Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005

Reliable Determination of Sleep Versus Wake from Heart Rate Variability Using Neural Networks

A.T. Lewicke[1], E.S. Sazonov[1], M.J. Corwin[2], S.A.C. Schuckers[1], CHIME study group
[1] Department of Electrical Engineering, Clarkson University, NY, USA
[2] Pediatrics, Boston Medical Center, Boston University School of Medicine, MA, USA

Abstract— **Heart rate, heart rate variability (HRV), and sleep state are some of the common physiologic parameters used in studies of infants. HRV is easily derived from infant electrocardiograms (ECG), but sleep state scoring is a time consuming task using many physiological signals. We propose a technique to reliably determine sleep and wake using only the ECG. The method will be tested with simultaneous ECG and polysomnograph (PSG) determined sleep scores from the Collaborative Home Infant Monitoring Evaluation (CHIME) study. The advantages include high accuracy, simplicity of use, and low intrusiveness, with design including rejection to increase reliability valuable for determining sleep-wake states in highly sensitive groups such as infants. Learning vector quantization and multi-layer perceptron (MLP) neural networks are tested as the predictors.**

The manual PSG scored test set has 38,121 (67.8%) sleep and 18,076 (32.2%) wake epochs for a total of 56,197 epochs. The MLP classification of the entire test set resulted in 77.7% agreement with the PSG sleep epochs and 79.0% with wake. The rejection scheme applied to the MLP resulted in 28.9% of sleep and wake epochs meeting the rejection criterion. Of the remaining 39,946 epochs 86.0% are in agreement with the PSG sleep epochs and 85.4% with wake.

After systematic rejection of difficult to classify segments, this model can achieve 85%-86% correct classification while rejecting only 30% of the data. This is an improvement of about 7.8% over a traditional model without rejection.

Keywords—**reliability, sleep classification, heart rate variability, neural network, infants**

I. INTRODUCTION

Identification of sleep/wake states is used in several areas of medical science. For infants, sleep state identification in combination with other parameters may be useful in the study of Sudden Infant Death Syndrome (SIDS) [1]. Polysomnography (PSG), which includes an electroencephalogram (EEG), electrooculogram (EOG) electromyogram (EMG), and electrocardiogram (ECG), is the most accurate procedure for determining sleep states and is considered to be the “gold standard”. The largest shortcoming of PSG is that it is rather expensive and too complex to be used by an untrained person. Sleep scoring from a PSG is a time intensive effort since trained technicians need to analyse each epoch manually. Relatively high intrusiveness of the PSG method is also the cause of its low tolerance by nursing-home patients and infants. An appealing alternative is presented by electrocardiogram methods.

Heart rate and heart rate variability (HRV) is one of the commonly used cardiorespiratory signals in studies of infants. HRV is determined by a complex interaction of the sympathetic and parasympathetic divisions of the autonomic nervous system (ANS). It is considered a noninvasive and inexpensive way to analyze the ANS. Many types of parameters have been used to measure heart rate variability, including time domain, frequency domain, time-frequency domain, and nonlinear methods. Heart rate and heart rate variability are sleep state dependent variables. It is assumed that wake state has the highest heart rate and heart rate variability, and quiet sleep (QS) has the lowest value of heart rate and variability. The value for rapid eye movement (REM) sleep is in between QS and wake. Landes et al. found that heart rate is the slowest and contains the highest frequencies in quiet sleep [2]. Sleep state differences are also shown in the frequency analysis of HRV. Medigue et al. found that the high frequency component of HRV (HF) dominates during QS, while low frequency component of HRV (LF) dominates in REM [3]. Kantelhardt et al. used ECG to test the correlation between sleep states [4].

