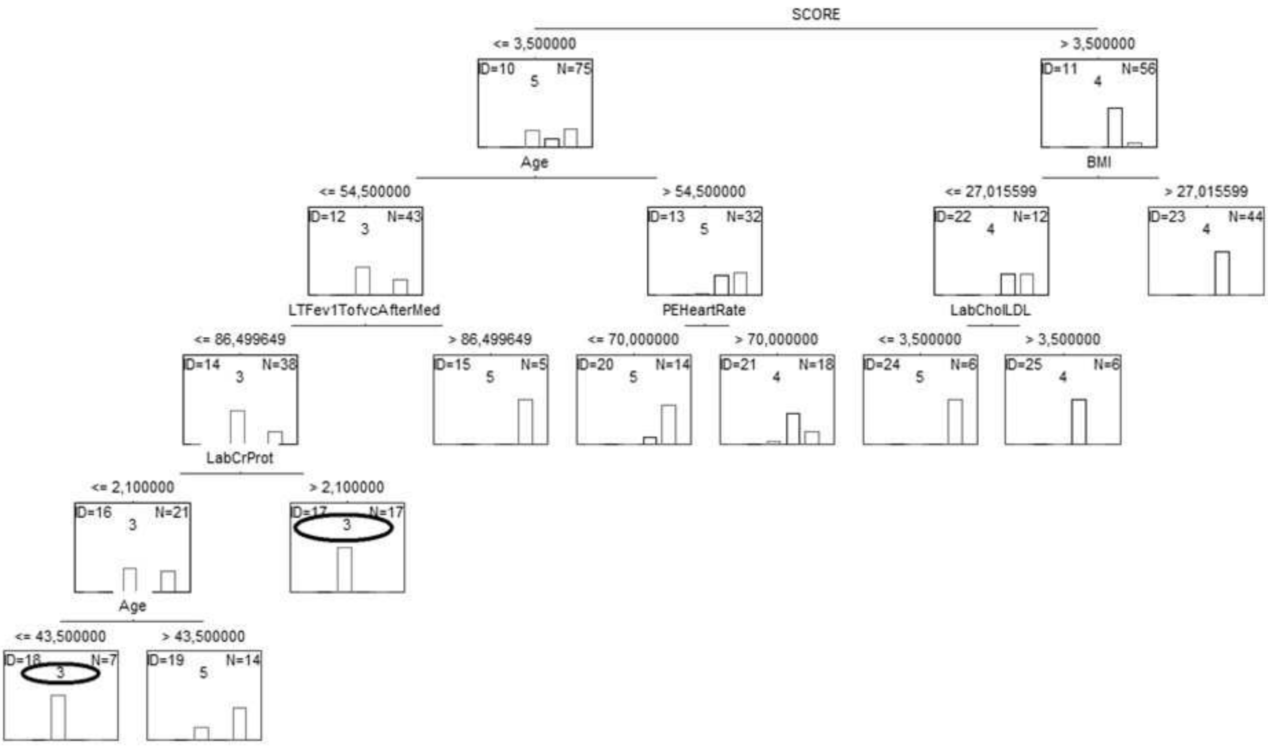


Fig. 7. Fragment of the classification tree

References

[1] H.Y. Chiu, P.Y. Chen, L.P. Chuang, N.H. Chen, Y.K. Tu, Y.J. Hsieh, Y.C. Wang, and C. Guilleminault. Diagnostics accuracy of the berlin questionnaire, stop-bang, stop, and epworth sleepiness scale in detecting obstructive sleep apnea: A bivariate meta-analysis. *Sleep Medicine Reviews*, 36:57–70, 2016.

[2] P. Escourrou, L. Grote, T. Penzel, W. T. Mcnicholas, J. Verbraecken, R. Tkacova, and F. Barbé. The diagnostic method has a strong influence on classification of obstructive sleep apnea. *Journal of sleep research*, 24(6):730–738, 2015.

[3] A. Glowacz. Acoustic based fault diagnosis of three-phase induction motor. *Applied Acoustics*, 137:82–89, 2018.

[4] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

[5] J. Hedner, L. Grote, M. Bonsignore, W. McNicholas, P. Lavie, G. Parati, P. Sliwinski, F. Barbé, W. De Backer, P. Escourrou, I. Fietze, J. A. Kvamme, C. Lombardi, O. Marrone, J. F. Masa, J. M. Montserrat, T. Penzel, M. Pretl, R. Riha, D. Rodenstein, T. Saaresranta, R. Schulz, R. Tkacova, G. Varoneckas, A. Vitols, H. Vrints, and J. Zielinski. The european sleep apnoea database (esada): report from 22 european sleep laboratories. *Eur Respir J*, 38:635–42, 2011.

