Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm

Chen Chen, Adrien Ugon, Chenglu Sun, Wei Chen, Carole Philippe, Andrea Pinna

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 12, 2018, accepted December 3, 2018, date of publication December 17, 2018,
date of current version January 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2887082

Towards a Hybrid Expert System Based on Sleep Event's Threshold Dependencies for Automated Personalized Sleep Staging by Combining Symbolic Fusion and Differential Evolution Algorithm

CHEN CHEN[1], ADRIEN UGON[2], CHENGLU SUN[1], WEI CHEN[1,3], (Senior Member, IEEE), CAROLE PHILIPPE[4], AND ANDREA PINNA[5]

[1]Center for Intelligent Medical Electronics, Department of Electronic Engineering, Fudan University, Shanghai 200433, China
[2]École SupÉrieure d'Ingénieurs en Électronique et Électrotechnique-Paris, 93160 Noisy-le-Grand, France
[3]Shanghai Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention, Shanghai 200032, China
[4]Unité des Pathologies du Sommeil, Groupe Hospitalier Pitié Salpêtrière, 75012 Paris, France
[5]Laboratoire d'Informatique de Paris 6, Sorbonne UniversitÉ, 75005 Paris, France

Corresponding authors: Wei Chen (w_chen@fudan.edu.cn) and Andrea Pinna (andrea.pinna@lip6.fr)

This work was supported in part by the IUIS (Institut Universitaire d'Ingénierie en Santé) of Sorbonne Universités and the CSC (ChinaScholarship Council), in part by the China Postdoctoral Science Foundation under Grant 2018T110346 and Grant 2018M632019, in part by the National Key R&D Program of China under Grant 2017YFE0112000, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01.

ABSTRACT Identification of sleep stages is a fundamental step in clinical sleep analysis. Existing automatic sleep staging systems ignore two major issues: 1) Most of existing automatic sleep staging systems are using numerical classification methods without involving medical knowledge. These kinds of systems are not yet understood and accepted by physicians. 2) Individual variability sources are ignored. However, individual variability is observed in many aspects of sleep research (such as polysomnography recordings, sleep patterns, and sleep architecture). In this paper, a hybrid expert system is proposed to mimic the decision-making process of clinical sleep staging in accordance with the medical knowledge by using symbolic fusion. To formalize the medical guideline and knowledge, thresholds are used for translating the sleep events into symbols and the sleep event's threshold dependencies are analyzed for fully understanding the thresholds dependencies among different sleep stages and subjects. Meanwhile, the differential evolution algorithm is adopted to automate the setting-up of thresholds that are used in the symbolic fusion model and to provide personalized thresholds, which allows taking the individual variability into consideration. The robustness and clinical applicability of the proposed system are evaluated and demonstrated on a clinical dataset. The dataset is composed of 16 patients (nine males and seven females) and scored by physicians. Only 5% of the dataset is used for the training process to obtain the personalized thresholds. Then, these personalized thresholds are passed to the classification process, and the overall accuracy on the identification of five sleep stages reaches 80.09%. Using a small dataset for the training process, the proposed system not only drastically reduces the training set but also achieves favorable results compared with most of the existing works.

INDEX TERMS Knowledge-based system, symbolic fusion, personalization, automatic sleep staging system.

I. INTRODUCTION

Sleep consumes around one-third of our lives. However, sleep disorders involving signs and symptoms like excessive daytime sleepiness, irregular breathing or increased movement during sleep, difficulty in sleeping, and abnormal sleep behaviors are affecting more and more people [1], [2]. For the diagnosis and treatment of sleep disorders, an overnight polysomnographic (PSG) test including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), is usually done via a technician by

placing sensors on the patient's body. Based on recorded PSG signals, a detailed analysis and interpretation will be provided by a physician with recommendations for the diagnosis and treatment. Sleep staging, as a fundamental step of PSG interpretation, is usually visually performed by the physician according to the American Academy of Sleep Medicine (AASM) manual, considered as the international reference guidelines in sleep medicine. Based on the AASM manual, each 30 seconds of recordings — called epoch — can be classified into five different sleep stages, including wakefulness (stage W), Non-Rapid Eye Movement (stages N1, N2, and N3) and Rapid Eye Movement (stage R) [3].

Clinical sleep staging is a time-consuming task. Normally, physicians need 1 to 4 hours to visually score an overnight PSG recording into different sleep stages. Meanwhile, inter-rater variability concerns also exist due to subjective interpretation and decision by different physicians. In [4], 80.6% and 82.0% inter-rater agreements were reported by using Rechtschaffen & Kales (R&K) (old reference guidelines for sleep study) and AASM (new reference guidelines for sleep study), respectively. In order to reduce the burden of physicians, automatic sleep staging systems have attracted extensive attention. Numerous attempts have been undertaken to automate the interpretation of PSG recordings.

Regarding the existing automatic sleep staging systems, they can be roughly categorized into two types, namely machine learning-based systems and knowledge-based expert systems. In machine learning-based systems, a wide range of typical machine learning methods have been applied for the sleep staging, like Decision Tree (DT) [5]–[7], Support Vector Machine (SVM) [8]–[10], Artificial Neural Network [11], [12], etc. Meanwhile, the combinations of two machine learning methods have also been explored in [13] and [14]. Besides these methods, deep learning methods like Convolutional Neural Networks (CNN) [15]–[17], Long Short-Term Memory (LSTM) [18] are increasingly used for the sleep stages classification in recent years. Most of these studies emphasize the importance of selecting a suitable classifier or suitable parameters for the network to improve the accuracy of the classification. Thus, they may gain better or competitive performance in the sleep staging field. However, they may ignore the consideration or acceptance from the physician or the medical perspectives. Without taking medical knowledge or physicians experience into consideration, machine learning methods are used to learn the patterns between features and corresponding stages classes. The pattern recognition is usually established by interaction with a set of training data. Patterns used in the classification of sleep stages are mainly dependent on the features extracted from raw data. Insignificant patterns may be selected independently of medical knowledge and without validation from physicians. Moreover, the resulting predictive model is elaborated so that the application of the model to input data generates the same answer as an expert, but without any consideration of the rules that should be applied to link input data and decision appropriately. Hence, physicians cannot be easily convinced by these kinds of systems without involving any medical knowledge; thus, these kinds of systems are not really used by physicians in clinical practice.

To be compliant with the sleep knowledge or international sleep guidelines, knowledge-based expert systems have also been investigated. These studies dedicated to modeling the medical knowledge by undertaking distinct formalisms, like symbolic fusion framework [19]–[21]. Compared with the machine learning systems, these expert systems may lose some competitive in terms of the performance, which mainly because not all the medical knowledge and expert experience can be exhaustively formalized. While these kinds of expert systems conform to the decision-making process of the visually PSG interpretation, which can be understood and validated by the physicians. Meanwhile, preliminary studies of using symbolic fusion framework have proved their possibility and feasibility in sleep staging applications [19]–[21]. However, there still exist several issues which need to be addressed in order to enrich this knowledge-based sleep staging system: 1) Thresholds were used to translate the sleep events into symbols. While manual interpretation of these thresholds was adopted and the sleep event's threshold dependencies were ignored; 2) Not all the sleep events described by AASM manual were implemented; 3) Pre-processing of PSG signals and smoothing of Hypnogram were not included. In comparison to machine learning-based methods, symbolic fusion rigorously obeys the medical guidelines from the sleep events extraction to the decision through the modelization of the inference framework. From the physician's perspective, symbolic fusion-based sleep staging system can be easily understood and validated by translating and formalizing the AASM guidelines into computer logic.

However, individual variability sources were not taken into account in both machine learning-based methods and symbolic fusion-based method. While individual variability was observed in many aspects of sleep research (such as PSG recordings, sleep patterns, sleep architecture, etc.). To improve the accuracy of the sleep staging system, individual variability should be taken into consideration. Is it possible to have a combined system which is able to compute the symbolic fusion thresholds automatically and to take the individual variability into consideration?

In this paper, a hybrid expert system is proposed based on the symbolic fusion. It aims to assist the physicians in sleep analysis and diagnosis. To conceive this system, firstly, a detailed sleep event's threshold dependencies analysis is performed to fully understand the thresholds used for translating the sleep events into symbols. Secondly, to compute and automate the thresholds setting-up procedure, an automatic thresholds setting-up method using differential evaluation algorithm is evaluated. Finally, the hybrid expert system composed by combining symbolic fusion model and differential evolution algorithm is put forward. The new paradigm of the proposed system allows an automated personalized sleep staging by taking into account the individual variability.

The rest of this paper is organized as follows: Section II describes an existing symbolic fusion-based sleep staging system. Section III analyzes the thresholds and their dependencies on sleep stages. Section IV shows how a differential evolution algorithm could be used to compute thresholds value for the symbolic fusion model and how it needs to be set-up. Section V shows how we have taken into account the individual variability. Section VI shows the overall hybrid expert system by combining the symbolic fusion and the differential evolution algorithm. Based on the proposed system, a personalized sleep staging system is presented. Section VII presents the evaluations of the personalized sleep staging system. Followed by a brief discussion on the results which can be found in Section VIII. At last, the conclusion is presented in Section IX.

II. SLEEP STAGING SYSTEM BASED ON SYMBOLIC FUSION

In this section, a brief description of symbolic fusion and symbolic fusion-based sleep staging system is presented. It is followed by the analysis of the limitations in the existing symbolic fusion-based sleep staging system.

Symbolic fusion is an efficient decision-making technique involving interdisciplinary methods among signal processing, artificial intelligence, inference, statistics and so on. It has been widely applied in image processing [22], [23] or medical analysis [24], which proved it to be efficient to fuse information from different sources.

The three-level Dasarathy architecture allows to abstract information using symbolic fusion. Three layers are considered: Data, Features, and Decision. Inference rules fuse information from one level to the next level so that, at the end, we can take a decision.

A. SYMBOLIC FUSION-BASED SLEEP STAGING SYSTEM (SF-SSS)

An existing system based on symbolic fusion has been designed to realize the classification of sleep stages [19], [20], which follows Dasarathy architecture, as shown in Fig. 1. It starts from the extraction of sleep events (digital parameters) from raw PSG signals and goes up-to high-level symbolic interpretation of feature parameters. Finally, rules are used to make the decision. Digital parameters, symbolic interpretation, and rules in SF-SSS are inspired by international guidelines in sleep medicine. A brief introduction is presented below, more details can be found in our previous work [19], [20].

1) DATA FUSION

In data fusion, nine digital parameters are extracted using time-domain, frequency-domain, and non-linear analysis. These parameters are used to symbolize the sleep events which are described in the AASM manual. In AASM, sleep events like sleep spindle, K complex, chin EMG tone, eyes movement, etc. are described for guiding the physicians to perform manual interpretation of sleep stages.