This study aims at validating the use of HRV in sleep/wake identification. Soft computing methods (neural networks) perform classification using a set of HRV features. Special attention is given to the reliability of sleep scoring, since sleep states are usually used as predictors, and may impact the reliability of subsequent estimates. Reliability is improved by incorporating rejection into the model. By rejecting epochs systematically, performance can be improved by designing a neural network model that rejects epochs in the overlap region between two classes. For example, the reliability of sleep states used in studying life-threatening events is crucial for accurate results. The ECG was recorded as a part of the Collaborative Home Infant Monitoring Evaluation (CHIME) NIH study, which studied home infant monitors for apnea and bradycardia for over 1000 infants [5,6,7]. Sleep state information would be helpful in analysing the over 700,000 hours of data recorded on the home monitor where traditional PSG information is not available [7]. Because of the amount of data available it is not required that a decision be made for

0-7803-9048-2/05/$20.00 ©2005 IEEE

every epoch. Therefore, the reliability method can be used to reject epochs by which a confident sleep classification could not be made.

II. Methodology

A. Data

Data used in this study were collected from infants as part of the CHIME study [5]. An 8-hour polysomnograph (PSG) from the CHIME dataset was performed which includes simultaneous manually scored sleep/wake (30 sec epoch) and raw RR interval sampled at 1000 Hertz. A total of 191 infants (ages 34 to 62 weeks PCA) are divided evenly to create training, validation, and testing sets, each with ~57,000 epochs. Two neural networks are compared; multi-layer perceptron (MLP), and learning vector quantization (LVQ), where reliability is incorporated into the model through rejection of epochs where sleep/wake cannot be determined with confidence [8]. The neural networks are designed using the training and validation sets. Results were confirmed by an untouched testing dataset.

Trained technicians performed PSG-based sleep state identification for the infants at 30-second intervals (epochs) [7]. The PSG-identified sleep states (Awake, Active sleep, Quiet sleep and Indeterminate) were used as a baseline for training of the neural network. For this study the Active and Quiet sleep states were combined into "sleep" (ASP) state, the Awake state remained "wake" (AWK), and the indeterminate states were discarded.

This data is preprocessed to create reliable data as follows:

RR interval signal:

1. An ECG artifact rejection routine was run to find artifactual heartbeats and correct the data [9].
2. Both the beginning and end of the RR interval recording were synchronized in time with the PSG data on an epoch (30 seconds) boundary.
3. Standard HRV measures were extracted from the signal for each epoch.
4. A fuzzy C-means (FCM) clustering algorithm [10] was run to find which of the features best separated the data into clusters of sleep and wake. Many of the HRV measures were not expected to be good predictors because the epoch length was too short for the measures to develop. The best HRV measure from the FCM experiment was mean.
5. The mean RR interval for the last 8 consecutive 30-second epochs, formed M_{-8} , … , M_{-2} , M_{-1} , M_0 as separate predictors. Thus, the input to the neural networks was a 9-element vector of epoch mean values.

The PSG records of AWK or ASP represent the response variable during each 30-second epoch. Utilization of lagged metrics as predictors was based on a simple reasoning that there could be no sudden change in the wake/sleep state. Similar lagged models were used in related studies on adults [11].

B. Prediction-Multilayer Perceptron

A Multilayer Perceptron (MLP) neural network was one of the two neural predictors used in this experiment. A MLP uses hyperplanes to separate the data into different classes, and consist of three parts: input layer, hidden layers, and output layers [21]. The inputs to the system are the lagged metrics M_{-n} , … , M_{-2} , M_{-1} , M_0 described in the preprocessing. The network weights are then adjusted to minimize a cost function by a training algorithm. In this case we used the Optimized Levenberg-Marquardt with Adaptive Momentum (OLMAM) training algorithm. This is the traditional LM algorithm with an additional adaptive momentum term that offers excellent convergence [12].

Each structure was tested 100 times with randomly selected initial weights. All results are given by the mean percent correct classification from the 100 experiments. The structure of the MLP was determined by a performance curve where the percent correct classification was plotted vs. the number of hidden neurons (Figure 1). The point where the validation set starts to decrease in performance while the training set continues to increase is the point that gives the maximum training and validation performance.