[6] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[7] P. Jennum, R. Ibsen, and J. Kjellberg. Morbidity prior to a diagnosis of sleep-disordered breathing: a controlled national study. *J Clin Sleep Med*, 9(2):103–108, 2013.

[8] P. J. Jennum, P. Larsen, C. Cerqueira, T. Schmidt, and P. Tønnesen. The danish national database for obstructive sleep apnea. *Clinical Epidemiology*, 8:573–576, 2016.

[9] Riha R.L. Jennum, P. Epidemiology of sleep apnoea/hypopnoea syndrome and sleep-disordered breathing. *Eur Respir J*, 33:907–14, 2009.

[10] B. D. Kent, L. Grote, M. R. Bonsignore, T. Saaresranta, J. Verbraecken, and P. Lévy. Sleep apnoea severity independently predicts glycaemic health in nondiabetic subjects: the esada study. *European Respiratory Journal*, 44(1):130–139, 2014.

[11] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Symposium on Math*, pages 281–297, Statistics, and Probability, Berkeley, CA, 1967. University of California Press.

[12] O. Marrone, S. Battaglia, P. Steiropoulos, O.K. Basoglu, J.A. Kvamme, S. Ryan, J.L. Pepin, J. Verbraecken, L. Grote, J. Hedner, and M.R. Bonsignore. Chronic kidney disease in european patients with obstructive sleep apnea: the esada cohort study. *J Sleep Res*, 25:739–45, 2016.

[13] E. Nawarecki, S. Kluska-Nawarecka, and K. Regulski. *Multi-aspect character of the man-computer relationship in a diagnostic-advisory system*, pages 85–102. Springer-Verlag, 2012.

[14] American Academy of Sleep Medicine. *International classification of sleep disorders: diagnostic and coding manual*. Amer Academy of Sleep Medicine, Westchester, Illinois, USA, 2005.

[15] A. Opaliński, P. Nastałek, B. Mrzygłód, N. Celejewska-Wójcik, M. Głowacki, G. Bochenek, K. Regulski, K. Sładek, and A. Kania. The system for integration of heterogeneous data sources in the domain of obstructive sleep apnea. In Economic Advance In Behavioral and Sociocultural Computing (B. E. S. C.) eds Economic, editors, *Proc.Conf. 4th International Conference on Behavioral*, pages 1–6. Demazeau Y, 2017.

[16] D. Passali, G. Caruso, L.C. Arigliano, F.M. Passali, and L. Bellussi. Arigliano lc, passali fm, bellussi i. database application for patients with obstructive sleep apnoea syndrome. *Acta Otorhinolaryngol Ital*, 32:252–255, 2012.

[17] J. R. Quinlan. *Induction on Decision Trees, Machine Learning*. Kluwer Academic Publishers, Boston, 1986.

[18] K. Regulski, D. Wilk-KoÅĆodziejczyk, and G. Gumienny. Comparative analysis of the properties of the nodular cast iron with carbides and the austempered ductile iron with use of the machine learning and the support vector machine. *The International Journal of Advanced Manufacturing Technology*, 87(1):1077–1093, 2016.

[19] T. Saaresranta, J. Hedner, M. R. Bonsignore, R. L. Riha, W. T. McNicholas, T. Penzel, U. Anttalainen, J. A. Kvamme, M. Pretl, P. Sliwinski, and J. Verbraecken. Clinical phenotypes and comorbidity in european sleep apnoea patients. *PLoS One*, 11(10), 2016.

[20] White D.P. Amin R. et al. Somers, V.K. Sleep apnea and cardiovascular disease: an american heart association/american college of cardiology foundation scientific statement from the american heart association

council for high blood pressure research professional education committee, council on clinical cardiology, stroke council, and council on cardiovascular nursing. in collaboration with the national heart, lung, and blood institute national center on sleep disorders research (national institutes of health). *Circulation*, 118:1080–111, 2008.

[21] Y. Song and Y. Lu. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry*, 27(2):130–135, 2015.

[22] R. Tkacova, W. T. McNicholas, M. Javorsky, I. Fietze, P. Sliwinski, G. Parati, L. Grote, and J. Hedner. Nocturnal intermittent hypoxia predicts prevalent hypertension in the european sleep apnoea database cohort study. *Eur Respir J*, 44:931–41, 2014.