IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received September 3, 2019, accepted September 16, 2019, date of publication September 20, 2019,
date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942641

Epileptic EEG Signals Recognition Using a Deep View-Reduction TSK Fuzzy System With High Interpretability

YUANPENG ZHANG[2,3], (Member, IEEE), XIANGZHE LI[1], JUNQING ZHU[4], CHUNYING WU[4], AND QINFENG WU[1]

[1]Rehabilitation Medicine Center, Suzhou Science and Technology Town Hospital, Suzhou 215153, China
[2]Department of Medical Informatics, Medical School, Nantong University, Nantong 226001, China
[3]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong
[4]Department of Radiology, Case Center for Imaging Research, Case Western Reserve University, Cleveland, OH 44106, USA

Corresponding author: Qinfeng Wu (wuqinfeng0911@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 81701793, and in part by the Hong Kong Scholars Program under Grant XJ2019056.

ABSTRACT Takagi-Sugeno-Kang (TSK) fuzzy systems are well known for their good balance between approximation accuracy and interpretability. In this paper, we propose a deep view-reduction TSK fuzzy system termed as DVR-TSK-FS in which two powerful mechanisms associating with a deep structure are developed: 1) during the multi-view learning in each component, a sample-distribution-dependent parameter is defined to control the learning of the weight of each view. The parameter is not fixed by users, it is set according to the feature space in advance such that the learnt weight of each view indeed reflects the amount of pattern information involved in each view; 2) during the iteration of DRV-TSK-FS in each component, weak views are automatically reduced by comparing the learnt weight with a fixed threshold which is also automatically set according to the number of objects and the dimension of the feature space. 3) All components are linked in a stacked way based on the stacked generalization principle such that the outputs of all previous components are augmented into the current one which can help open the manifold structure of the original feature space. DRV-TSK-FS is testified on a multi-view EEG dataset for epileptic EEG signals recognition.

INDEX TERMS Multi-view learning, stacked generalization principle, view reduction, TSK fuzzy systems.

I. INTRODUCTION

Epilepsy is a finite episode of brain dysfunction caused by abnormal discharge of cerebral neurons. With regards to the clinical diagnosis of Epilepsy, electroencephalogram (EEG) signals are often employed to decide its presence and type [1]. With the development of clinical decision systems (CDS), how to design an effective CDS and hence automatically detect seizures from EEG signals becomes very significantly in clinical practice. As a result, many machine learning-based approaches including SVM [2], fuzzy systems [3]–[7], KNN [14]–[16], decision trees [17], [18] have been developed and successfully applied in epileptic EEG signals recognition [3]–[7]. As stated in [4], two essential steps are required when a CDS works for EEG signals recognition. The first one is to extract valuable features from EEG signals by appropriate feature extraction approaches. The second one is to design and train a classifier for EEG signals recognition using the extracted features. From the above two steps, it is obvious that there exist at least two main factors which may affect the recognition performance. One is that sufficient and effective features extracted from EEG signals may bring positive affection to the recognition performance. Recently, some studies focus on using different feature extraction approaches to extract EEG features simultaneously, then combining there different kinds of features together to drive a multi-view CDS for EEG signals recognition. In [4], the authors firstly employed different extraction approaches, e.g., WPD, STFT, KPCA [19]–[22] to extract EEG features to construct multi-view EEG data. Then they introduced the

The associate editor coordinating the review of this manuscript and approving it for publication was Yongtao Hao.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

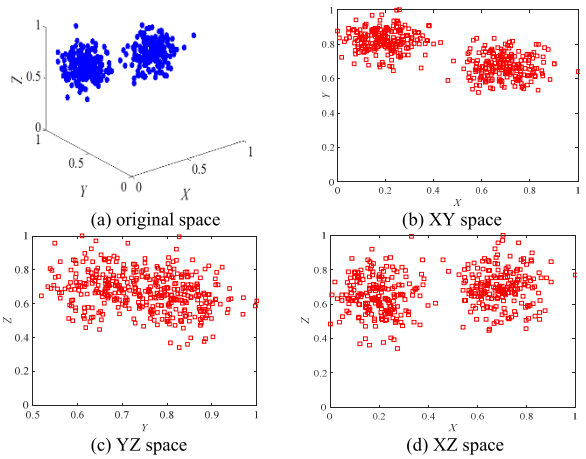


FIGURE 1. An example of weak views.

Shannon entropy and developed a multi-view TSK fuzzy system to recognize abnormal EEG signals based on multi-view collaborative learning framework. Their experimental results indicated that comparing with traditional single view EEG data, multi-view EEG data contain more significant and effective pattern information. When combining with multi-view collaborative learning, promising performance for EEG signals recognition is able to be expected. Besides, in [2], multi-view data are obtained by explicitly considering both the EEG reconstruction and seizure detection errors to unleash the power of multi-channel information. Except multi-view features, the performance and interpretability of the adopted classifier also plays a significant role for EEG signals recognition. Although some classical classifiers (e.g., SVM, KNN, C4.5) can perform well for classification tasks, they are all non-transparent for users. In other words, they are black boxes and work in a black way. Recently, TSK fuzzy systems are widely used for EEG signals recognition because of their promising performance and high interpretability. For example, in [4]–[7], the authors proposed different kinds of TSK fuzzy systems for EEG signals recognition.

Despite multi-view TSK fuzzy systems can generate promising performance and good interpretability for EEG recognition, there still exists several challenges to be addressed. Firstly, it is well known that, some views extracted from EEG signals may contain redundant information and become irrelevant or noneffective (We call these views weak views). If they are included in the recognition procedure, then they may not help discriminate between patterns and accordingly yield invalid recognition results. Fig.1 gives a toy example to illustrate the negative influences from weak views.

Fig.1(a) shows 400 objects distributed in a 3-dimentional space that can be grouped into 2 classes. All objects in Fig.1(a) are projected into three 2-dimentional spaces, i.e., the XY space, the YZ space and the XZ space, respectively. Therefore, we can consider each feature space as a view and analysis the original data in Fig.1(a) from 3 views. Obviously, it is hard to obtain effective pattern information from the feature space in Fig.1(c) to classify objects into 2 groups. Hence, during the procedure of multi-view collaborative learning of the 3 views, the second one in Fig1.(c) may exert negative influences on the final classification result.

Secondly, in most TSK fuzzy systems, a fuzzy grid is often employed to group the input space into different subsets and generate fuzzy rules. However, such a grid can cause the rule-explosion problem so that the interpretability will be inevitably degraded with the increasing number of features [8]. To solve the rule-explosion problem, hierarchical TSK fuzzy systems are often used. Generally speaking, a hierarchical TSK fuzzy system is constructed by several classical TSK fuzzy systems as components in a layer-by-layer way. The original feature space is divided into different parts as the input of each component [8]. In addition, the output of a component in one layer is also taken as the input of a component in the next layer. Although hierarchical TSK fuzzy systems can solve the rule-explosion problem, the output of each component is not endowed with explicit physical meaning such that fuzzy rules in each component become incomprehensible. This problem become severe with the increasing number of layers in hierarchical TSK fuzzy systems.