The next step was to incorporate reliability into the model. An algorithm developed by De Stefano et al. [8] was used to reject epochs that the MLP could not reliably classify. This is achieved by having two outputs in the structure one for sleep and one for wake. The outputs take on values between 1 and 0. Ideally, the network would only have one neuron's output be 1 while the other output neuron would show zero. This is to say that one neuron tells us the output is ASP and the other neuron tells us the output is not AWK. In practice this is not the case. Sometimes both outputs are close to 1 or both outputs are close to zero. This is when the network cannot reliably make a decision. The model uses a measure which takes the difference between the winning output neuron (the maximum of the two) and the losing neuron. We now have a single value that describes the MLP's reliability for a given input epoch [8]. A threshold can then be applied to the reliability measure to determine which epochs are rejected.

A decision must be made as to the maximum acceptable amount of data to be rejected this is an application-specific decision. In the case of the CHIME data, 30% has been selected in consultation with physicians as an acceptable limit. Data that is rejected is not counted as an error because no decision was made towards classifying that data. Do to the large amount of data available and the target use of the data, rejection of portions of the data is acceptable for this application.

C. Prediction-Learning Vector Quantization

A Learning Vector Quantization (LVQ) neural network [13], a subclass of so-called Kohonen networks, was the other neural network tested. The LVQ networks are primarily used as non-linear classifiers, which make this method a perfect candidate for the task at hand. The neural predictor was initially trained utilizing the LVQ-1 algorithm and fine-tuned with LVQ-3.

Classification by both LVQ-1 and LVQ-3 is based on a codebook of vectors m_i, where each codebook vector belongs to a certain class. The input vector X is compared to each codebook vector m_i and assigned to the same class to which the closest codebook vector belongs. The distance d between X and m_i is calculated according to the following formula:

$$d = \min_i \left\| X - m_i \right\| \qquad (1)$$

More information on the LVQ neural networks can be found in [14].

Similarly to MLP, each structure was tested 100 times with random initial codebook vectors, and all results are given by the mean percent correct classification of the 100 structures. Again, the limit for the amount of data rejected in this application is 30%. The structure of the LVQ was determined by a performance curve where the percent correct classification was plotted vs. the number of codebook vectors (Figure 3).

The reliability applied to the LVQ network works by a distance measure. Epochs that are approximately equidistant to codebook vectors of different classes are unreliable. The reliability measure comes from taking the difference between the closest codebook vector to an epoch and the next closest codebook vector of the opposing class. Again, like the MLP, we now have a single metric that can have a threshold applied to it to remove epochs that are not reliably determined [8].

III. Results

A. Multilayer Perceptron

The structure determined from the training and testing sets was nine inputs, 32 hidden layers, and two outputs. The performance plot for the hidden layers can be seen in Figure 1.

A threshold was applied to the output reliability measure. A plot of threshold values against percent correct classification can be seen in Figure 3. As can be seen in the plot, the point where maximum percent correct classification occurs only leaves about 25% of the original data. A rejection of threshold was set to remove 30.9% of the training data. This was the maximum amount of data we were willing to reject. Since the distribution of data is very similar between sets, the validation set rejected 28.2% of the data and the testing set rejected 28.9% of the data. The performance of the MLP comparing no rejection to approximately 30% rejection is summarized in Table 1.

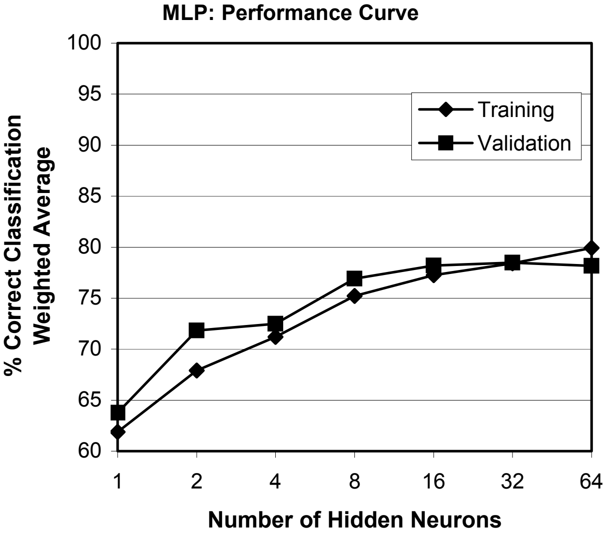

Figure 1. The performance curve for the multilayer perceptron neural network. The number of hidden neurons is plotted against the percent correct classification of the weighted average of asleep and awake epochs.