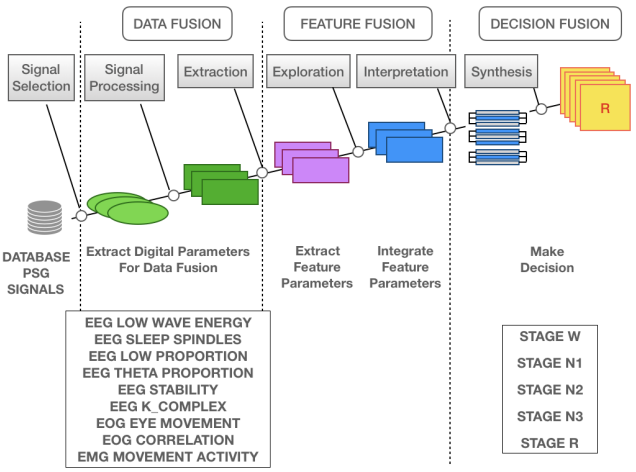


FIGURE 1. Framework of symbolic fusion-based sleep staging system.

These sleep events can be visually observed from the polysomnographic curves. To model the symbolic fusion framework which obeys the medical guideline from the sleep event extraction to the decision-making, the digital parameters are extracted in the data fusion to symbolize the sleep events. As mentioned in our previous work [20], nine digital parameters, namely *EEGLowWaveEnergy*, *EEGSleepSpindles*, *EEGLWProportion*, *EEGThetaProportion*, *EEGStability*, *EOGEyeMovement*, *EOGCorrelation*, and *EMGMovementActivity* were extracted. These digital parameters were used to symbolize the sleep events, like sleep spindle, K complex, chin EMG, etc. that described in AASM to characterize the sleep events for distinguishing different sleep stages. A detailed description of each parameter was described in [20].

2) FEATURE FUSION

In feature fusion, digital parameters are transformed into symbolic features using thresholds. Nine digital parameters are transformed into 24 features via 15 thresholds as shown in Table 1. These thresholds are inspired by the medical

TABLE 1. Thresholds used in SF-SSS model.

Digital Parameters	Features	No. TH[a]
EEGLowWaveEnergy	High - Middle - Low	2
EEGSleepSpindles	Confidently Have - Not Confident	1
EEGLWProportion	High - Low	1
EEGThetaProportion	High - Low	1
EEGStability	Stable - Not Confident - Unstable	2
EEGKComplex	High - Middle - Low	2
EOGEyeMovement	High - Middle - Low - Lowest	3
EOGCorrelation	Conjugate - Disconjugate	1
EMGActivity	High - Normal - Low	2

[a]Number of Thresholds.

knowledge guidelines AASM manual. E.g. in the AASM, the chin EMG activity is described as three different semantic descriptions, namely *High*, *Normal* and *Low*. Thus, two thresholds are used to transform the *EMGActivity* digital parameter into three symbols (*High*, *Normal* and *Low*).

3) DECISION FUSION

In decision fusion, inference method is used to aggregate symbols in order to recognize one of the sleep stages. In this way, we can consider having an expert system for each sleep stage described by symbols aggregation. Each expert system is executed one after the other [W, N2, N3, R]. The execution of one of the expert systems depends on the result obtained by the previous system. As the expert system recognizes its stage, the execution of the remaining expert systems will not be performed. If any system expert recognizes a sleep stage, the N1 stage will be selected for exclusion.

Stage N1 is considered as a transition between wake and sleep. It occurs upon falling asleep and during brief arousal periods within sleep and usually accounts for 2 – 5% of total sleep time. The detection of N1 is always a problematic aspect of the sleep stages in both clinical sleep staging and automatic sleep staging system. Only 63.0 % inter-scorer reliability for stage N1 is reported among different scorers in [25]. Moreover, finding a significant feature that could separate N1 from W, N2, N3, and R, is rather difficult for automatic sleep staging system, because N1 is a transition phase in the changes of wakefulness and other sleep stages. In this paper, a classifier for stage N1 is proposed as shown in Fig. 2. It is proposed under the guideline of AASM and the medical knowledge and experience provided by the physicians while using existing digital and feature parameters. Once *EMGActivity* is *Low*, *EEGSleepSpindles* is *Not Confident*, *EEGStability* is *Not Confident*, *EEGLowWaveEnergy* is *Low* and *EEGKComplex* is *Low* then this epoch can be considered as stage N1.

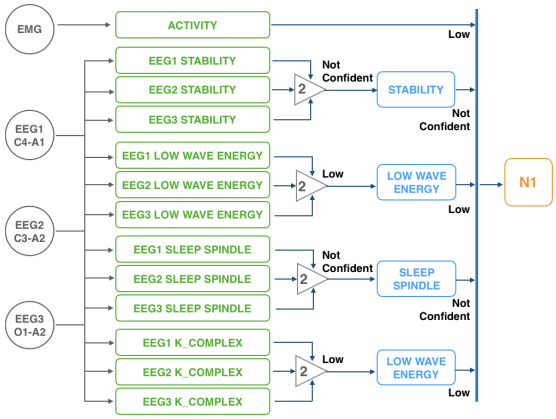


FIGURE 2. Classifier for stage N1 using symbolic fusion.

B. LIMITATIONS OF SF-SSS

In [19] and [20], interpretations of digital parameters into their symbolic features are performed manually by authors via thresholds. Thresholds are widely applied in decision support systems, not only in [19] and [20] but also in [26] and [27], for transforming digital parameters into high-level features — linguistic or symbolic — to realize interpretation included in the inference process under the guideline of medical knowledge. While in clinical practice, boundaries of linguistic or symbolic features are very flexible. Physicians may adjust the boundaries for each linguistic or symbolic feature according to their experience and patient information. However, as far as we know, there is no fully satisfying automatic setting - up thresholds method in the existing automatic sleep staging systems. Manually predefined values of thresholds have been widely used due to the following reasons:

- Thresholds dependencies should be carefully considered;
- Building a mathematical model or a threshold function in setting-up thresholds is very challenging as it requires a set of data with sufficient quantity and adequate quality;
- There is a lack of uniformity between subjects, how take it into account?

The following part of the paper shows how we have addressed these limitations and found a solution that can generate the hybrid expert system.

III. SLEEP EVENT'S THRESHOLD DEPENDENCIES ANALYSIS

Before analyzing which algorithm is better for the thresholds computation, we need to understand what are the thresholds dependencies in order to know how many thresholds are needed. For that, in this section, thresholds dependencies among sleep stages are discussed. In the analysis of thresholds dependencies among sleep stages can help us to understand whether same thresholds can be used in different sleep stages or different thresholds are required for different sleep stages.

In this section, we describe how to define and use thresholds to transform digital parameters into the symbolic interpretation of feature parameters. According to AASM rules, we analyze how it is possible to generate different symbols from one sleep event through thresholds.

A. DESCRIPTION IN AASM

Taking chin EMG as an example, there exist several rules in AASM which are described below.

- *Rule E3.c* Score epoch as stage W when Irregular, conjugate rapid eye movements associated with **normal or high chin muscle tone**.
- *Rule F.N3* During stage N1, **the chin EMG amplitude** is variable, but often **lower than in stage W**.
- *Rule I.2b* Score stage R sleep in epochs with the following phenomena: **low chin EMG tone** for the majority of the epoch.

In AASM, chin EMG/chin muscle tone has been mentioned in three rules for guiding physicians to score stage W, N1 and R respectively.

B. FROM AASM TO SF-SSS MODEL USING THRESHOLDS

In order to formalize these rules into SF-SSS model, the description of **chin muscle tone**, **the chin EMG amplitude** and **chin EMG tone** in AASM are assumed as the same sleep event which represents the *chin EMG activity* in SF-SSS. Data Fusion, Feature Fusion and Decision Fusion on *chin EMG activity* are described as follows.

- Data Fusion of *chin EMG activity*
  In SF-SSS, digital parameter *EMGActivity* is extracted by using the mean absolute value of the chin EMG signal as shown in (1), where $x(t)$ is the chin EMG signal.

$$EMGActivity = mean\,(abs\,(x\,(t))) \tag{1}$$

  This parameter can be used to indicate the activity level of chin EMG, which can be used as an indicator of the chin muscle tone during sleep as shown in Fig. 3. It shows the digital parameter *EMGActivity* of subject 3774 as an example. For stage W, the digital parameter *EMGActivity* is relatively high as shown in black box; for stage N1 and R, *EMGActivity* is relatively low as shown in green and pink boxes respectively.

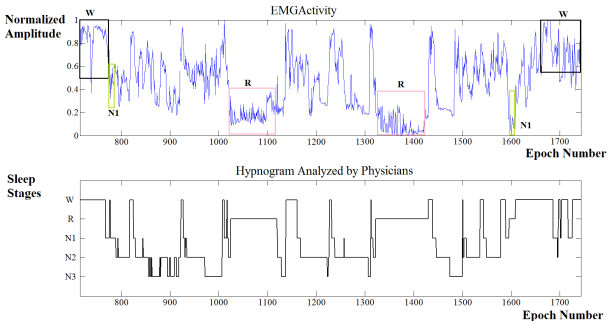


FIGURE 3. Digital parameter: EMGActivity.

- Feature Fusion of *chin EMG activity*
  In feature fusion, thresholds are used to transform digital parameters into symbolic interpretation of feature parameters. To build the correspondence between symbolic interpretation of feature parameter with AASM manual, three symbolic interpretations of feature parameters are used: *High*, *Normal* and *Low*.
  To transform *EMGActivity* into symbolic interpretation of *High*, *Normal* and *Low*, two thresholds *EMGTh1* and *EMGTh2* are used in Fig. 4. Values of digital parameter *EMGActivity* higher than *EMGTh1* are interpreted as *High*, values between *EMGTh1* and *EMGTh2* are interpreted as *Normal*, values lower than *EMGTh2* are interpreted as *Low*.
- Decision Fusion of *chin EMG activity*
  In decision fusion, rules inspired by AASM to make decisions are shown below.
  For stage W: *EMGActivity* is *Normal* or *High* (in addition to other required criteria).
  For stage N1: *EMGActivity* is *Low* (in addition to other required criteria).
  For stage R: *EMGActivity* is *Low* (in addition to other required criteria).

In SF-SSS model, digital parameters (e.g.*EMGActivity*) in the low-level, interpretation (e.g.*High*, *Normal*, *Low*) and decision rules in the high-level are inspired from AASM. However, in the medical guideline, there is no definition or description of the process for transforming low-level digital parameters into high-level symbolic interpretation. In other words, from the medical guideline, there is no definition or description for guiding the setting-up of these thresholds to realize the interpretation, physicians may adjust these thresholds according to their experience. The setting up of the thresholds are the simulation of the decision-making process of the physician for translating the sleep events into the corresponding semantic symbols. To better understand the thresholds, dependencies of thresholds on sleep stages are discussed in the following parts.

C. DEPENDENCIES OF THRESHOLDS ON SLEEP STAGES

In SF-SSS model, nine digital parameters are transformed into 24 features via 15 thresholds as shown in Table 1. Take *EMGActivity* as an example, two thresholds *EMGTh1* and *EMGTh2* are used to distinguish three different symbolic features: *High*, *Normal* and *Low* as shown in Fig. 4. Then, these symbolic features are used in classifying stage W, N1 and R. In previous work [19], [20], [28], [29], values of *EMGTh1* and *EMGTh2* are the same for each stage without considering thresholds dependencies.