To handle the aforementioned challenges, in this study, a novel deep view-reduction TSK fuzzy system termed as DVR-TSK-FS is proposed. In DVR-TSK-FS, a sample-distribution-dependent parameter is defined to control the learning of the view weight during multi-view collaborative learning in each component. This parameter is user-free and set according to the feature space in advance such that the learnt weight of each view indeed reflects the amount of effective / valuable pattern information involved in each view. Moreover, a view-reduction principle is set out that weak views are automatically reduced by comparing the learnt weight with a fixed threshold which is also automatically set according to the number of samples and the dimension of feature space. Besides, based on the stacked generalization principle [23], we design a special hierarchical structure in which each basic component (1-TSK-FS) is linked in a stacked way such that the generalization capacity of DVR-TSK-FS is enhanced. Moreover, unlike classic hierarchical fuzzy systems in which the outputs of previous layers have less physical meaning, DVR-TSK-FS can hide the outputs of previous layers in certainly factors such that its high interpretability keeps.

The rest paper is organized as follows. We prepare some basic knowledge about TSK fuzzy systems and multi-view EEG data in Section II. In Section III, DVR-TSK-FS is designed. In Section IV, experimental results on epilepsy EEG data are reported and the paper is concluded in the last section.

II. PRELIMINARY

Since the basic component in DVR-TSK-FS is the classic 1-order TSK fuzzy system (1-TSK-FS) and DVR-TSK-FS is a special hierarchical fuzzy system, here we briefly introduce

the fuzzy rule and training of 1-TSK-FS, and three classic hierarchical structures used in fuzzy systems.

A. 1-TSK-FS

Suppose we have a dataset $\chi = \{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T \in R^d$ and N is the number of samples invovled in χ, then the kth fuzzy rule in the feature space of the 1-order TSK fuzzy system can be expressed as

$$\begin{aligned} &\text{If } x_{i1} \text{ is } A_1^k \wedge x_{i2} \text{ is } A_2^k \wedge \ldots \wedge x_{id} \text{ is } A_d^k, \\ &\text{then } f^k(\mathbf{x}_i) = p_0^k + p_1^k x_{i1} + \ldots + p_d^k x_{id}, \quad k = 1, 2, \ldots, K. \end{aligned} \tag{1}$$

In (1), A_j^k is a fuzzy set subscribed by the input feature x_{ij} for the kth rule, $\wedge$ is a operator for fuzzy conjunction and K is the number of fuzzy rules. Each fuzzy rule is premised on the feature space ($\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T \in R^d$) and maps the fuzzy sets in the feature space to a varying singleton represented by $f^k(\mathbf{x}_i)$. After series of operation steps and defuzzification procedures, the decision result of the 1-order TSK fuzzy system can be formulated as

$$y^o(\mathbf{x}_i) = \sum_{k=1}^{K} \frac{\mu^k(\mathbf{x}_i)}{\sum_{k'=1}^{K} \mu^{k'}(\mathbf{x}_i)} f^k(\mathbf{x}_i) = \sum_{k=1}^{K} \tilde{\mu}(\mathbf{x}_i) f^k(\mathbf{x}_i), \tag{2}$$

where

$$\mu^k(\mathbf{x}_i) = \prod_{j=1}^{d} \mu_{A_j^k}(x_{ij}). \tag{3}$$

The Gaussian function is often employed as the fuzzy membership function such that $\mu_{A_j^k}(x_{ij})$ can be defined as

$$\mu_{A_j^k}(x_{ij}) = \exp\left(\frac{-(x_{ij} - c_j^k)}{2\delta_j^k}\right), \tag{4}$$

where c_j^k and δ_j^k denote the kernel center and kernel width, respectively.

From (2), we see that c_j^k and δ_j^k in the antecedent and $\mathbf{p}^k = [p_0^k, p_1^k, \ldots, p_d^k]^T$ are two kinds of parameters needed to learn in the training procedure of 1-order TSK fuzzy system. Generally speaking, antecedent learning and consequent learning are carried out independently. As for antecedent learning, clustering techniques [24]–[28] are often used. For example, if FCM is employed, c_j^k and δ_j^k can be caclulated as

$$c_j^k = \sum_{i=1}^{N} \mu_{ik} x_{ij} \sum_{i=1}^{N} \mu_{ik}, \tag{5}$$

$$\delta_j^k = h \sum_{i=1}^{N} u_{ik} (x_{ij} - c_j^k)^2 \sum_{i=1}^{N} u_{ik}, \tag{6}$$

where μ_{ik} denotes the fuzzy membership degree $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T$ belonging to cluster k. h is a regularized constant that is often set to 0.5 manually or determined by cross-validation strategies. As for antecedent learning, suppose parameters in antecedent are determined, let

$$\mathbf{x}_e = (1, (\mathbf{x}_i)^T)^T, \tag{7}$$

$$\tilde{\mathbf{x}}_i^k = \tilde{\mu}^k(\mathbf{x}_i)\mathbf{x}_e, \tag{8}$$

$$\mathbf{x}_{gi} = ((\tilde{\mathbf{x}}_i^1)^T, (\tilde{\mathbf{x}}_i^2)^T, \ldots, (\tilde{\mathbf{x}}_i^K)^T)^T, \tag{9}$$

$$\mathbf{p}^k = (p_0^k, p_1^k, \ldots, p_d^k)^T, \tag{10}$$

$$\mathbf{p}_g = ((\mathbf{p}^1)^T, (\mathbf{p}^2)^T, \ldots, (\mathbf{p}^K)^T)^T, \tag{11}$$

then the decision result of the 1-order TSK fuzzy system can be rewritten as

$$y^o(\mathbf{x}_i) = \mathbf{p}_g^T \mathbf{x}_{gi}. \tag{12}$$

From (12), it is obvious that the consequent learning can be considered as solving a linear regression problem. According to different criteria, many solution strategies can be used. In [2], $\mathbf{p}_g$ is solved by the following objective function,

$$J_{1-order-TSK}(\mathbf{p}_g) = \frac{1}{2}(\mathbf{p}_g)^T \mathbf{p}_g + \frac{\eta_{\mathbf{p}_g}}{2} \sum_{i=1}^{N} \left\| (\mathbf{p}_g)^T \mathbf{x}_{gi} - y_i \right\|^2, \tag{13}$$

where $\frac{1}{2}(\mathbf{p}_g)^T \mathbf{p}_g$ is a regularization item than can improve the generalization ability of the 1-order TSK fuzzy system for classification tasks. $\sum_{i=1}^{N} \left\| (\mathbf{p}_g)^T \mathbf{x}_{gi} - y_i \right\|^2$ is the error item and $\eta_{\mathbf{p}_g} > 0$ is used to control the the complexity of the 1-order TSK fuzzy system and the tolerance of errors. By setting $\partial J_{1-order-TSK}(\mathbf{p}_g)/\partial \mathbf{p}_g$ to 0, the optimal $\mathbf{p}_g$ can be analytically obtained as

$$\mathbf{p}_g = \left(\mathbf{I}_{k(d+1)\times k(d+1)} + \sum_{i=1}^{N} \mathbf{x}_{gi} (\mathbf{x}_{gi})^T \right)^{-1} \times \left(\eta_{\mathbf{p}_g} \sum_{i=1}^{N} \mathbf{x}_{gi} y_i \right). \tag{14}$$