The manual PSG scored test set has 38,121 (67.8%) sleep and 18,076 (32.2%) wake epochs for a total of 56,197 epochs. The MLP classification of the entire test set resulted in 77.5% agreement with the PSG sleep epochs and 79.0% with wake. The rejection scheme applied to the MLP resulted in 28.9% of sleep and wake epochs meeting the rejection criterion. Of the remaining 39,946 epochs 86.0% are in agreement with the PSG sleep epochs and 85.4% with wake. Thus, rejection of 30% of the data offered a 7.8% increase in correct classification.

B. Learning Vector Quantization

The LVQ network of 128 codebook vectors (64 ASP and 64 AWK) was trained by the data from 64 infants combined into a single data set. Nine input features were used (the mean of the epoch and the mean of the previous 8 epochs). The network was trained for 70,960 iterations and fine-tuned by LVQ_3 for 70,960 iterations. The performance curve for graphically showing over training occurs. This is seen by a decrease in the test set accuracy while the training set continually increases in accuracy.

TABLE 1. MLP NEURAL NETWORK PERFORMANCE: ZERO REJECTION COMPARED TO 30% REJECTION

		Training		Validation		Testing	
		Zero Rejection	30% Rejection	Zero Rejection	30% Rejection	Zero Rejection	30% Rejection
Sleep	Epochs	39827	27228 (68.4%)	37840	26998 (71.3%)	38121	26418 (69.3%)
	% Correct	77.2	85.4	78.5	86.4	77.5	86.0
Wake	Epochs	17742	12559 (70.8%)	18958	13807 (72.8%)	18076	13528 (74.8%)
	% Correct	76.1	83.6	78.1	85.1	79.0	85.4%
Total	Epochs	57569	39787 (69.1%)	56798	40805 (71.8%)	56197	39946 (71.1%)
	% Correct	76.9	84.8	78.4	86.0	78.0	85.8%

The value of 128 codebook vectors was decided from Figure 2.

A threshold was applied to the output reliability measure. The results are quite similar to the MLP in the sense that as more data was rejected, the better overall classification became. A plot of threshold values against percent correct classification can be seen in Figure 3. A rejection of threshold was set to remove 30.4% of the training data. Since the distribution of data is very similar between sets, the validation set rejected 31.1% of the data and the testing set rejected 29.3% of the data. The performance of the LVQ comparing no rejection to 30% rejection is summarized in Table 2.

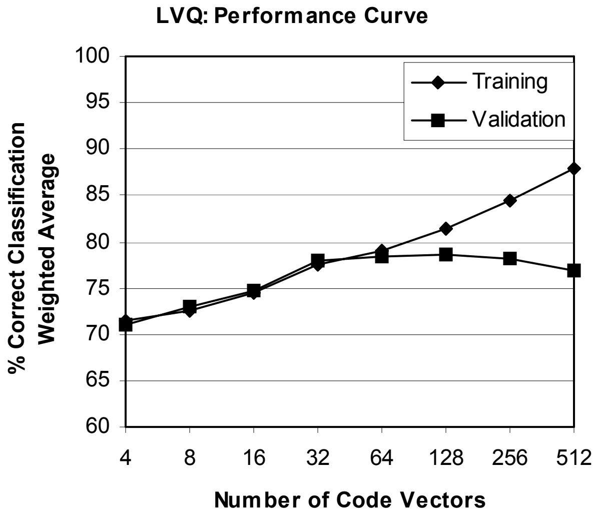

Figure 2. The performance curve for the learning vector quantization neural network. The number of the hidden neurons is plotted against the percent correct classification of the weighted average of asleep and awake epochs.