However, in practice, we observed with physicians that the threshold values for transforming the same digital parameter into a same symbolic feature for two different sleep stages are different. E.g. for classifying stage N1, 0.8 and 0.6 can be considered as the appropriate values for thresholds *EMGTh1* and *EMGTh2* respectively, as shown in Fig. 4. While for classifying stage W and R, decreasing *EMGTh2* value can allow more precise classification results by reducing misclassified stages of stage W and R. 0.55 and 0.38 are con-

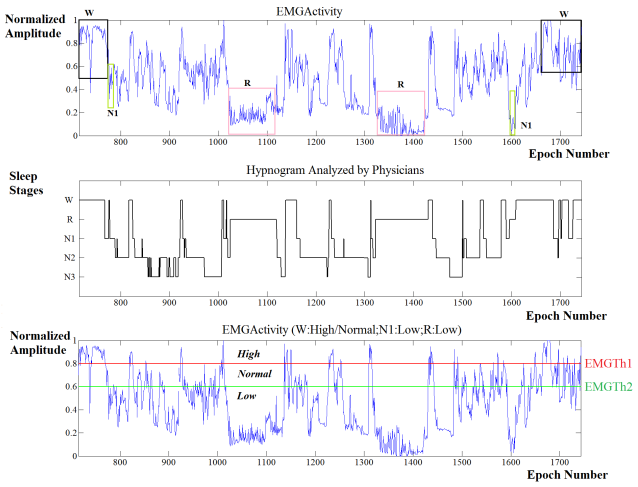


FIGURE 4. Thresholds for EMGActivity.

sidered to be more suitable for classifying stage W and R, respectively. So that, thresholds dependencies exist and values of *EMGTh2* for classifying stage W, N1 and R should be different. As for *EMGTh1*, from the technical point of view, no matter how we adjust *EMGTh1*, it has no impact on the classification result for stage W, N1 or R. *EMGTh1* is unnecessary for SF-SSS model, and in the view of the complexity of SF-SSS, *EMGTh1* can be subtracted. Thus, three thresholds are required by taking thresholds dependencies on sleep stages into consideration. Fig. 5 presents the appropriate threshold *EMGTh2* of *EMGActivity* we observed and verified with the physician for each stage (stage W, N1 and R) of the 16 subjects.

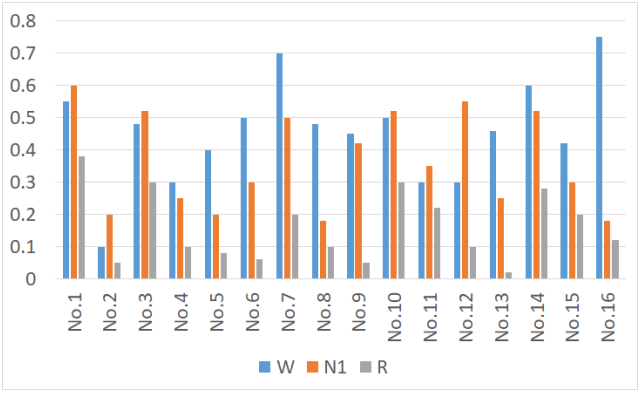


FIGURE 5. Threshold *EMGTh2* of *EMGAcitivity* visually observed and verified with physicians.

D. DETERMINATION OF THRESHOLDS CONFIGURATION

In this paper, only the digital parameter *EMGAcitivity* is taken as an example for the thresholds dependencies analysis. For the remaining eight digital parameters, thresholds dependencies are also analyzed. By taking these thresholds dependencies into consideration, the number of thresholds which need to be set-up for each stage is listed in Table 2. E.g. to classify stage W, five digital parameters are extracted and five corresponding thresholds are required in transforming these digital parameters into different symbolic features. For stage N1, N2, N3, and R, the number of thresholds needed to be set-up are 6, 8, 7 and 8 respectively. Totally 34 thresholds are required to be set-up.

After figuring out the number of thresholds that need to be set-up while considering the thresholds dependencies, a method to compute the thresholds setting-up is required and proposed in the next section.

IV. THRESHOLDS SETTING-UP FUNCTION

Stochastic Search Algorithms (SSAs) have been widely used in solving combinatorial optimization problems. In [30], SSAs have been applied to navigate through the large parametric space of different drugs to identify optimal low-dose drug combinations in manipulating the cellular network toward a therapeutic goal. In [31], SSAs are used to find optimal input variable combination for guiding the complex system toward the desired state. Recently, SSAs are explored to find the optimal parameters for different machine learning models (e.g. SVM, Extreme Learning Machine, etc.) in different applications, like the diagnosis of Alzheimer's disease [32], load forecasting [33], and image processing [34].

TABLE 2. Thresholds configuration for each sleep stage.

	Digital Parameters	Symbols	No. TH
Stage W	EEGThetaProportion	Low	5
	EEGStability	Unstable	
	EEGKComplex	High	
	EOGEyeMovement	High(Middle)	
	EMGActivity	High(Normal)	
Stage N1	EEGStability	Not Confident	6
	EEGLowWaveEnergy	Low	
	EEGKComplex	Low	
	EEGSleepSpindles	Not Confident	
	EMGActivity	Low	
Stage N2	EEGThetaProportion	High	8
	EEGStability	Stable	
	EEGLowWaveEnergy	Low	
	EEGKComplex	Middle	
	EEGSleepSpindles	Confidently Have	
	EOGEyeMovement	Low(Lowest)	
	EOGCorrelation	Disconjugate	
Stage N3	EEGLWProportion	High	7
	EEGStability	Stable	
	EEGLowWaveEnergy	High	
	EEGSleepSpindles	Not Confident	
	EOGEyeMovement	Low(Middle)	
	EOGCorrelation	Disconjugate	
Stage R	EEGThetaProportion	Low	8
	EEGLWProportion	Low	
	EEGStability	Not Confident	
	EEGLowWaveEnergy	Low	
	EEGSleepSpindles	Not Confident	
	EOGCorrelation	Conjugate	
	EMGActivity	Low	

There exists several typical SSAs, like *Tabu Search*, *Gur Game*, *Simulated Annealing*, *Switching Particle Swarm Optimization*, *Differential Evolution* and *Cross Entropy*, each algorithm has its own pros and cons. Instead of building a mathematical model or a threshold function from scratch, new solutions to solve thresholds setting-up problems have been presented in [28] and [29] by using stochastic search algorithms. The thresholds setting-up problems can be described as a combinatorial optimization problem that aims at finding the optimal thresholds combination among possible thresholds combinations space regarding the objective value of sleep staging systems. Among the typical SSAs, *Differential Evolution* is adopted.

A. DIFFERENTIAL EVOLUTION MODEL

Differential Evolution (DE) was proposed by Stron and Price [35] in 1997. As an effective and efficient stochastic optimization technique, it has been successfully applied in diverse domains [30], [36].

In this paper, DE was applied due to the following advantages: 1) DE can mimic natural biological evolution and provide a fast and stable convergence. 2) It is less sensitive to the initial population. 3) It is a parallel search method. 4) It can improve objective function value iteratively. To deal with optimization problems, DE starts with a set of initial population (as parents) which is usually drawn randomly from the uniform distribution within the variable space. Then DE operators (*mutation* and *crossover*) are applied consecutively to each individual in the population to produce another population (as offspring). Both populations are then evaluated using a fitness (objective) function. The subjects who obtain the higher values of fitness (objective) function survive for further *reproduction*, *evaluation*, and *selection* until the *termination criterion* is met.

A brief procedure description of DE is introduced as follows, more details of DE can be found in [35].

- **Initialization**
  As a population based search algorithm, DE starts with the initial population vector $X_{i,G} = \left\{x^1_{i,G}, x^2_{i,G}, \ldots, x^D_{i,G}\right\}$, where the index i denotes the i^{th} individual of the population ($i \in \{1, 2, \ldots, NP\}$), NP denotes population size, D is the dimension of the population, and G denotes the generation to which the population belongs. The initial population is generated using (2), where x^L_j, x^U_j denote the lower and upper limits of the variable of the j^{th} dimension ($j \in \{1, 2, \ldots, D\}$), and $rand(0, 1)$ represents a uniformly distributed random value within $[0, 1]$.

$$x^j_{i,0} = rand(0, 1) \times (x^U_j - x^L_j) + x^L_j \tag{2}$$

- **Mutation**
  Then, $V_{i,G+1} = \left\{v^1_{i,G+1}, v^2_{i,G+1}, \ldots, v^D_{i,G+1}\right\}$, as a mutant vector, is generated according to (3). The indexes r_1,r_2,r_3 are mutually exclusive integers randomly chosen within the range $[1, NP]$ and they are all different from base index i. Mutation Scale Factor F is a real and constant factor belonging to $[0, 2]$ which controls the amplification of the differential variation.

$$V_{i,G+1} = X_{r1,G} + F \times \left(X_{r2,G} - X_{r3,G}\right) \tag{3}$$

- **Crossover**
  In order to increase the diversity of the DE population, crossover is introduced. A crossover vector $U_{i,G+1} = \left(u^1_{i,G+1}, u^2_{i,G+1}, \ldots, u^D_{i,G+1}\right)$ is formed. Among the vector, each variable is generated using (4), where $randb(j)$ is the j^{th} evaluation of a random number generator which belongs to $\in [0, 1]$, $rnbr(i)$ is an integer randomly generated from $[1, D]$, and Crossover Rate CR is a crossover constant that belongs to $[0, 1]$. The crossover vector takes the variable $v^j_{i,G+1}$ from the mutation vector when the generated random number is equal or less than the CR and also guarantees at least one variable is from the mutation vector.

$$u^j_{i,G+1} = \begin{cases} v^j_{i,G+1} & \text{if } randb(j) \le CR \quad \text{or } j = rnbr(i) \\ x^j_{i,G} & \text{otherwise} \end{cases} \tag{4}$$

- **Selection**
  To decide whether individual can become a member of Generation $G + 1$ or not, greedy criterion is used by assessing the value of the objective function f on $U_{i,G+1}$ and $X_{i,G}$ as shown in (5), where f is the objective function for evaluating the individuals of the population.

$$X_{i,G+1} = \begin{cases} U_{i,G+1} & \text{if } f\left(U_{i,G+1}\right) \ge f\left(X_{i,G}\right) \\ X_{i,G} & \text{otherwise} \end{cases} \tag{5}$$

B. DIFFERENTIAL EVOLUTION IMPLEMENTATION

To compute the thresholds automatically, the way to implement the differential evolution model into the symbolic fusion system is proposed. As shown in Fig. 6, Automatic Thresholds Setting-Up (ATSU) is mainly involving the following steps:

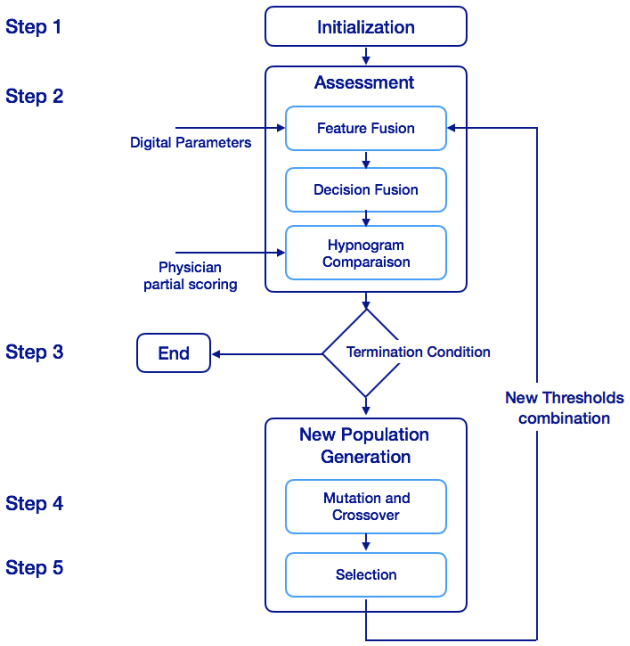


FIGURE 6. Automatic thresholds setting-up model by fusing differential evolution into symbolic fusion.