B. HIERARCHICAL STRUCTURES

As stated in the first section, hierarchical structures designed for fuzzy systems aim at solving the rule-explosion problem. Generally speaking, a hierarchical fuzzy system is constructed by many low dimensional fuzzy systems termed as basic components that are connected in a layer-by-layer manner [13]. Although different kinds of hierarchical fuzzy systems have been designed for classification or regression tasks, their frameworks, i.e., the hierarchical structures can be divided into three categories, i.e., *incremental, aggregated, and cascaded* [29], [30] as illustrated in Fig.2. With regards to the incremental structure, the original features are broken down into several parts, then each part is taken as the input to the low-dimensional fuzzy systems. Also, except the first layer, the low-dimensional fuzzy systems in other layers receive the outputs of the previous layers and take as their inputs [8], see Fig.2(a) which shows a 2-input incremental hierarchical structure. For the aggregated structure shown in Fig.2(b), the original features are also broken down into

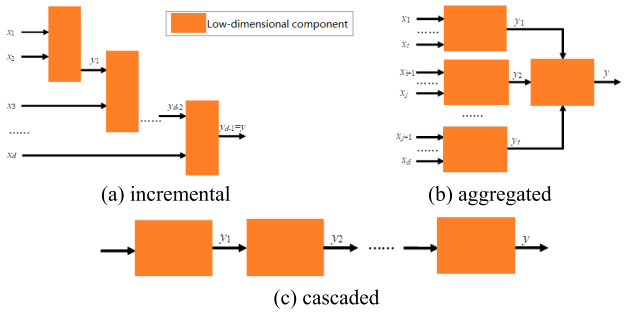


FIGURE 2. Structures of hierarchical fuzzy systems.

several parts, then each part is taken as the input only to low-dimensional fuzzy systems in the first layer. Outputs from the low-dimensional fuzzy systems in the first layer are taken as inputs to the low-dimensional fuzzy system in the next layer. Such a structure can be extended to a larger number of low-dimensional fuzzy systems in the first layer and multiple layers. Fig.2(c) shows the cascaded structure which is another typical hierarchical structure. In cascaded structure-based fuzzy systems, all input features are presented to the first component whose output are then taken as the input to the next component in a layer-by-layer manner.

Although the three types of hierarchical fuzzy systems can better handle the rule-explosion problem mentioned above, we should keep in mind that the severe deterioration in the interpretability is creeping in. Firstly, it is difficult to give physical meaning to intermediate variables (y_i in Fig.2). As a result, we cannot easily interpret each fuzzy rule in which intermediate variables are embedded. Secondly, although the hierarchical fuzzy systems are able to significantly reduce the number of fuzzy rules, the curse of dimensionality still exists. For example, in Fig.2(a), the number of layers increases as the number of input features increases. As a result, the increasing number of layers further deteriorates the interpretability of a hierarchical fuzzy system.

C. MULTI-VIEW EEG DATA

The original EEG data[1] we used in this study are provided by the University of Bonn which consist of five groups, i.e., group A to group E, with each one containing 100 single channel EEG segments of 23.6 duration. The sampling rate is 173.6Hz. Segments in group A and group B are obtained from healthy volunteer subjects and segments in groups C, D and E are acquired from volunteer subjects with epilepsy. Table 1 gives the detailed description about the epileptic EEG data, and Fig.3 illustrates some representative original epileptic EEG signals in five groups.

Since the original EEG data are highly stochastic, nonstationary, nonlinear and contains background noises, it has been demonstrated that directly using original EEG data for recognition may result in unstable and even bad performance [4].

[1]https://www.meb.uni-bonn.de/epileptologie/science/physik/eegdata.html

TABLE 1. Detailed information about the epileptic EEG data.

Subjects	Group	#Size	Description
Healthy	A	100	Signals captured from volunteers with eyes open
	B	100	Signals captured from volunteers with eyes closed
Epileptic	C	100	Signals captured from volunteers during seizure silence intervals.
	D	100	Signals captured from volunteers during seizure silence intervals.
	E	100	Signals captured from volunteers during seizure activity.

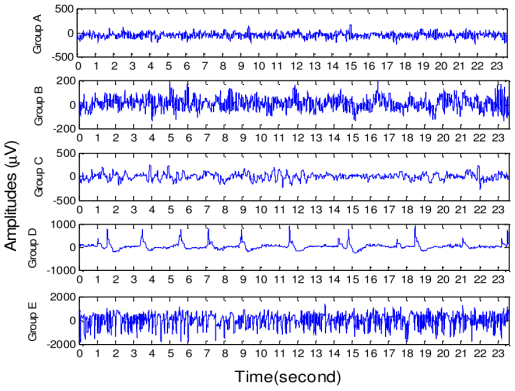


FIGURE 3. Original signals in five groups.

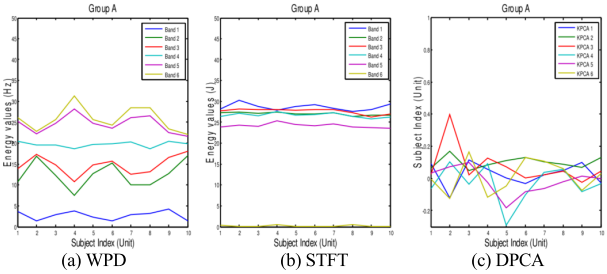


FIGURE 4. Extracted features by using WPD, STFT and DPCA methods, respectively.

Feature extraction methods are widely used for signals processing. Generally speaking, there exists three main types of features involved in signals that can be extracted and used for pattern recognition, i.e.,

1) time-domain features, e.g., principle component features;
2) time-frequency features, e.g., features obtained by wavelet analysis;
3) frequency-domain features, Fourier transform features;

In this study, we aim at learning pattern information in EEG data from multiple views. Thus, we employ different feature extraction methods, i.e., WPD (wavelet packet decomposition) [22], STFT (short time Fourier transform) [21] and KPCA (kernel principal component analysis) [20] to extract different kinds of features to construct a multi-view epileptic EEG dataset for our next experiments. Fig.4 gives the extracted features using the three methods from Group A.