The manual PSG scored test set has 38,121 (67.8%) sleep and 18,076 (32.2%) wake epochs for a total of 56,197 epochs. The LVQ classification of the entire test set resulted in 79.8% agreement with the PSG sleep epochs and 74.5% with wake. The rejection scheme applied to the LVQ resulted in 29.3% of sleep and wake epochs meeting the rejection criterion. Of the remaining 39729 epochs 87.2% are in agreement with the PSG sleep epochs and 80.9% with wake. Thus, rejection of 30% of the data offered a 7.3% increase in correct classification.

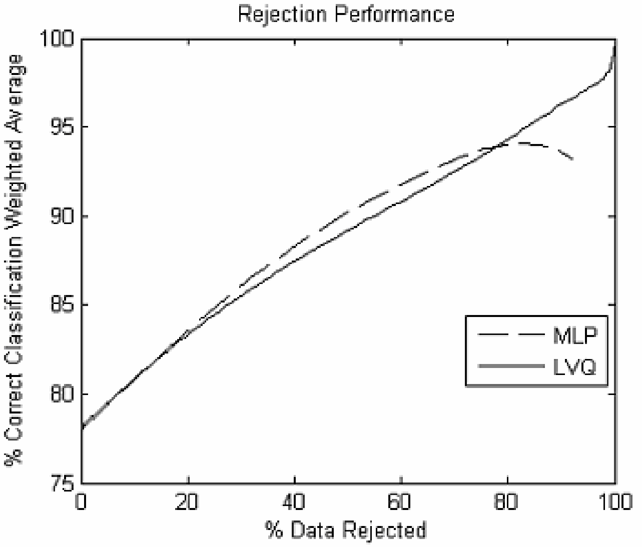

Figure 3. The rejection performance curves for the multilayer perceptron and learning vector quantization neural networks. The percent of the data rejected is plotted against the percent correct classification of the weighted average of asleep and awake samples

C. Distribution of data

The difference in classification between the training, validation, and testing sets is very small suggesting that the data is similarly distributed between the three sets. Looking at Figure 3 we can see that MLP and LVQ offer almost identical results. MLP has slightly better overall classification abilities.

The way the data is distributed between sleep and wake can be seen when we look at the reliability. Figure 4 includes three graphs. The first graph is a frequency plot of

TABLE 2. LVQ NEURAL NETWORK PERFORMANCE: ZERO REJECTION COMPARED TO 30% REJECTION

		Training		Validation		Testing	
		Zero Rejection	30% Rejection	Zero Rejection	30% Rejection	Zero Rejection	30% Rejection
Sleep	Epochs	39827	28929 (72.6%)	37840	27548 (72.8%)	38121	28157 (73.9%)
	% Correct	80.3	88.2	81.1	89.3	79.8	87.2
Wake	Epochs	17742	11162 (62.9%)	18958	11601 (61.2%)	18076	11572 (64.0%)
	% Correct	71.9	76.8%	73.2	79.6	74.5	80.9
Total	Epochs	57569	40091 (69.6%)	56798	39149 (68.9%)	56197	39729 (70.7%)
	% Correct	77.7	85.1	78.4	86.5	78.1	85.4

the reliability measure from MLP for all of the correctly classified samples. The x-axis value ranges from zero to one with zero being the least reliable and one being the most reliable. The graph is then split into two other plots for misclassified epochs. The second plot is where ASP epochs are classified as AWK epochs, and the third plot is when AWK samples are classified as ASP epochs. Please note the difference in the frequency scales on the y-axis, and that the correctly classified samples have a very large spike at reliability measure equal to one. To interpret the three graphs a vertical line is drawn at the reliability threshold. For example a threshold of 0.31 was used to reject 30% of the MLP data. Any epoch that has a reliability measure less than this value is rejected. Notice for the correctly classified epochs the majority is above the threshold and will be considered “reliable”. For incorrectly classified epochs a significant portion has a reliability measure below the threshold and will be considered “unreliable”.