- **Step 1. Initialization:** Determine several control parameters: *population size*, *mutation scale factor*, *crossover rate*. Generate initial population (initial thresholds combinations). Details of the selection of these control parameters are explained in [**?**].
- **Step 2. Assessment:** The population is assessed by an objective function. In this step, the performance of different thresholds combinations on SF-SSS is evaluated by comparing the whole Hypnogram that analyzed by the physicians with the Hypogram generated using

SF-SSS with different thresholds. F-Measure is used to assess the performance, as shown in (6), which provides a balance between precision and recall.

$$F - Measure = 2 \times \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

Precision (also called positive predictive value) is the ratio of all positive predictions among the predictions, which equals to (TP/(TP + FP)). Recall (also known as sensitivity or true positive rate) is the ratio of positive predictions among true events, which equals to (TP/(TP + FN)). TP, FN, and FP are used to evaluate how good the observations (predictions) reflect the actual events for a classification. TP are observations which were correctly predicted. FN are observations considered as negative where the actual events are positive. FP are observations labeled as positive where the actual events are negative. In the assessment, feature fusion and decision feature are needed to be performed in each iteration. However, regarding the data fusion, it can be performed only once to extract digital parameters at the beginning and to be reused in each iteration, because different thresholds combinations have no impact on it.

- **Step 3. Check:** Verify whether the terminate condition is satisfied or not. If one of the terminate conditions — for instance, objective function reaches the desired value (F-Measure≥0.98) or iteration reaches the pre-defined maximum iteration number (G = 100) — is satisfied, then the procedure stops and the optimal population (optimal thresholds combination) is given.
- **Step 4. Mutation and Crossover:** Generate the provisional population by mutation and crossover operations.
- **Step 5. Selection:** Evaluate the objective function of the provisional population. Compare objective function of the initial population with the objective function of the provisional population to generate new population (new thresholds combinations).
- **Step 6. Repeat:** Repeat from step 2.

For the symbolic fusion framework, totally five stage-specific expert systems are established for classifying stage W, N1, N2, N3, and R. For each specific expert system, the number of thresholds needs to set-up is 5, 6, 8, 7, and 8 respectively, as shown in Table 2. To illustrate, for stage W, totally 5 thresholds needed to be set-up and the detailed thresholds setting-up procedure can be described as follows. Initially, the thresholds combinations are generated randomly. Then these thresholds combinations are passed to the specific expert system and evaluated by assessing the F-Measures on the classification results of the specific expert system. Based on the calculated F-Measures, the terminate conditions are checked. Once one of the F-Measures reaches 0.98 or the iteration number reaches the pre-defined value (G = 100), then the procedure stops. Otherwise, provisional thresholds combinations will be generated by applying mutation and crossover operations. For the provisional thresholds combinations, they will be passed to the specific expert system for evaluating the impact of the thresholds combinations on the symbolic fusion framework by calculating the F-Measures. Then the F-Measures of provisional thresholds combinations are compared with the F-Measures of initial thresholds combinations to select new thresholds combinations for the next generation. The new thresholds combinations will repeat the aforementioned procedure until one of the terminate conditions is satisfied. After mixing the differential evolution into symbolic fusion, the analysis of control parameter and training set selection of differential evolution is described as follows.

1) CONTROL PARAMETER SELECTION OF DIFFERENTIAL EVOLUTION

There are three main control parameters of DE: Population Size (NP), Mutation Scale Factor (F) and Crossover Rate (CR). Details of each control parameter are described as follows and the range of these control parameters are listed in Table 3:

TABLE 3. Main control parameters of differential evolution.

Main Control Parameters of DE	Range
Population Size(NP)	$[5D, 10D]^{b}$
Mutation Scale Factor(F)	$[0, 2]$
Crossover Rate(CR)	$[0, 1]$

[b]Population Size suggested in [35]

- **Population Size (NP)**: NP may play a crucial role in the efficiency and effectiveness of DE. Large population size potentially increases the population diversity. However, when the computational budget is limited, increasing the population size implies to decrease the number of iterations (generations).
- **Mutation Scale Factor (F)**: F controls the amplification of the differential variation. Too small F values increase the risk of premature convergence (i.e. converge to an undesirable point), while too large F values decrease the convergence speed that degrades DE efficiency and may result in early termination.
- **Crossover Rate (CR)**: CR controls the number of components inherited from the mutant vector; it can be interpreted as a mutation probability. Small CR values can boost convergence speed when a few decision variables are interacting with each other. In turn, large CR values are more effective when lots of decision variables are interacting.

The selection of appropriate parameters can affect the efficiency of the ATSU model. Due to the variability of the underlying mathematical properties of different problems, a fixed set of control parameters that suits well for one problem or a class of problems does not guarantee that it will work well for another class, or range of problems [37]. That is, the selection of control parameters is problem dependent.

To ensure the performance of ATSU, selection of control parameters is extremely important.

To select the appropriate control parameters, the trial-and-error approach, is widely used. Several sets of control parameters are tested, then appropriate control parameters are selected based on the average performance of the problem.

In this paper, three different population sizes (NP = 5D, 10D and 500, where D is the number of thresholds which differs for each stage), three different values of mutation scale factor (F = 0.5, 1 and 1.5), three different values of crossover rate (CR = 0.1, 0.5 and 0.9) are analyzed to select optimal control parameters for DE. Fig. 7 shows F-Measure dependence on generation number for classifying stage W with population size 5D as an example. Due to the stochastic ability of DE, F-Measure in Fig. 7 is the mean values of 20 independent runs. As F increases, the convergence speed decreases. Among the three different values of F (0.5, 1 and 1.5), 0.5 provides the best convergence speed as shown in Fig. 7. For the same NP and $F = 0.5$, $CR = 0.9$ is relatively better than $CR = 0.1/0.5$. Population size with 10D and 500 came to the same results (as F increases, the convergence speed decreases).

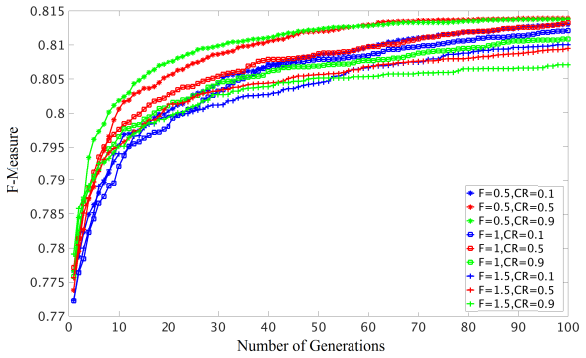


FIGURE 7. DE control parameters selection: NP = 5D.

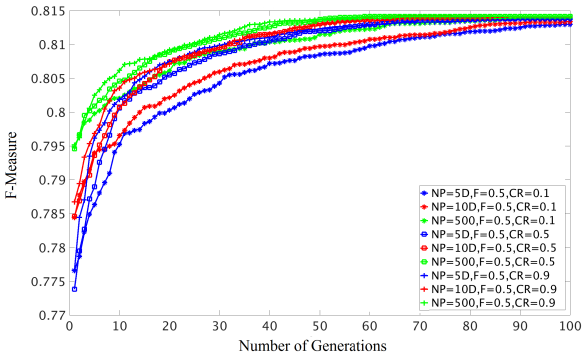


FIGURE 8. DE control parameters selection.

In order to evaluate the impact of different population sizes, Fig. 8 shows the comparison among different NP and CR when $F = 0.5$. Large population size potentially increases the population diversity and may provide fast convergence speed. However, it also increases the computational time. $NP = 500$, $F = 0.5$, $CR = 0.5/0.9$ have faster convergence speed than others. For $NP = 500$, $F = 0.5$, $CR = 0.1$ almost have same convergence speed as $NP = 5D/10D$, $F = 0.5$, $CR = 0.9$.

According to the simulation results, in order to balance F-Measure and computational complexity, $NP = 5D$, $F = 0.5$, $CR = 0.9$ can be suggested as optimal parameters which have less computational complexity compared to $NP = 500$ and have slightly better convergence speed than $NP = 5D$, $F = 0.5$, $CR = 0.1/0.5$.

2) PARTIAL PHYSICIAN SLEEP STAGING SCORING (TRAINING SET EPOCH SELECTION)

Until now all the Hypnogram scored by the physician was used for the assessment objective step, in order to evaluate and select the control parameters of the differential evolution model. Thus, it could not help the physician in the PSG reading if she/he always needs to score all the recorded signals and the automation process loses its meaning. We need to find the minimum number of epochs scored by the physician that is enough to compute the threshold by the differential evolution model and for which the evaluation of the overall sleep staging is good enough (F-Measure greater than 70%). We named this the training set of epochs needed by the differential evolution model. The training set is absolutely necessary to learn the optimized value of thresholds used in the symbolic fusion model for each patient. Different sizes of the training set are analyzed to select the optimal one. Table 4 illustrates the corresponding Recall, Precision, and F-Measure of evaluation set for each sleep stage using 5%, 10%, 15% and 20% as the training set, respectively.

TABLE 4. Recall, precision and F-Measure of evaluation set in accordance to different training set.

		Training Set Epoch Selection			
		Training Set (%) & Evaluation Set (%)			
		5 & 95	10 & 90	15 & 85	20 & 80
W	Recall	75.15	75.20	75.96	76.39
	Precision	67.95	71.59	72.35	72.74
	F-Measure	71.37	73.35	74.11	74.52
N1	Recall	23.96	23.96	25.22	25.23
	Precision	20.49	22.09	23.07	23.86
	F-Measure	22.09	22.99	24.10	24.53
N2	Recall	83.19	85.82	86.03	86.66
	Precision	63.92	64.34	64.61	64.86
	F-Measure	72.29	73.54	73.80	74.19
N3	Recall	78.56	78.61	79.23	79.57
	Precision	69.28	71.47	72.44	73.16
	F-Measure	73.63	74.87	75.68	76.23
R	Recall	66.42	67.76	68.30	68.39
	Precision	64.90	68.60	70.37	70.75
	F-Measure	65.65	68.18	69.32	69.55

The epochs are selected in equal part at the beginning, middle and at the end of recorded signals. As the size of the

training set increases, Recall, Precision, and F-Measure on the evaluation set are marginally increased. For the stage W, N1, N2 and N3, using 5%, 10%, 15% and 20% as the training set, the recalls are slightly higher than the corresponding precisions. It indicates that for these stages with different training set, the classifier finds more false positives than false negatives. For the stage R, using 10%, 15% and 20% as the training set, the precisions are slightly higher than the corresponding recall. To balance the recall and precision of the stages, F-Measure is used for the evaluation. In this paper, for physicians to consume less time in scoring training set, 5% can be considered as the optimal value for the training set. By using the 5% as the training set, the F-Measures of most stages like stage W, N2 and N3 are higher than 70%. Meanwhile, it consumes the minimum time for the physician to score in comparison with using 10%, 15% and 20% as the training set.