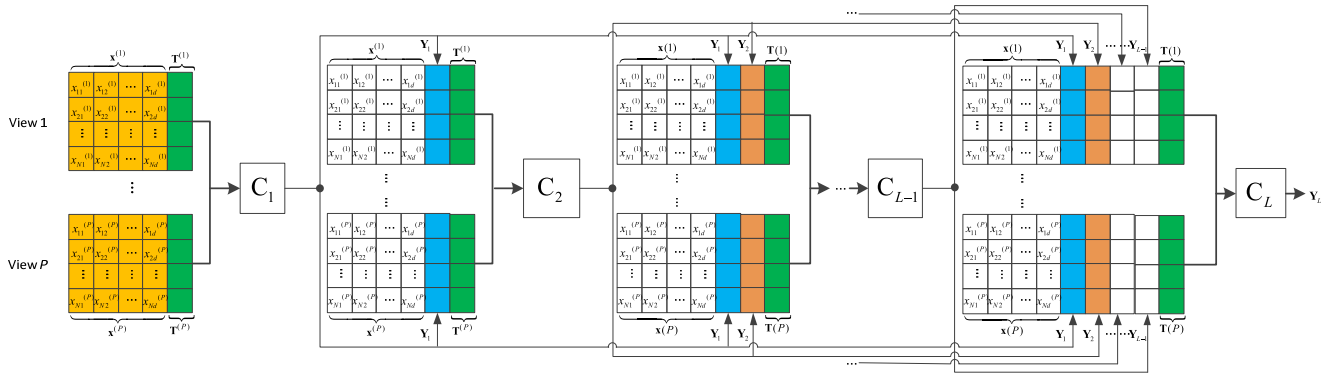


FIGURE 5. Deep structure of DVR-TSK-FS.

III. DEEP VIEW-REDUCTION TSK FUZZY SYSTEM FOR EPILEPTIC EEG RECOGNITION

In this section, a deep view-reduction TSK fuzzy system (DVR-TSK-FS) with high interpretability is proposed for epileptic EEG recognition.

A. NOTATION AND PROBLEM STATEMENT

Suppose we have a multi-view EEG training dataset χ in which features in each view are extracted from original EEG signals by different feature extraction algorithms. The multi-view dataset consists of P views, and samples and corresponding label vectors in each view are denoted as $\mathbf{X}^{(p)} = \{\mathbf{x}_i^{(p)}\}_{i=1}^N$ and $\mathbf{T}^{(p)} = \{t_i^{(p)}\}_{i=1}^N$, respectively, where $\mathbf{x}_i^{(p)} = [x_{i2}^{(p)}, x_{i3}^{(p)}, \ldots, x_{id}^{(p)}]^T$, $1 \le p \le P$. With the multi-view EEG data, we expect to recognize the epileptic subjects. To be specific, a special hierarchical (deep) fuzzy system is designed in which each component denoted as C_l, $1 \le l \le L$ is connected in a stacked way, where L is the number of the components. Each component as the basic unit is a 1-TSK-FS which can achieve multi-view collaborative learning and weak view reduction. Besides, the intermediate output of the proposed fuzzy system should be endowed with physical meanings such that the proposed fuzzy system is highly interpretable.

B. DEEP STRUCTURE

Based on the stacked generalization principle [23], the deep structure of DVR-TSK-FS is shown in Fig.5.

In Fig.5, all components are concatenated in a stacked manner according to the stacked generalization principle. Specifically, the output of the first component (layer), i.e., $\mathbf{Y}_1$ of component C_1 is augmented into the original input feature spaces as the input to the next all components, i.e., $\mathrm{C}_2, \mathrm{C}_3, \ldots, \mathrm{C}_L$. Similarly, the output of the lth component, i.e., $\mathbf{Y}_l$ is augmented into the original input feature spaces as the input to next all components, i.e., $\mathrm{C}_{l+1}, \mathrm{C}_{l+2}, \ldots, \mathrm{C}_L$. Finally, the input feature space of component C_L consists of the original feature space $\mathbf{x}_i^{(p)}$ and the outputs of all previous components. In this way, comparing with the classic TSK fuzzy system only having a single component, the generalization capability of DVR-TSK-FS is improved since that the manifold structure of the original input space is constantly opened by outputs of previous components.

In such a deep structure, the kth fuzzy rule of the lth component C_l in view p can be written as

$$\begin{aligned}&\text{If } x_1^{(p)} \text{ is } A_1^k \wedge x_2^{(p)} \text{ is } A_2^k \wedge \ldots \wedge x_d^{(p)} \text{ is } A_d^k \wedge y_1 \text{ is } \\ &\qquad A_{y_1}^k \wedge \ldots \wedge y_{l-1} \text{ is } A_{y_{l-1}}^k, \\ &\text{then } f_l^k(\mathbf{x}^{(p)}) = p_1^{k,(p)} x_1^{(p)} + \ldots + p_{l-1}^{k,(p)} x_{l-1}^{(p)} + r_l^{k,(p)}, \quad (15)\end{aligned}$$

where $A_{y_i}^k$ is a fuzzy subset for the output $y_i \in \mathbf{Y}_i$ from the ith component in the kth fuzzy rule ($i = 1, \ldots, l-1$), $k = 1, \ldots, K_l$, and K_l is the number of fuzzy rules in component C_l. $p_1^{k,(p)}, p_2^{k,(p)}, \ldots, p_{l-1}^{k,(p)}$ and $r_l^{k,(p)}$ are the coefficients of the consequent linear function of the kth fuzzy rule in view p.