IV. DISCUSSION

The two neural networks presented in this paper offer very similar results with the MLP network slightly better than the LVQ network. Figure 3 shows that the MLP performance is generally higher than the LVQ until more than 80% of the data is rejected.

The determination of sleep state using heart rate parameters in previous studies has a range from 82% (25 infants with different discriminant functions at 1, 2, 3, 4, and 6 months of age) [15] to with wavelet packet modeling 75%-90% (1 infant at 2, 3, 4, and 5 months of age) correct classification [16]. Both studies account for age by creating different classifiers for different age groups. Our study uses one general classifier on a population with mean post conception age (PCA) of 45.5 weeks and a range from 34.3 PCA to 62.3 PCA, and can achieve 85-86% correct classification. Lisenby et al. use heart rate data to separate REM from non-REM states [17]. Other studies have been done to see how heart rate changes during sleep [18, 19, 20]. Our results confirm that an automated technique is plausible to distinguish between AWK and ASP in infants.

In general, AWK prediction using HRV measures may be difficult due to lack of motion (infant being wake and resting quietly) and low, steady HR.

V. CONCLUSION

In conclusion, a method has been developed for sleep/wake classification that incorporates reliability. After systematic rejection of difficult to classify segments, this model can achieve 85%-86% correct classification while rejecting only 30% of the data. This is an improvement of about 7.8% over a traditional model without rejection. This level of prediction may be useful for physiologic data where EEG is not available. Since the distributions for the three datasets (191 infants: over 1500 hours of data) are very similar, a general model is justified. This study is a step towards sleep state classification in infants with a limited set of physiologic parameters.

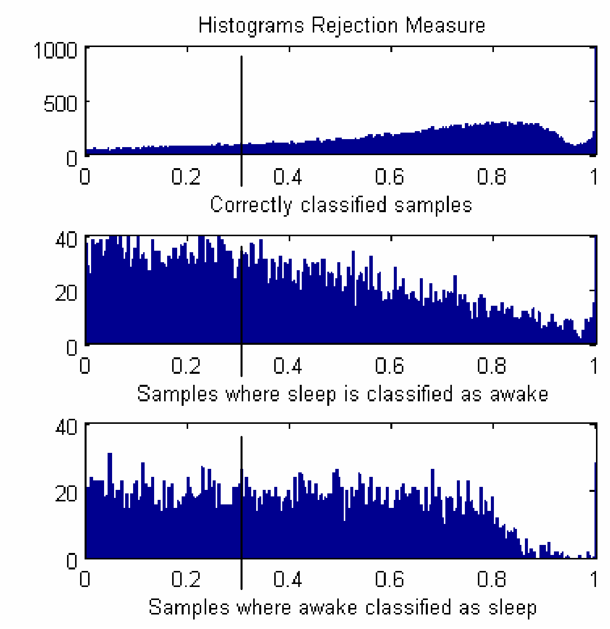

Figure 4. Frequency plots of the reliability measure from the multilayer perceptron neural network. The x-axis represents the range of the reliability measure. Zero is the most unreliable and one is the most reliable. The y-axis is the frequency of occurrence. A vertical line is drawn at the example threshold value, 0.31, used for rejection.

REFERENCES

[1] X. Xueyan, "Prediction of life-threatening events in infants using heart rate variability measurements," Ph.D. dissertation, West Virginia University, 2002.

[2] R.A. Landes, M.S. Scher, M. Sun, R.J. Sclabassi, "Characterization of heart rate dynamics in infants as a probe for neural state and age," in *18th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, IEEE/EMBS'96,* pp. 1662-1663.

[3] C. Medigue, J. Bestel, S. Renard, et al., "Discrete wavelet transform applied to heart rate variability analysis in iron-deficient anemic infant," in *Proc. of 19th Int. Conf. on IEEE/EMBS'97*, pp. 1613-1616.