V. INDIVIDUAL VARIABILITY: DEPENDENCIES OF THRESHOLDS ON SUBJECTS

After determining the number of thresholds, the dependence of thresholds on subjects has been taken into account. The analysis of thresholds dependencies among subjects can help us to understand whether generalized thresholds are sufficient or specific thresholds for different subjects are required. Individual variability exists in many aspects of sleep. To investigate the thresholds dependencies on subjects, another subject 55341 is taken as an example. For subject 55341, *EMGTh2* = 0.1 can be considered as appropriate thresholds values in classifying stage W as shown in Fig. 9. Values of *EMGTh2* between subject 3774 and 55341 are quite different because of the individual variability of chin EMG signals. In the previous visual analysis in Fig. 5, we observed that thresholds are different among the sixteen subjects. Meanwhile, to verify this observation, threshold *EMGTh2* of *EMGActivity* is calculated using ATSU proposed in this paper, as shown in Fig. 10. Both visual thresholds setting-up method and automatic thresholds setting-up using Differential Evolution prove that thresholds dependencies on subjects exist.

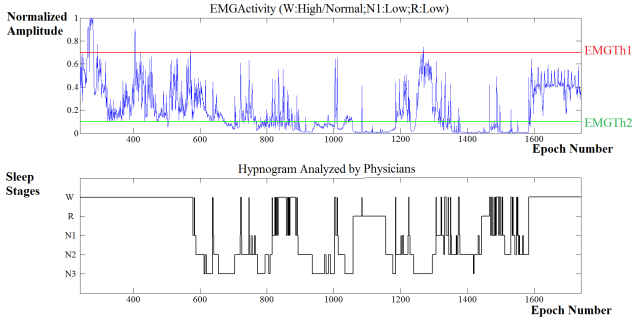


FIGURE 9. Thresholds for EMGActivity (stage W) of patient 55341.

In other words, the symbolic fusion model for the sleep staging is based on generic rules, but the calculation of these

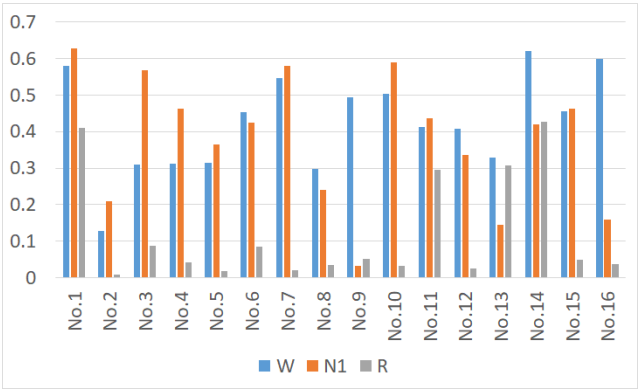


FIGURE 10. Threshold *EMGTh2* of *EMGAcitivity* calculated by ATSU.

thresholds takes the individual variability into consideration. To set-up these thresholds with the differential evolution model, partial scoring of epochs by the physician is required. In this way, the individual variability is taken into account indirectly, which makes a personalized sleep staging possible.

VI. HYBRID EXPERT SYSTEM: THE NEW PARADIGM

Based on the symbolic fusion model and differential evolution algorithm, a hybrid expert system conception is proposed as shown in Fig. 11. We resume in this section how the system works. Firstly, several epochs (5%) will be selected and analyzed by a physician (Hypnogram generation). Based on the selected epochs, thresholds can be set-up using the differential evolution algorithm. Then, these thresholds are used to score all remaining epochs by using the symbolic fusion model and full Hypnogram can be generated at the end.

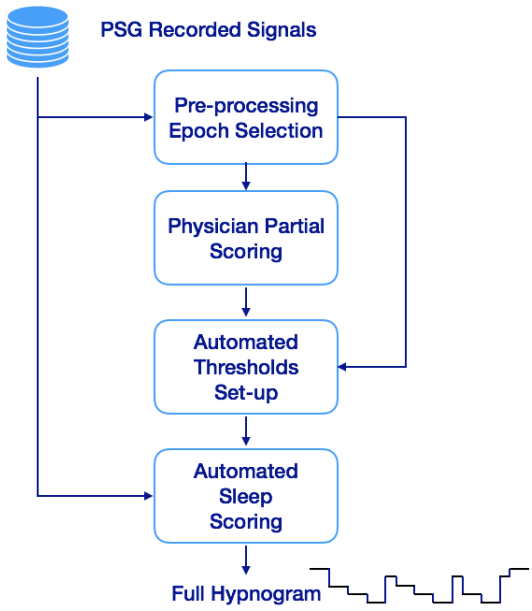


FIGURE 11. Workflow of hybrid expert system.

The proposed system was named as a hybrid expert system because the differential evolution algorithm is mixed in the

symbolic fusion model for computing the thresholds, meanwhile, the symbolic fusion is integrated inside the differential evolution algorithm for evaluating the impact of these thresholds. The proposed expert system is implemented and evaluated using Matlab R2016a (The MathWorks, Inc.).

Details of the hybrid expert system for personalized sleep staging by applying the workflow shown in Fig. 11 is presented below.

- **Pre-processing and Personal Epoch Selection:** pre-processing is designed to eliminate noise and artifacts. According to AASM manual, band-pass filters with a cut-off frequency of 0.3–35 Hz, 0.3–35 Hz and 10–100 Hz are suggested to perform pre-processing filtering of curves for EEG, EOG, and EMG respectively [3]. As recommended by AASM, a Butterworth band-pass filter between 0.3 Hz and 35 Hz is designed for EEG and EOG; a Butterworth band-pass filter between 10 Hz and 100 Hz is designed for EMG. Meanwhile, a band-stop filter with a cut-off frequency of 50 Hz is adopted for EMG to eliminate power-line artifacts. By applying filters proposed in pre-processing, it can effectively eliminate some noise and artifacts like movement artifacts and power-line artifacts.
  Personal Epoch Selection is performed for selecting epochs for the partial physician scoring. The number of epochs that need to be selected is investigated in Section IV-B2.
- **Physician Partial Scoring:** epochs that selected in the Personal Epoch Selection will be scored by the physician to generate the partial Hypnogram.
- **Automated Thresholds Setting-Up:** we have designed five different classifiers, one for each sleep stage (W, N1, N2, N3, and R). For each classifier, the optimal thresholds combination is generated using the differential evolution algorithm. Details are shown in Section IV-B.
- **Automated Sleep Staging:** by using the optimal personalized thresholds combination for each sleep stage classifier, we are able to apply the classification on the remaining PSG recording. Hence, the classification results for each stage can be obtained, which followed by a smoothing function. The smoothing function is proposed to consider the temporal effects of sleep staging process, and to detect and correct false sleep transitions. In smoothing, temporal contextual information and sleep transitions are considered.
  Temporal Contextual Information Smoothing: in smoothing, we implement generally accepted smoothing rules: the ''3-minutes rule'' [38]. If a sequence of six epochs has only one epoch (isolated sleep stage) scored differently from the others (major sleep stage), such as, if the major sleep stage is R/W, then the isolated sleep stage is changed into the major sleep stage R/W.
  Sleep Transitions Detection and Correction: it is designed to detect and correct the impossible transitions, meanwhile, provide the warning on the irregular transitions. As for the sleep transitions, we verified with physicians about the likelihood of occurrence of the transitions. Impossible Transitions are the transitions will never happen, like W to N3, N1 to N3 and R to N3; irregular Transitions are the transitions that rarely happen, like N3 to N1.

VII. EVALUATIONS OF THE PERSONALIZED HYBRID EXPERT SYSTEM

In this section, the dataset for evaluating the personalized hybrid expert system is introduced. Meanwhile, results of the personalized hybrid expert system are presented.

A. DATASET DESCRIPTION

In this paper, overnight PSG signals were recorded in La PitiÉ-Salpêtrière Hospital (AP-HP) in 2016. PSG recordings were segmented into 30s epoch and manually scored into five different sleep stages (W, N1, N2, N3, and R) by physicians using AASM manual. All the PSG signals were recorded using the Grael HD-PSG device including three EEG channels (Fp1–A2, C3–A2 and O1–A2 according to the international 10âĂ"20 standard system), two EOG horizontal channels, and a chin EMG channel. The sampling rates for EEG, EOG, and EMG were 256 Hz. Before going through the PSG recording procedure, all the subjects gave their informed consent, approved by the AP-HP, to participate in this research. Both healthy subjects and subjects who were suspected to suffer from sleep disorders follows instructions to refrain from alcohol and caffeine ingestion and avoid engaging in prolonged and/or strenuous exercise in the daytime before study nights, the subjects underwent one night of polysomnographic recording in a quiet, darkened room. Total 16 subjects (9 males and 7 females) ranging from 22 to 82 years old (mean = 45.6, std = 18.1) are included and the AHI (average number of apneas and hypopneas per hour of sleep) ranges from 0 to 40.2 (mean = 22.0, std = 16.1).

B. PERFORMANCE OF THE PERSONALIZED HYBRID EXPERT SYSTEM

To provide a more robust estimate of the sleep staging performance as compared to the simple agreement percentage and give an overall evaluation of all the stages, Cohen's Kappa coefficient κ is estimated, as shown in (7). p_o is the relative observed agreement among proposed system and experts analysis, and p_e is the hypothetical probability of chance agreement. If results from the proposed system are in complete agreement with expert analysis then $\kappa = 1$. If there is lower agreement among personalized hybrid expert system classification results and experts analysis other than what would be expected by chance, then $\kappa \leq 0$.

$$\kappa = \frac{p_o - p_e}{1 - p_e}; \tag{7}$$

As suggested by Landis and Koch [39], Kappa values of 0.21–0.4 indicate fair agreement, 0.41–0.6 moderate agreement, 0.61–0.8 substantial agreement, and 0.81–0.99 almost perfect agreement.

TABLE 5. Performance of low and high AHI group.