From (15), we see that the outputs of the previous components, i.e., $y_1, y_2, \ldots, y_{l-1}$ are involved in the antecedent of the fuzzy rule. It seems that the fuzzy rule shown in (15) becomes incomprehensible because the outputs $y_1, y_2, \ldots, y_{l-1}$ have no physical meanings. However, such an issue can be resolved by considering the component C_l with the fuzzy rule shown in (15) as another equivalent TSK fuzzy system with high interpretable fuzzy rules. The output of the component C_l can be expressed as

$$y_l = \sum_{k=1}^{K_l} f_l^{k,(p)}(\mathbf{x}^{(p)})\Big(\prod_{i=1}^{d} \mu_i^k(x_i^{(p)}) \prod_{j=1}^{l-1} \mu_{y_j}^k(y_j(\mathbf{x}^{(p)}))\Big). \quad (16)$$

Please note, according the discussion in [29], we omit the denominator $\sum_{k=1}^{K_l} \prod_{i=1}^{d} \mu_i^k(x_i^{(p)}) \prod_{j=1}^{l-1} \mu_{y_j}^k(y_j(\mathbf{x}^{(p)}))$ in (16) when the Gaussian function is adopted as the fuzzy membership function. It is easy to see that a new TSK fuzzy system can also achieve the same output shown in (16) whose kth fuzzy rule can be expressed as,

$$\begin{aligned}&\text{If } x_1^{(p)} \text{ is } A_1^k \wedge x_2^{(p)} \text{ is } A_2^k \wedge \ldots \wedge x_d^{(p)} \text{ is } A_d^k, \\ &\text{then } f_l^k(\mathbf{x}^{(p)}) = p_1^{k,(p)} x_1^{(p)} + \ldots + p_{l-1}^{k,(p)} x_{l-1}^{(p)} + r_l^{k,(p)}, \\ &\text{with } CF(\mathbf{x}^{(p)}) = \prod_{j=1}^{l-1} \mu_{y_j}^k(y_j(\mathbf{x}^{(p)})). \quad (17)\end{aligned}$$

In (17), we see that only the original features are involved in the antecedent and the outputs of previous components now are hidden in a function $\mathrm{CF}(\mathbf{x}^{(p)})$ termed as the dynamic certainty factor. As a result, the antecedent is still comprehensible. As a new concept, $\mathrm{CF}(\mathbf{x}^{(p)})$ here is a function of the input $\mathbf{x}^{(p)}$, which can be interpreted as the confidence degree that the fuzzy rule can act on $\mathbf{x}^{(p)}$.

Therefore, based on the above equivalence between the two fuzzy systems, the interpretability of DVR-TSK-FS can be insured.

C. MULTI-VIEW LEARNING AND VIEW REDUCTION

With regards to multi-view learning for each component in DVR-TSK-FS, it it is expected that for an unseen sample, its decision result of each view should be as consistent as possible. Therefore, the multi-view learning mechanism on the multi-view EEG training dataset χ can be expressed as

$$\Theta = \frac{\alpha}{2}\sum_{p=1}^{P}\sum_{i=1}^{N}\left\|(\mathbf{p}_g^{(p)})^T\mathbf{x}_{gi}^{(p)} - \frac{1}{P-1}\sum_{l=1,l\neq p}^{P}(\tilde{\mathbf{p}}_g^{(l)})^T\mathbf{x}_{gi}^{(l)}\right\|^2, \tag{18}$$

where $\mathbf{x}_{gi}^{(p)}$ denotes the input vector in view p mapped from $\mathbf{x}_i^{(p)}$ through (9), $\mathbf{p}_g^{(p)}$ denotes the consequent parameter in view p. $\tilde{\mathbf{p}}_g^{(l)}$ can be taken as the prior knowledge of each view that is obtained by (14). In (18), the item $\frac{1}{P-1}\sum_{l=1,l\neq p}^{P}(\tilde{\mathbf{p}}_g^{(l)})^T\mathbf{x}_{gi}^{(l)}$ represents the mean of the prior decision result of all views, $(\mathbf{p}_g^{(p)})^T\mathbf{x}_{gi}^{(p)}$ denotes the expected decision result of the pth view. Therefore, the multi-view learning mechanism can be implemented by minimizing (18) such that the decision result of each view can be consistent. α in (18) is used to control the degree of consistency between each view. It is often determined by cross-validation accroding to the corresponding EEG training dataset.

As we stated in the first section, weak views may exert negative influences on the final decision results during the multi-view learning procedures. In order to reduce weak views, the variant Shannon entropy [9], [10] introduced to learn the weight of each view. Comparing with the Shannon entropy used in other weight learning strategies, the variant one is very different in that it employs a sample-distribution-dependent parameter to control the view weight learning such that the weight can reflect the amount of pattern information involved in each view more veritably. The view reduction mechanism can be formulated as

$$\begin{aligned}\Delta = &\frac{\beta}{2}\sum_{p=1}^{P}w_p\delta_p\sum_{i=1}^{N}\left\|(\mathbf{p}_g^{(p)})^T\mathbf{x}_{gi}^{(p)} - t_i\right\|^2\\ &+\frac{N}{P}\sum_{p=1}^{P}w_p\log\delta_p w_p,\\ &s.t.\ \sum_{p=1}^{P}w_p = 1,\quad 0\le w_p\le 1.\end{aligned} \tag{19}$$

In (20), we introduce a weight vector $\mathbf{w} = [w_1, w_2, \ldots, w_P]^T$ in which each element represents the weight of each view. This first item in Δ is used to control the training errors, and the second item is a variant Shannon entropy used to learn the view weight. Here, please note that comparing with the classic Shannon entropy, a sample-distribution-dependent parameter δ_p is embedded to control the weight learning. We say that δ_p is sample-distribution-dependent, it means that δ_p should be derived from the sample distribution in each view. In the field of probability statistics, the deviation, variance and mean or their combinations are commonly-used indicators used to measure the object distribution. For example, in [11], the ratio of the variance and mean, i.e., σ^2/μ are used to measure the dispersion degree of objects. Smaller dispersion degree indicates a compact object distribution. In (20), we expect that the view having large dispersion degree should be discarded. Therefore, in this study, we set $\delta_p = \mu/\sigma^2$ to control the weight learning.

In order to reduce the weak views, we introduce a threshold as the upper bound of w_p. The view reduction principle can be asserted that when $w_p \le 1/\sqrt{NP}$, the correpsonding view is reduced. It is abvious that when the number of view P is very big, it is more reasonable to use $w_p \le 1/P$ as the reduction condition. However, when P is not very big, the reduction condition $w_p \le 1/P$ is not suitable. As all we know that $1/P = 1/\sqrt{P^2} = 1/\sqrt{PP}$, so in order to search for the balance between the bigger P and the samller one, we use N, i.e., the number of objects in each view to replace one P. Thus, the the reduction condition becomes $w_p \le 1/\sqrt{NP}$ which is also suitable for multi-view datasets with smaller P.