[4] J. Kantelhard, A. Yosef, P. Ivanov, A. Bunde, S. Havlin, T. Penzel, J.H. Peter, H. Stanley, "Characterization of sleep stages by correlations in the magnitude and sign of heartbeat increments," *Physical Review E*, 65:051908:1-6, 2002.

[5] T. Hoppenbrouwers, M. Neuman, M. Corwin, et al., "Multivariable cariorespiratory monitoring at home: Collaborative Home Infant Monitoring Evaluation (CHIME)," in *Proc. 18th Annual Conf. Of the IEEE Engineering in Medicine and Biology Society, IEEE/EMBS'96,* Amsterdam.

[6] M. R. Neuman, "Cardiopulmonary monitoring at home: the CHIME monitor," *Physiol. Measurements*, vol. 22, no. 2, pp. 267-286, 2001.

[7] D. H. Crowell, L. J. Brooks, T. Colton, M. J. Corwin, T. T. Hoppenbrouwers, C. E. Hunt, CHIME Steering Committee, et al., "Infant polysomnography: reliability," *Sleep*, vol. 20 pp. 553-560, 1997.

[8] C. De Stefano, C. Sansone, M. Vento. "To Reject or Not to Reject: That is the Question-An Answer in Case of Neural Classifiers," *IEEE Trans on Sys, Man, Cybe*r, vol. 30, no. 1, pp. 84-94, 2000.

[9] X. Xu, S. Schuckers, CHIME study group. "Automatic detection of artifacts in heart period data," *J. Electrocardiol*, vol. 33, pp. 205-210, 2001.

[10] J. C. Bezdek. "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York, 1981.

[11] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney, J. C. Gillin, "Automatic sleep/wake identification from wrist actigraphy," *Sleep*, vol. 15, pp. 461-469, 1992.

[12] N. Ampazis, S.J. Perantonis, "Two Highly Efficient Second-Order Algorithms for Training Feedforward Networks," *IEEE Trans NN,* vol. 13, no. 5, pp. 1064-1074, 2002.

[13] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, K. Torkkola, *LVQ PAK: The Learning Vector Quantization Program Package*. Laboratory of Computer and Information Science, Helsinki University of Technology, Finland, 1995.

[14] T. Kohonen, *Self-Organizing Maps*. Springer Series in Information Sciences, vol. 30, NY: 200; Third Extended Ed.

[15] R. Harper, V. Schechtman, K. Kluge, "Machine classification of infant sleep state using cardiorespiratory measures," *Electroencephalogr Clin. Neurophysilogy,* vol. 67, pp. 379-387, 1987.

[16] G. Nason,, T. Sapatinas, A. Sawczenko, "Wavelet packet modeling of infant sleep state using heart rate data," *Sankhay: The Indian Journal of Statistics,* vol. 63, pp. 199-217, 2001.

[17] M. Lisenby, P. Richardson, A. Welch, "Detection of cyclic sleep phenomena using instantaneous heart rate," *Electroencephalogr Clin. Neurophysilogy*, vol. 40, pp. 169-177, 1976.

[18] A. Welch, P. Richardson, "Computer sleep stage classification using heart rate data," *Electroencephalogr Clin. Neurophysilogy,* vol. 34 pp. 145-152, 1973.

[19] V. L. Schechtman, R. K. Harper, R. M. Harper, "Development of heart rate dynamics during sleep-wake states in normal infants," *Pediatric Research,* vol. 34, no. 5, pp. 618-623, 1993.

[20] R. M. Harper, D. S. Kelly, D. O. Walter, T. Hoppenbrouwers, "Cardiac waveform alterations during sleep in the infant," *Psychophysiology*, vol. 13, no. 4, pp. 318-322, 1976.

[21] R. Duda, P. Hart, D. Stork. *Pattern Classification.* Wiley-Interscience Publication, NY: 2001; Second Ed.