	AHI (Mean ± Std)	Accuracy (Mean ± Std)	Kappa (Mean ± Std)
All	22.02 ± 16.10	80.33 ± 6.77	0.71 ± 0.10
Low AHI	2.00 ± 1.99	82.55 ± 5.29	0.75 ± 0.07
High AHI	31.60 ± 7.71	73.70 ± 6.94	0.60 ± 0.08

To figure out a suitable epoch number that needs to be scored by the physician, different percentage (5%, 10%, 15%, and 20%) of whole recordings to set-up the thresholds are evaluated in this paper. As experimental results presented in Table 4, by using the 5% as the training set, the F-Measures of stages like stage W, N2 and N3 are higher than 70%, which do not show a significant reduction in comparison with using 10%, 15%, and 20% as the training set. Meanwhile, it consumes the minimum time for the physician to score. In this paper, in the consideration of consuming less time for the physician to score, for each subject, 5% of the whole epochs are randomly selected as the training set to provide personalized thresholds for scoring the remaining epochs. The average accuracy of these sixteen subjects is 80.33% with the standard deviation of 6.77%, and it had an average kappa of about 0.71 with the standard deviation of 0.10, as shown in Table 5. Moreover, by selecting 5% as the training set, the overall identification of the five sleep stages of all the subjects can still achieve over 80%, as the confusion matrix presented in Table 6. The recalls for all the stages, except stage N1, are higher than 80%. The precisions of stage W, N3, and R reach over 80%. The Cohen's Kappa coefficient also shows a substantial agreement (0.7224). Therefore, in this paper, we suggest 5% can be considered as the optimal value for the training set which consumes less time for the physician to score the training set. Meanwhile, it still can reach a comparable and favorable performance (overall accuracy is about 80%) in comparison with the inter-scorer variability among different physicians (overall accuracy is about 80.6%-82%) [4]. While, if a short-time PSG recording is required to be analyzed (e.g. if we only need to analyze 1 hour PSG recordings), then the size of training set should be increased to ensure that there are enough epochs for the training step instead of only selecting 5%. However, if the physician dedicates to obtaining a higher accuracy, more epochs can be scored and used in the training process, and the result would be also improved. The balance between the consuming timing in scoring the partial epochs and accuracy of the classification can be adjusted by different physicians according to their exact requirements.

In this paper, the dataset contains both healthy subjects and subjects who suspected to suffer from sleep disorders, which spans a wide range in apnea-hypopnea index (AHI). To investigate whether the difference in the performance of these subjects is related to the AHI or not, the results are also analyzed and presented in two groups, namely low AHI group (AHI $<$ 5) and high AHI group (AHI $\geq$ 5), as shown in Table 5. The mean accuracy is decreased to 73.70% for High AHI group, while the mean kappa can still achieve 0.6. It indicates that the proposed system can still achieve a substantial agreement with the manual analysis from the physician for the High AHI group. To fully understand why the accuracy and the kappa coefficient of High AHI group are lower than Low AHI group, we go through the sleep architectures of high AHI subjects. A high percentage of the stage N1 is observed in the High AHI group, which is in accordance with the observation in [40] and [41]. In the clinical observation study, the sleep architecture changes in sleep apnea patients in comparison with normal subjects. An increment of stage N1 is observed in the sleep apnea patients. To figure out whether the overall decrement in the performance of the High AHI group is caused by the increment of stage N1, the correlation coefficient between the percentage of stage N1 and the accuracy for all the subjects is calculated. A negative coefficient -0.9154 is obtained. The correlation coefficient presents, as the percentage of stage N1 increases, the accuracy of the performance decreases. This is mainly due to the poor classification performance of stage N1 (overall accuracy for stage N1 is less than 20%), as shown in Table 6. However, the detection of N1 is always the most problematic aspect of the sleep stages in both clinical sleep staging and automatic sleep staging system. In the proposed system, the poor classification performance of stage N1 is mainly because of some of the sleep events which can characterize the stage N1 like vertex sharp waves are not yet extracted and fused in the proposed system. However, these sleep events would be involved in the proposed system in the future. Thus, for enhancing the classification performance of stage N1, more sleep events and rules should be taken into consideration in our further work, which would also potentially improve the performance of High AHI group.

C. PERFORMANCE OF VARIOUS CLASSIFICATION MODELS

To compare the proposed personalized hybrid expert system with various machine learning classification methods, several typical models like, DT, random forest (RF), discriminant analysis (DA), SVM, k-Nearest Neighbors (kNN) are implemented on the same database by using the nine digital parameters we extracted for the personalized hybrid system. In order to assess the performance of these classifiers, subject-dependent training process and *k*-fold cross-validation are used. The data of each subject is randomly partitioned into *k* subsets. Among these *k* subsets, *k*-1 subsets are used as training set, the remaining one subset is retained as the validation dataset for testing the model. The cross-validation process is repeated *k* times, the overall performance is averaged on these *k* times results. In this paper, we adopt *k* folds cross-validation for evaluating the classification models instead of using 5% of the data as the training set is mainly due to the following reasons: 1. Using only 5% of the data as the training set may result in over-fit models and may bias the classification results. However, *k* folds cross-validation offers a relatively unbiased

TABLE 6. Confusion matrix of the dataset using the personalized hybrid expert system.

		Personalized Hybrid Expert System							
		W	**N1**	**N2**	**N3**	**R**	**Total**	**Recall(%)**	**F-Measure(%)**
Expert Analysis	**W**	**2593**	58	410	39	92	3192	81.23	84.01
	N1	138	**161**	806	13	204	1322	12.18	20.29
	N2	151	34	**6541**	344	258	7328	89.26	81.74
	N3	19	0	493	**2595**	2	3109	83.47	84.65
	R	80	12	426	31	**2634**	3183	82.75	82.66
	Total	2981	265	8676	3022	3190	18134		
	Precision(%)	86.98	60.76	75.39	85.87	82.57			
	Accuracy(%)	80.09							
	Kappa	0.7224							

result by randomly split the data into k folds and repeat the cross-validation process k times to provide an averaged performance. 2. The functionalities of the training process are totally different in the proposed method and machine learning methods. In the personalized hybrid expert system, the training process is used to find the optimal personalized thresholds and then pass these thresholds to the model and realize the sleep staging of the remaining dataset. While in the machine learning models, the training set is used to train the model and to learn the pattern between features and corresponding sleep stages classes. The accuracies of various classification models with 2, 5 and 10 folds cross-validation are presented in Table 7. All the implementation of these classifiers is carried out using Statistics and Machine Learning Toolbox of Matlab. For the decision tree model, the standard CART algorithm [42] is used to generate the decision trees. Meanwhile, the Gini's diversity index is adopted as the splitting criterion and the maximal number of decision splits is set to 20. Random forest, as proposed by Breiman [43], is an ensemble of decision trees, where each tree is trained by a different subset of the training dataset. In this paper, the number of trees in the random forest is set to 20. For the discriminant analysis model, a popular Linear Discriminant Analysis (LDA) [44], also known as Fisher's Linear Discriminant is adopted in this paper. It searches for a linear combination of features to distinguish the class from others. For a multi-class classification with SVM [45], one-vs-all approach combined with a linear kernel function is applied. It constructs 5 SVM models. Each classifier is trained to separate one class from the remaining 4 classes. For the kNN [46], the Euclidean distance metric is used with k neighbors equal to 1 to 6. In this study, the highest accuracy value for kNN was obtained when $k = 5$, as presented in Table 7. More details of these classifiers can consult [47].

With the number of folds increase, the accuracy of each classification model also slightly increases. Compared with DT and LDA classification models with 10 folds cross-validation, the proposed hybrid expert system achieves favorable results while using only 5% as the training set to obtain the personalized thresholds. While RF, SVM, and kNN outperform the proposed system in terms of the overall accuracy, which mainly due to the following reasons. 1) RF, SVM, kNN, and the proposed system adopts different models/frameworks. For the proposed system, it dedicates to emulating the decision-making process of the expert. While for the machine learning methods, they learn the patterns between features and corresponding stages classes from the training set. Thus, the resulting predictive model is elaborated and insignificant patterns may be selected independently of medical knowledge. 2) The different training process of the proposed method and machine learning methods. In this paper, we adopt k folds cross-validation for evaluating the machine learning models instead of using 5% of the data as the training set. Meanwhile, for machine learning methods, the training set is used to train and generate a predictive model. While in our method, the training set is used to figure out the personalized thresholds for realizing the personalized sleep staging process. In the training process, the model itself will not be modified in our proposed method. However, the proposed system applies the knowledge-based model and mimics the clinical sleep staging process which would be easily understood and validated by physicians. Meanwhile, the proposed system can achieve a favorable result while using only 5% as the training set.

TABLE 7. Accuracy of various classification models.

	DT	RF	DA	SVM	kNN
2 folds	78.37	83.50	79.38	82.13	81.82
5 folds	79.45	84.25	79.40	82.54	82.47
10 folds	79.85	84.27	79.44	83.00	82.70

VIII. DISCUSSION

In clinical sleep analysis, the interpretation of PSG is manually performed by the physician in scoring the polysomnographic curves by applying the AASM manual. Meanwhile, due to the individual variability (like polysomnographic recordings, sleep pattern, sleep architecture, etc.) among subjects, the PSG interpretation is normally done by the physician while considering the subject's dependence like subject

profile, the shape of sleep patterns, etc. Moreover, for the patient who diagnosed with sleep disorders, to set-up the follow-up treatment plan or to assess the therapeutic efficacy, continuous overnight PSG tests may be required. Thus, the PSG test and interpretation may need to be performed more than once for the same subject. To be compliant with the medical guideline while considering individual variability and guide for the follow-up treatment for a patient, a personalized sleep staging method is required. In this paper, a hybrid expert system conception for sleep staging is proposed by combining symbolic fusion and differential evolution algorithm. The proposed system is targeted for helping physicians in clinical sleep analysis, it mimics the clinical sleep staging process by translating AASM and medical knowledge into computer logic which can be understood, accepted and validated by physicians according to their knowledge and experience. In the proposed system, only several epochs are needed to be selected and analyzed by a physician. Then thresholds can be set-up based on the selected epochs and partial Hypnogram that scored by the physician using ATSU model which involves differential evolution algorithm. Finally, these thresholds are used in scoring the remaining epochs by symbolic fusion and full Hypnogram can be generated at the end. Based on this conception, personalized sleep staging system is implemented and evaluated.

In previous studies, machine learning-based systems and knowledge-based expert systems have been explored to automate the sleep staging process [5]–[21], [48]. The overall accuracies of the related studies were in the range of 54.6% to 92.23%. The training sets applied in the machine learning-based systems were in the range of 50% to 97% of the whole dataset. Several typical machine learning methods and symbolic fusion-based method which mainly based on the single EEG or multi-channel signals for classifying sleep into different stages are listed and summarized in Table 8 for the comparison. Several public databases (e.g. the Sleep-EDF database, MIT-BIH Polysomnographic database, the Cleveland Children's Sleep and Health Study, the Cleveland Family Study, etc.) were used for evaluating the models. However, for these public databases, old medical guideline (R&K manual) was used as the gold standard for scoring the polysomnographic data. Several studies applied AASM manual as the gold standard for scoring their databases, while these databases were relatively small and do not offer a public access. Moreover, for most of the studies, the subject-independent approach and a relatively larger training set were used to establish the models for sleep staging. In this paper, only 5% data of each subject was used as the training set to generate the personalized thresholds for the further classification. Meanwhile, the proposed hybrid expert system was evaluated on a clinical database which applied the new medical guideline as the gold standard for scoring the polysomnographic data.