D. OBJECTIVE FUNCTION

Based on multi-view learning and view reduction, the objective function of each component in DVR-TSK-FS can be formulated as follows,

$$\begin{aligned}&J(\mathbf{p}_g^{(p)}, \mathbf{w})\\ &= \frac{1}{2}\min_{\mathbf{p}_g^{(p)},\mathbf{w}}\sum_{p=1}^{P}(\mathbf{p}_g^{(p)})^T\mathbf{p}_g^{(p)}\\ &\quad+\frac{\alpha}{2}\sum_{p=1}^{P}\sum_{i=1}^{N}\left\|(\mathbf{p}_g^{(p)})^T\mathbf{x}_{gi}^{(p)} - \frac{1}{P-1}\sum_{l=1,l\neq p}^{P}(\tilde{\mathbf{p}}_g^{(l)})^T\mathbf{x}_{gi}^{(l)}\right\|^2\\ &\quad+\frac{\beta}{2}\sum_{p=1}^{P}w_p\delta_p\sum_{i=1}^{N}\left\|(\mathbf{p}_{g,c}^{(p)})^T\mathbf{x}_{gi}^{(p)} - t_i\right\|^2\\ &\quad+\frac{N}{P}\sum_{p=1}^{P}w_p\log\delta_p w_p,\\ &\quad s.t.\ \sum_{p=1}^{P}w_p = 1,\quad 0\le w_p\le 1.\end{aligned} \tag{20}$$

In order to search for the extremum of $J(\mathbf{p}_g^{(p)}, \mathbf{w})$ subjected to the condition of w_p, a Lagrangian multiplier λ is introduced and the corresponding Lagrangian objective function is given

as

$$L = J(\mathbf{p}_g^{(p)}, \mathbf{w}) + \lambda(1 - \sum_{p=1}^{P} w_p). \tag{21}$$

Let $\partial L/\partial \mathbf{p}_g^{(p)} = 0$ and $\partial L/\partial \mathbf{w} = 0$, we can get two updated rules as

$$\begin{aligned}\mathbf{p}_g^{(p)} = &\left(\delta_p w_p \sum_{i=1}^{N} (\mathbf{x}_{gi}^{(p)})^T \mathbf{x}_{gi}^{(p)} + \beta \mathbf{I}_{((d_p+1)K)\times((d_p+1)K)}\right.\\ &\left. + \alpha \sum_{i=1}^{N} \mathbf{x}_{gi}^{(p)} (\mathbf{x}_{gi}^{(p)})^T\right)^{-1}\\ &\times \left(\delta_p w_p \sum_{i=1}^{N} \mathbf{x}_{gi}^{(p)} t_i + \frac{\alpha}{K-1} \sum_{l=1, l\neq p}^{P} \sum_{i=1}^{N} \mathbf{x}_{gi}^{(l)} \tilde{\mathbf{p}}_g^{(l)}\right)\end{aligned} \tag{22}$$

$$w_p = \frac{\frac{1}{\delta_p} \exp(-P \sum_{i=1}^{N} \left\| (\mathbf{p}_g^{(p)})^T \mathbf{x}_{gi}^{(p)} - t_i \right\|^2 / N)}{\sum_{h=1}^{P} \frac{1}{\delta_h} \exp(-P \sum_{i=1}^{N} \left\| (\mathbf{p}_g^{(p)})^T \mathbf{x}_{gi}^{(p)} - t_i \right\|^2 / N)}. \tag{23}$$

With above two updated rules in terms of $\mathbf{p}_g^{(p)}$ and w_p, a iteration strategy also used in FCM is employed to search for their optimal values. Please note, some weak views may be reduced during the iteration procedure. But we should keep in mind that the objective function is subject to the condition $\sum_{p=1}^{P} w_p = 1$. So, during the iteration procedure, we should dynamically adjust w_p by

$$w_p' = \frac{w_p}{\sum_{p'=1}^{P'} w_{p'}}, \tag{24}$$

where P' is the number of views after view reduction.

With the obtained $\mathbf{p}_g^{(p)}$ and w_p by a EEG training dataset, for an unseen sample in each component, its decision result can be computed as

$$y_i = f(\mathbf{x}_i) = \sum_{p=1}^{P} w_p (\mathbf{p}_g^{(p)})^T \mathbf{x}_{gi}^{(p)}. \tag{25}$$

E. DEEP LEARNING ALGORITHM

Since components in DVR-TSK-FS are concatenated in a stacked way, we develop a deep learning algorithm, i.e., the layer by layer learning for DVR-TSK-FS. The pseudocode of the deep learning algorithm is listed in Algorithm 1.

Algorithm 1 Deep Learning Algorithm for DVR-TSK-FS

Input:
1. EEG training dataset $\chi = \{\mathbf{X}^{(p)}\}_{p=1}^{P}$ where $\mathbf{X}^{(p)} = \{\mathbf{x}_i^{(p)}\}_{i=1}^{N}$ and the corresponding label vector $\mathbf{T} = \{t_i\}_{i=1}^{N}$.
2. Number of components, i.e., the depth DP.
3. Number of fuzzy rules in each component, i.e., K_l, $1 \leq l \leq L$.
4. Regularization parameters α and β in the objective function of each component.
Output:
$\mathbf{p}_g^{(p)}$ and w_p of each view in the last component.

1. Set $l = 1$ which indicates the current component.
2. Initialize $\mathbf{Y}$ to empty which represents the output of the current component.
Repeat
3. Initialize the weight vector $\mathbf{w}$ of current component by setting $w_p = 1/P$.
4. Compute δ_p by $\delta_p = \mu/\sigma^2$ for each view in current component.
5. Use FCM to obtain the antecedent parameters, then obtain the consequent parameters $\tilde{\mathbf{p}}_g$ by (14) as prior knowledge.
6. Compute $\mathbf{p}_g^{(p)}$ by (23).
7. Compute w_p by (24).
8. If $w_p \leq 1/\sqrt{NP}$, then reduce view p and set $P = P - 1$.
9. If step 6 is executed, then update w_p by (25).
10. If the difference of $J(\mathbf{p}_{g,c}^m, \mathbf{w})$ bewteen two iterations is less than ε, then the iteration stops, set $\mathbf{Y} = \sum_{p=1}^{P} w_p (\mathbf{p}_{g,c}^{(p)})^T \mathbf{x}_{gi}^{(p)}$ and augment $\mathbf{Y}$ into current feature space; Otherwise, go to step **6** and continue;
Until $l = L$

IV. EXPERIMENTAL RESULTS

In this section, the multi-view epileptic EEG data we introduced in Section II.C are used in the following experiments. For comparison studies, two classical classification algorithms, i.e., SVM [12] and 1-TSK-FS [6], two multi-view classification algorithms i.e., MV-L2-SVM [2], and MV-TSK-FS [4] and a deep TSK fuzzy system D-TSK-FS [13] are introduced.

A. SETUP

In our experiments, the multi-view EEG dataset is randomly grouped into three partitions, the first one contains 20% samples is used for 5 cross-validation to search for the optimal parameters, the second one contains 60% objects is used for model training with the optimal parameters obtained in cross-validation stage, and the last one is used for testing. In the testing procedure, each algorithm is repeatedly executed 10 times and the average testing accuracy and the corresponding standard deviation are reported. Parameters in all benchmarking algorithms are set according to Table 2.