Among the state of the art, the approaches that aimed at representing symbolic qualitative rules or using a hybrid system or focusing on personalized measures or applying the symbolic fusion-based expert framework were discussed. In [12], adaptive neuro-fuzzy inference systems (ANFIS), which implemented fuzzy inference systems in the framework of adaptive networks, have been applied for the sleep stages classification. It applied 10-fold cross-validation and the overall accuracy reached 92.23%. The implemented fuzzy inference system can act as a symbolic qualitative approach by using a set of if-then rules. These if-then rules can be used to express the qualitative aspects of human knowledge and reasoning process. While these rules were generated automatically from the learning process of the training data and the number of rules was relatively large (for each sleep stage, on average, 176 ± 28 rules were generated and pruned). In [14], a hybrid sleep staging system was proposed, which mainly involved two parts, namely a random forest classifier and correction rules. It applied a typical machine learning method, random forest, to realize the classification of five sleep stages. Then, a Markov model-based correction rules were applied to the classification results for the consideration of the dynamic characteristic of sleep transitions. In essence, this hybrid expert system combines two machine learning methods without the requirement of much prior knowledge. In [48], a personalized feature scaling method was proposed to normalize the features before passing the features to the machine learning models. It personalized the features as follows. Firstly, the distribution of a feature was calculated. Then, the upper and lower limits of the distribution of a feature were determined. The upper and the lower limits defined an interval containing most of the extracted measured values of a subject. The absolute measured values were then normalized with respect to the individual interval defined by the upper and lower limits. This way, each feature of a subject was standardized into the range of [0, 1] by a personalized range of measured values. In fact, it is a kind of normalization method to scale the feature into the range of [0, 1] by a personalized range of measured values, which totally differs from our proposed system by taking individual variability into consideration. In [21], a knowledge-based decision system for automatic sleep staging based on symbolic fusion was proposed. A new five-abstraction-layers framework of symbolic fusion, as an extension of the framework introduced by Dasarathy, was presented. Meanwhile, a Turing Machine-like decision process to handle sleep stages transitions was proposed. The overall accuracy reached 54.60%. While in our work, we were focusing on proposing a hybrid expert system, which adopted a typical three-level Dasarathy model and considered the individual variability by using differential evolution algorithm to set-up personalized thresholds. Both of these two papers were based on the symbolic fusion model, while one adopted a new five-level architecture and one used a classical three-level architecture. Meanwhile, two papers were focusing on two different perspectives, one explored the sleep transitions decision method, one dedicated to handling the individual variability concerns. However, these two papers can also be integrated together to generate a comprehensive knowledge-based expert system.

TABLE 8. Comparison with existing works.

Authors (Year)	Signals	Classifier	Dataset (Training VS Testing)	Scoring Rules	Sleep Stages	Accuracy
Syed Anas Imtiaz *et al.* (2015) [5]	EEG	Decision Tree	The Sleep-EDF Database (Training: 50%; Testing: 50%)	R&K	W, S1, S2, (S3, S4), R	78.85%
C. Panagiotou *et al.* (2015) [8]	EEG	Support Vector Machine (SVM)	MIT-BIH Polysomnographic Database (10-fold cross validation)	R&K	W, (S1, S2, R), (S3, S4)	83%
M. Prucnal *et al.* (2017) [11]	EEG	Artificial Neural Network (ANN)	The Sleep-EDF Database (Training: 70%; Testing: 15%; Validation: 15%)	R&K	W, S1, S2, (S3, S4), R	81.10%
S. Raiesdana (2018) [12]	EEG	Adaptive Neuro-Fuzzy Inference System (ANFIS)	Private Database: 14462 epochs (10-fold cross validation)	R&K	W, S1, S2, (S3, S4), R	92.23%
T. Lajnef *et al.* (2015) [13]	EEG, EOG, EMG	Hybrid Method (DT + SVM)	Private Database:12202 epochs (10-fold cross validation)	R&K	W, S1, S2, (S3, S4), R	88%
X. Li *et al.* (2018) [14]	EEG	Hybrid Method (Random Forest + Markov Model)	CCSHS* and CFS** Databases (Training: 50%; Testing: 50%)	R&K	W, S1, S2, (S3, S4), R	85.95%
O. Tsinalis *et al.* (2016) [15]	EEG	Convolutional Neural Network (CNN)	The Sleep-EDF Database (20-fold cross validation)	R&K	W, S1, S2, (S3, S4), R	74%
C. Chen *et al.* (2015) [20]	EEG, EOG, EMG	Symbolic Fusion	Private Database: 12 subjects (Manual Thresholds Setting-Up)	AASM	W, N1, N2, N3, R	76%
Proposed System	EEG, EOG, EMG	Hybrid Method (Symbolic Fusion + Differential Evolution)	Private Database: 16 subjects (Training: 5%; Testing: 95%)	AASM	W, N1, N2, N3, R	80.09%

*CCSHS:Cleveland Children's Sleep and Health Study); **CFS: Cleveland Family Study

For most of the machine learning methods, the accuracy can achieve over 80% [9], [10], [12]–[14], [16]–[18], while for the knowledge-based method, the accuracy can only achieve around 54.6% to 76% [19]–[21]. The relatively high performance of the machine learning-based methods may mainly due to the exhaustive learning process from the training set. The training sets for the machine learning methods are relatively large. While the knowledge-based systems are compliant with the sleep knowledge or international sleep guidelines, which can be understood and validated by the physicians. These expert systems may lose some competitive in terms of the performance, which mainly because not all the medical knowledge can be exhaustively formalized. However, it also worth to mention that comparing performances of the different methods is complicated, since the datasets used are different, patient profiles are different, epochs selection may also be different. Simply comparing the accuracy of the studies may not fair. In this paper, a knowledge-based symbolic fusion framework is adopted. Only 5% data of each subject is required to generate the personalized thresholds for the further classification. It also provides a personalized solution for sleep staging for the consideration of the individual variability instead of simply normalizing/scaling the features. In our previous work [20], Manual Thresholds Setting-Up (MTSU) was adopted in the symbolic fusion framework to transform digital parameter into linguistic features and the overall accuracy was 76%. Fig. 12 presents the

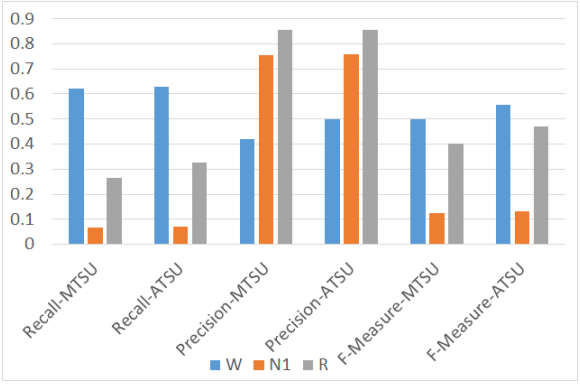


FIGURE 12. Recall, precision and F-Measure using different thresholds setting-up methods of *EMGTh2*.

corresponding recall, precision and F-Measure using different thresholds setting-up methods for setting-up *EMGTh2*. It has been proved that using ATSU proposed in this paper can achieve higher performance in comparison to MTSU in terms of recall, precision, and F-Measure. In this paper, an ATSU model is proposed and it is computationally efficient. It consumes approximately 0.2 sec for each iteration on a server with 2xXeon E5-2640 CPU, 12 cores and 128 GB RAM. Thus, a maximum of 20 sec is required to find the optimal personalized thresholds. Moreover, compared with previous works, the proposed hybrid expert system reached an acceptable and favorable result while using only 5% of the dataset for the training process.

In comparison to existing works, this paper presents a novel automated personalized sleep staging method. To the best of our knowledge, it is the first study of automatic sleep staging system which considers the individual variability. The contributions of this paper can be concluded as follows. Firstly, the proposed system adopts a three-level symbolic fusion framework, it is compliant with the sleep medical guidelines from the sleep events extraction to the decision-making process. By applying this knowledge-based framework, the proposed system can be understood, accepted and validated by physicians according to their knowledge and experience. Secondly, to take the individual variability into consideration, the differential evolution algorithm is explored and integrated with the symbolic fusion framework. Instead of reducing the individual variability by simply normalizing the raw signal or the features that extracted from the raw signal, the proposed system offers a personalized method that can overcome the concerns of the individual variability. Thirdly, only a few epochs need to be scored by the physicians for setting-up the personalized thresholds. In comparison to most of machine learning methods which need a relatively large training set (i.e. the training set normally among 50% to 80% of the whole data), only a few epochs need to be scored for setting-up the personalized thresholds. Finally, the proposed system is evaluated a clinical dataset which involves both healthy subjects and patients with sleep disorders. The overall accuracy on the clinical dataset is over 80% and the kappa coefficient is 0.7224. It can reach a comparable and favorable result in comparison with the inter-scorer variability among different physicians (overall accuracy about 80.6%-82%) [4].

However, there are some limitations of the proposed system: 1) Not all the sleep events and rules described in the AASM have been involved. 2) The proposed system was validated on a relatively small dataset. 3) Several epochs need to be selected and scored by the physician. Thus, our study could still be improved by taking the following points into consideration: 1) Proposed system can still be enhanced by involving more sleep events and rules described by AASM. E.g. for the classification of stage N1, the recall and precision are relatively lower than others stages. This would be improved by involving and fusing more sleep events, like vertex sharp waves in the proposed system. Meanwhile, the proposed system can still be strengthened by improving the identification accuracy of sleep events. E.g. as the confusion matrix presented in Table 6, the precision of stage N2 is less than 80%. To improve the precision of stage N2, the identification accuracy of sleep events (i.e. sleep spindle or K complex) that used to characterize the stage N2 can be further enhanced. 2) More subjects should be involved and sleep efficiency difference between healthy subjects and patients with sleep disorders can be also researched. Meanwhile, by involving more subjects, the robustness and reliability of the proposed system can be further validated. 3) Only a few epochs need to be scored by the physician for setting-up the personalized thresholds. At present, these epochs are selected randomly to verify the performance of the proposed system, nevertheless, from the clinical perspective, score a single epoch out of the context may be a little bit challenge. In the further work, successive epoch selection would be taken into consideration. A preliminary step supporting the selection of epochs to be scored by the expert needs to be investigated.

IX. CONCLUSION

In this paper, a hybrid expert system conception for sleep staging was proposed by combining symbolic fusion and differential evolution algorithm. Symbolic fusion was designed for extracting sleep events and formalizing sleep rules from medical guideline and knowledge; differential evolution algorithm was dedicated to realizing automatic optimization of the thresholds combination that was used in the symbolic fusion model to transform digital parameters into symbolic features. Based on this conception, personalized sleep staging system was implemented and evaluated while taking individual variability into consideration. The overall accuracy and kappa coefficient on the identification of five sleep staging using the proposed hybrid expert system to overnight PSG on 16 subjects have achieved 80.09% and 0.7724, respectively. In the future, the proposed system can still be enhanced by involving more sleep events and rules that described by AASM. Meanwhile, other stochastic search algorithms, like switching particle swarm optimization algorithm can also be explored and investigated in the hybrid expert system. Furthermore, the proposed system can be extended to realize complete personalized sleep disorders analysis and be integrated into an embedded system to realize remote sleep monitoring.

ACKNOWLEDGMENT

The authors would like to thank Prof. Patrick Garda, Prof. Jean-Gabriel Ganascia, Prof. Amara Amara and Dr. Xun Zhang for supporting the research. They are also grateful to Dr. Amina Kotti for her assistance with sleep medical knowledge.

REFERENCES

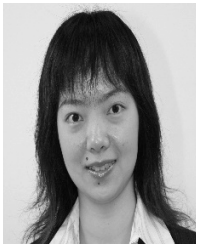
[1] J. L. Hossain and C. M. Shapiro, "The prevalence, cost implications, and management of sleep disorders: An overview," *Sleep Breathing*, vol. 6, no. 2, pp. 85–102, 2002.