All benchmarking algorithms and DVR-TSK-FS are coded in the MATLAB (version: 2012b) environment on a PC with 4 cores of I5-4950 with 32G of memory.

TABLE 2. Search intervals of all bechmarking algorithms.

Algorithms	Search intervals
SVM	$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2})$ where $\sigma\in\{2^{-5},2^{-4},...,2^0,2^1,...,2^4,2^5\}$. Penalty parameter: $C\in\{10^{-5},10^{-4},...,10^0,10^1,...,10^4,10^5\}$
1-TSK-FS	The number of fuzzy rules $K\in\{5,10,15,...,30\}$, The regularization parameter $\eta_{\mathbf{p}_g}\in\{10^{-3},10^{-2},...,10^3\}$.
MV-L2-SVM	$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2})$ where $\sigma\in\{2^{-5},2^{-4},...,2^0,2^1,...,2^4,2^5\}$. Penalty parameter: $C\in\{10^{-5},10^{-4},...,10^0,10^1,...,10^4,10^5\}$
MV-TSK-FS	The number of fuzzy rules $K\in\{5,10,15,...,30\}$, the regularization parameters $\lambda^{\mathbf{p}_g}\in\{10^{-3},10^{-2},...,10^3\}$, $\lambda^{co}\in\{10^{-3},10^{-2},...,10^3\}$ and $\lambda^{w}\in\{10^{-3},10^{-2},...,10^3\}$.
DVR-MV-TSK-FS	The number of fuzzy rules $K\in\{5,10,15,...,30\}$, The regularization parameter $\alpha\in\{10^{-3},10^{-2},...,10^3\}$, The regularization parameter $\beta\in\{10^{-3},10^{-2},...,10^3\}$.

B. ON MULTI-VIEW EEG DATA

Based on the original 5 groups EEG data and three feature extraction methods, we construct one multi-view EEG data contain Group A and B as health subjects and Group C as epileptic subjects.

The experimental results are observed from three aspects, i.e., view reduction, interpretability and the deep structure.

In order to observe the view reduction in each component, we select the first component, and print the iteration steps on multi-view EEG data. Some key steps in terms of view weights are shown in Table 3. From Table 3, we see that the reduction threshold is $1/\sqrt{300\times 3}=0.0333$(the size of training samples is 300 and the number of views is 3), when the iteration count of DVR-TSK-FS reaches to 98, the weight of the second view, i.e., the features extracted by STFT, is smaller than $1/\sqrt{NP}$ such that the second view should be reduced according to our view reduction principle. In the following iterations, the weights of the remnant views are adjusted by (25) to obey the condition ''sum to one''.

Therefore, we see that the weak view, i.e., the second view is automatically reduced by our adopted view reduction mechanism. With the trained component, the testing accuracy of this component is 83.43%. In addition, to emphasize the power of view reduction mechanism, we also train each component without view reduction mechanism. That is to say, when the weight of one view is smaller than $1/\sqrt{NP}$, we do not carry out view reduction. The testing accuracy of the firstly component without view reduction is 78.76% which is inferior to that of the first component with view reduction. From the comparison study, we believe that weak views indeed bring negative influences to the final performance.

TABLE 3. Iteration results of DVR-TSK-FS in terms of the weight of each view.

Iteration count	Weight vector $\mathbf{w}=[w_1,w_2,w_3]$	$1/\sqrt{NP}$
0	$\mathbf{w}=[0.3333,0.3333,0.3333]$	$1/\sqrt{300\times 3}$
1	$\mathbf{w}=[0.4813,0.1769,0.3418]$	
...	...	
...	...	
98	$\mathbf{w}=[0.5873,0.0061,0.4066]$	
99	$\mathbf{w}=[0.6214,0.3786]$	
...	...	
156	$\mathbf{w}=[0.5923,0.4077]$	

In order to observe the interpretability of DVR-TSK-FS, Table 4 list the some trained fuzzy rules on the second and third component for the first view. It is obvious that the output from the first component is not involved in the antecedents and hence does not complicate the interpretation of the consequents of the fuzzy rules because it is hidden in CF of these rules. Each row in Table 4 can be translated into a fuzzy rule with a linear consequent with CF. For example, the first fuzzy rule in component C_1 can be expressed as

If x_1 is $A_1^1(1.8762, 0.2134)\wedge x_2$ is $A_2^1(1.3431, 0.3767)\wedge x_3$ is $A_3^1(2.5637, 0.2342)\wedge x_4$ is $A_4^1(2.2310, 0.1176)\wedge x_5$ is $A_5^1(-2.1345, 0.7123)\wedge x_6$ is $A_6^1(-2.1345, 0.2123)$

then $f^1(\mathbf{x})=1.3066x_1+1.3207x_2-1.2970x_3-0.5135x_4+1.8413x_5-1.3517x_6+0.0141$, with $\mathrm{CF}(\mathbf{x})=e^{-\frac{1}{2}\left(\frac{y_1(\mathbf{x})-3.6782}{0.6753}\right)^2}$.

Furthermore, in order to deeply observe the promising performance of EEG signals recognition, we report the testing accuracy of each component and introduce other benchmarking approaches for comparison studies. Table 5 gives the comparison results in terms of average training accuracy, testing accuracy and their corresponding standard deviation.

In Table 5, since SVM and 1-TSK-FS are two single-view classification algorithms, we report the results of them from two aspects, i.e., the results on the original feature space and the average results on all views. From Table 4, we see that the classification performance of SVM and 1-TSK-FS in the original feature space is better than that of average of all views. This is because each view is considered as contributing equally to the classification accuracy and hence their performance on the second view significantly drags down the average results. In MV-TSK-FS and DVR-TSK-FS, the Shannon entropy and its variant are introduced to learn the view weight of each view, respectively. However, by comparing the weight vectors from MV-TSK-FS and DVR-TSK-FS (in Table 2), we find that with the parameter δ_p used in the Shannon entropy, the weight of the second view in DVR-TSK-FS is much smaller than that in MV-TSK-FS. Moreover, by introducing the reduction condition $w_p\leq 1/\sqrt{NP}$, the second view is reduced. Also, we find that with the reduction condition, the performance is enhanced compared with that without the condition.

D-TSK-FS is also a deep-structure based fuzzy system. However, by comparing the performance of each

TABLE 4. Rules obtained from multi-view EEG data.

The kth fuzzy rule: If $x_1^{(p)}$ is $A_1^k \wedge x_2^{(p)}$ is $A_2^k \wedge \ldots \wedge x_d^{(p)}$ is A_d^k, then $f_l^k(\mathbf{x}^{(p)}) = p_1^{k,(p)}x_1^{(p)} + \ldots + p_d^{k,(p)}x_d^{(p)} + r_l^{k,(p)}$, with $\mathrm{CF}(\mathbf{x}^{(p)}) = \prod_{j=1}^{l-1} \mu_{y_j}^k(y_j(\mathbf{x}^{(p)}))$ ..