[2] T. A. Hargens, A. S. Kaleth, E. S. Edwards, and K. L. Butner, "Association between sleep disorders, obesity, and exercise: A review," *Nature Sci. Sleep*, vol. 5, pp. 27–35, Mar. 2013.
[3] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan, "The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications," American Acad. Sleep Med., Westchester, IL, USA, Tech. Rep., 2007.
[4] H. Danker-Hopfe *et al.*, "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *J. Sleep Res.*, vol. 18, no. 1, pp. 74–84, 2009.
[5] S. A. Imtiaz and E. Rodriguez-Villegas, "Automatic sleep staging using state machine-controlled decision trees," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 378–381.
[6] A. R. Hassan and M. I. H. Bhuiyan, "A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features," *J. Neurosci. Methods*, vol. 271, pp. 107–118, Sep. 2016.
[7] K. D. Tzimourta *et al.*, "Eeg-based automatic sleep stage classification," *Biomed. J. Sci. Tech. Res.*, vol. 7, no. 4, pp. 1–6, 2018.
[8] C. Panagiotou, I. Samaras, J. Gialelis, P. Chondros, and D. Karadimas, "A comparative study between SVM and fuzzy inference system for the automatic prediction of sleep stages and the assessment of sleep quality," in *Proc. 9th Int. Conf. Pervasive Comput. Technol. Healthcare*. Gent, Belgium: ICST, 2015, pp. 293–296.
[9] S. Seifpour, H. Niknazar, M. Mikaeili, and A. M. Nasrabadi, "A new automatic sleep staging system based on statistical behavior of local extrema using single channel EEG signal," *Expert Syst. Appl.*, vol. 104, pp. 277–293, Aug. 2018.
[10] B. A. Savareh, A. Bashiri, A. Behmanesh, G. H. Meftahi, and B. Hatef, "Performance comparison of machine learning techniques in sleep scoring based on wavelet features and neighboring component analysis," *PeerJ*, vol. 6, p. e5247, Jul. 2018.
[11] M. Prucnal and A. G. Polak, "Effect of feature extraction on automatic sleep stage classification by artificial neural network," *Metrol. Meas. Syst.*, vol. 24, no. 2, pp. 229–240, 2017.
[12] S. Raiesdana, "Automated sleep staging of OSAs based on ICA preprocessing and consolidation of temporal correlations," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 1, pp. 161–176, 2018.
[13] T. Lajnef *et al.*, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *J. Neurosci. Methods*, vol. 250, pp. 94–105, Jul. 2015.
[14] X. Li, L. Cui, S. Tao, J. Chen, X. Zhang, and G.-Q. Zhang, "Hyclasss: A hybrid classifier for automatic sleep stage scoring," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 2, pp. 375–385, Mar. 2018.
[15] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou. (2016). "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1610.01683
[16] J. Zhang and Y. Wu, "Complex-valued unsupervised convolutional neural networks for sleep stage classification," *Comput. Methods Programs Biomed.*, vol. 164, pp. 181–191, Oct. 2018.
[17] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, and J.-F. Payen, "A convolutional neural network for sleep stage scoring from raw single-channel EEG," *Biomed. Signal Process. Control*, vol. 42, pp. 107–114, Apr. 2018.
[18] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.
[19] A. Ugon, "Fusion symbolique et données polysomnographiques," Ph.D. dissertation, Dept. Comput. Sci., Univ. Pierre Marie Curie, Paris, France, 2013.
[20] C. Chen *et al.*, "Symbolic fusion: A novel decision support algorithm for sleep staging application," in *Proc. 5th EAI Int. Conf. Wireless Mobile Commun. Healthcare*. London, U.K.: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2015, pp. 19–22.
[21] A. Ugon *et al.*, "Knowledge-based decision system for automatic sleep staging using symbolic fusion in a turing machine-like decision process formalizing the sleep medicine guidelines," *Expert Syst. Appl.*, vol. 114, pp. 414–427, Dec. 2018.
[22] I. Bloch and H. Maitre, "Data fusion in 2D and 3D image processing: An overview," in *Proc. 10th Brazilian Symp. Comput. Graph. Image Process.*, Oct. 1997, pp. 127–134.
[23] P. Lambert and T. Carron, "Symbolic fusion of luminance-hue-chroma features for region segmentation," *Pattern Recognit.*, vol. 32, no. 11, pp. 1857–1872, 1999.
[24] A. Ugon, J.-G. Ganascia, C. Philippe, H. Amiel, and P. Lévy, "How to use symbolic fusion to support the sleep apnea syndrome diagnosis," in *Proc. Conf. Artif. Intell. Med. Eur.* Bled, Slovenia: Springer, 2011, pp. 45–54.
[25] R. S. Rosenberg and S. Van Hout, "The American academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *J. Clin. Sleep Med.*, vol. 9, no. 1, pp. 81–87, 2013.
[26] S.-F. Liang, Y.-H. Chen, C.-E. Kuo, J.-Y. Chen, and S.-C. Hsu, "A fuzzy inference system for sleep staging," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, Jun. 2011, pp. 2104–2107.
[27] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, and Y.-S. Cheng, "A rule-based automatic sleep staging method," *J. Neurosci. Methods*, vol. 205, no. 1, pp. 169–176, 2012.
[28] C. Chen *et al.*, "Personalized sleep staging system using evolutionary algorithm and symbolic fusion," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 2266–2269.
[29] C. Chen *et al.*, "Cross entropy-based automatic thresholds setting-up method for sleep staging system," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2016, pp. 312–315.
[30] P. K. Wong, F. Yu, A. Shahangian, G. Cheng, R. Sun, and C.-M. Ho, "Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 13, pp. 5105–5110, 2008.
[31] C.-M. Ho, "Keynote: Personalized medicine enabled by FSC.X technology," in *Proc. IEEE Faible Tension Faible Consommation (FTFC)*, May 2014, pp. 1–2.
[32] N. Zeng, H. Qiu, Z. Wang, W. Liu, H. Zhang, and Y. Li, "A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease," *Neurocomputing*, vol. 320, pp. 195–202, Dec. 2018.
[33] N. Zeng, H. Zhang, W. Liu, J. Liang, and F. E. Alsaadi, "A switching delayed PSO optimized extreme learning machine for short-term load forecasting," *Neurocomputing*, vol. 240, pp. 175–182, May 2017.
[34] N. Zeng *et al.*, "Image-based quantitative analysis of gold immunochromatographic strip via cellular neural network approach," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1129–1136, May 2014.
[35] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–359, 1997.
[36] C. M. Selvi and K. Gnanambal, "Power system voltage stability analysis using modified differential evolution," in *Proc. Int. Conf. Comput., Commun. Elect. Technol.*, Mar. 2011, pp. 382–387.
[37] R. A. Sarker, S. M. Elsayed, and T. Ray, "Differential evolution with dynamic parameters selection for optimization problems," *IEEE Trans. Evol. Comput.*, vol. 18, no. 5, pp. 689–707, Oct. 2014.
[38] Q. K. Le, Q. D. K. Truong, and V. T. Vo, "A tool for analysis and classification of sleep stages," in *Proc. Int. Conf. Adv. Technol. Commun.*, Aug. 2011, pp. 307–310.
[39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
[40] D. Shrivastava, S. Jung, M. Saadat, R. Sirohi, and K. Crewson, "How to interpret the results of a sleep study," *J. Community Hospital Internal Med. Perspect.*, vol. 4, no. 5, p. 24983, 2014.
[41] K. Shahveisi, A. Jalali, M. R. Moloudi, S. Moradi, A. Maroufi, and H. Khazaie, "Sleep architecture in patients with primary snoring and obstructive sleep apnea," *Basic Clin. Neurosci.*, vol. 9, no. 2, pp. 147–156, 2018.
[42] L. Breiman, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, 2017.
[43] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[44] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, vol. 544. Hoboken, NJ, USA: Wiley, 2004.
[45] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.
[46] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 4, pp. 580–585, Jul./Aug. 1985.
[47] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.
[48] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated EEG sleep staging in the term-age baby using a generative modelling approach," *J. Neural Eng.*, vol. 15, no. 3, p. 036004, 2018.

CHEN CHEN received the M.S. degree in embedded system from the Institut Supérieur d'Electronique de Paris, in 2013, and the Ph.D. degree in computer science from Université Pierre et Marie Curie, in 2016. She is currently a Post-Doctoral Researcher with the Centre of Intelligent Medical Electronics, School of Information Science and Technology, Fudan University. Her research interests lie in biomedical engineering, focusing on biomedical signal processing, wearable sensor systems, sleep analysis, and personalized health monitoring.

ADRIEN UGON received the M.S. degree in computer science from Université d'Évry Val d'Essonne, in 2006, and the Ph.D. degree from Université Pierre et Marie Curie, in 2013. He is currently an Assistant Professor with the École Supérieure d'Ingénieurs en Électrotechnique et Électronique Paris. His research interests lie in artificial intelligence applied to healthcare and decision support systems based on data abstraction by knowledge integration through expert systems. His favorite application domain is the sleep medicine and the support of the sleep apnea syndrome.

CHENGLU SUN received the B.S. degree in mechatronic engineering from the Shanghai Normal University Tianhua College, Shanghai, China, in 2013, and the M.S. degree in mechanical manufacture and automation from the University of Shanghai for Science and Technology, Shanghai, in 2016. He is currently pursuing the Ph.D. degree with the Center for Intelligent Medical Electronics, Department of Electronic Engineering, School of Information Science and Technology, Fudan University, Shanghai. His research interests include signal processing, machine learning, and computer vision with an emphasis on unobtrusive physiological signal monitoring.



WEI CHEN received the B.S. and M.S. degrees from Xian Jiaotong University in 1999 and 2002, respectively, and the Ph.D. degree from The University of Melbourne, in 2007. From 2008 to 2015, she was an Assistant Professor with the Department of Industrial Design, Technical University of Eindhoven. She is currently a Full Professor with the Department of Electronic Engineering, Fudan University. Her research interests include wearable sensor system, biomedical signal processing, and health monitoring. She is a Vice Chair of the IEEE Sensors and Systems Council. She is an Associate Editor of the IEEE JOURNAL OF BIOMEDICAL HEALTH INFORMATICS. She is a Regional Representative of the IEEE EMBS Technical Committee on Wearable Biomedical Sensors and Systems.

CAROLE PHILIPPE received the M.D. degree from Université Pierre et Marie Curie, in 1990, and the Ph.D. degree from Université Paris-Est Créteil en Sciences de la Vie et de la Santé, in 2011. She is currently a Physician (Praticien Hospitalier) with the Unité des Pathologies du Sommeil, Groupe Hospitalier Pitié Salpêtrière, Paris, France. She takes care of patients suffering from sleep apnea syndrome. Her principal research interests sleep recording techniques and analysis.

ANDREA PINNA received the Ph.D. degree in electronics systems from University Pierre and Marie Curie, Paris, France, in 2003. He is currently an Associate Professor with the French Laboratory of Computer Science, Sorbonne UniversitÉ, Paris. From 2006 to 2011, he worked in private industry in semiconductor and information technology. His research activities are based on e-health embedded system for smart medical device design and on embedded system co-design.