Component	No. of rules	Antecedent parameters: $\mathbf{c}^k = [c_1^k, c_2^k, \ldots, c_d^k]$, $\boldsymbol{\delta}^k = [\delta_1^k, \delta_2^k, \ldots, \delta_d^k]$	Consequent parameters: $\mathbf{p}^k = [p_1^k, p_2^k, \ldots, p_d^k, r_l^k]$	$y(\mathbf{x})$	CF
C_2	1	$\mathbf{c}^1 = [1.8762, 1.3431, 2.5673, 2.231, -2.1345, -2.1345]$ $\boldsymbol{\delta}^1 = [0.2134, 0.3767, 0.2342, 0.1176, 0.7123, 0.2123]$	$\mathbf{p}^1 = [1.3066, 1.3207, -1.2970, -0.5135, 1.8413, -1.3517, 0.0141]$	$y_1(\mathbf{x}) = 3.7536$	$\mathrm{CF}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\left(\frac{y_1(\mathbf{x})-3.6782}{0.6753}\right)^2}$
	2	$\mathbf{c}^2 = [2.2234, -1.3212, 1.2321, 1.8973, 2.0094, -2.0214]$ $\boldsymbol{\delta}^2 = [0.1133, 0.4678, 0.3187, 0.1563, 0.6924, 0.3567]$	$\mathbf{p}^2 = [-0.0017, 0.1019, -0.2353, -0.0608, -0.1608, 0.2536, 0.0043]$		$\mathrm{CF}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\left(\frac{y_1(\mathbf{x})-3.5135}{0.1134}\right)^2}$
	3	$\mathbf{c}^1 = [2.9084, 2.5876, 2.1478, 1.7865, -1.7623, -1.9821]$ $\boldsymbol{\delta}^1 = [0.1873, 0.2789, 0.1761, 0.4321, 0.6521, 0.3123]$	$\mathbf{p}^1 = [2.4125, -1.0027, 1.4892, -0.3021, 1.5092, -2.4456, 0.0621]$		$\mathrm{CF}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\left(\frac{y_1(\mathbf{x})-3.6721}{0.2245}\right)^2}$
C_3	1	$\mathbf{c}^2 = [1.9822, -1.3232, 1.2367, 1.9085, 2.1111, -3.9822]$ $\boldsymbol{\delta}^2 = [0.2341, 0.5982, 0.4214, 0.2589, 0.7081, 0.4189]$	$\mathbf{p}^2 = [-1.3357, -0.3216, -0.1963, -0.1648, -0.1427, 0.8424, 0.0469]$	$y_1(\mathbf{x}) = 3.9785$ $y_2(\mathbf{x}) = 4.0135$	$\mathrm{CF}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\left(\left(\frac{y_1(\mathbf{x})-4.0156}{0.3245}\right)^2 + \left(\frac{y_2(\mathbf{x})-3.4438}{0.3873}\right)^2\right)}$
	2	$\mathbf{c}^2 = [3.8123, 2.2982, 2.2234, 2.0912, 2.1123, -3.1232]$ $\boldsymbol{\delta}^2 = [0.2376, 0.5012, 0.1234, 0.4256, 0.3009, 0.2098]$	$\mathbf{p}^2 = [-0.0252, 0.1152, -0.3000, -0.0412, -0.1764, 0.2621, -0.2135]$		$\mathrm{CF}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\left(\left(\frac{y_1(\mathbf{x})-3.6754}{0.7864}\right)^2 + \left(\frac{y_2(\mathbf{x})-4.0124}{0.3421}\right)^2\right)}$
	3	$\mathbf{c}^1 = [2.2212, 3.8756, -1.9712, 1.9879, -1.6753, -1.3456]$ $\boldsymbol{\delta}^1 = [0.1678, 0.2789, 0.1087, 0.2456, 0.6789, 0.3212]$	$\mathbf{p}^1 = [2.4223, 1.5421, -1.4213, -0.6732, 1.9127, -1.4217, 0.1171]$		$\mathrm{CF}(\mathbf{x}) = \mathrm{e}^{-\frac{1}{2}\left(\left(\frac{y_1(\mathbf{x})-3.5744}{0.7999}\right)^2 + \left(\frac{y_2(\mathbf{x})-3.8134}{0.5421}\right)^2\right)}$

TABLE 5. Comparison results on the EEG multi-view dataset (The result in parentheses is the standard deviation).

Algorithms		Training accuracy	Testing accuracy	Weight vector $\mathbf{w} = [w_1, w_2, w_3]$
SVM (original feature space)		0.8224 (0.0001)	0.8211 (0.0001)	---
SVM (average of all views)		0.7864 (0.0001)	0.7801 (0.0001)	---
1-TSK-FS (original feature space)		0.8198 (0.0005)	0.8024 (0.0004)	---
1-TSK-FS (average of all views)		0.7786 (0.0002)	0.7765 (0.0002)	---
MV-TSK-FS		0.8298 (0.0001)	0.82761 (0.0001)	[0.4921, 0.0425, 0.4654]
D-TSK-FS	C_1	0.8287 (0.0002)	0.8100 (0.0004)	---
	C_2	0.8345 (0.0002)	0.8356 (0.0003)	---
	C_3	0.8432 (0.0002	0.8421 (0.0001)	---
DVR-TSK-FS	C_1	0.8510 (0.0001)	0.8343 (0.0004)	[0.5874, 0.4126]
	C_2	0.8545 (0.0002)	0.8401 (0.0004)	[0.5542, 0.4458]
	C_3	0.8623 (0.0001)	0.8489 (0.0003)	[0.5712, 0.4288]

compoment (C_1, C_2, and C_3) with DVR-TSK-FS, we find that our approch DVR-TSK-FS performs better than D-TSK-FS. This is because, in our experiments, D-TSK-FS dose not adopt multi-view learning mechanism but only joints all features from each view together for its recgonitoin tasks. In addtition, in D-TSK-FS, only the output of the previous component is agumented into the next one, not all the prvious outputs. Therefore, the manifold structure cannot be opened enough.

Therefore, from the experimental results on the multi-view EEG dataset, we can draw the following conlusions:

(1) The parameter δ_p derived from the object distribution in each view can control the view weight learning in an effective manner. That is to say, comparing with MV-TSK-FS, the weight of each view in the proposed algorithm matches the pattern information involved in each view more successfully.

(2) The upper bound $1/\sqrt{NP}$ used in the reduction condition is more reliable than $1/P$. In this experiment, $1/P = 0.3333$ is near to the weight of the third view. Although it can also successfully reduce the second view, it does not reliable.

(3) The deep structure used in DVR-TSK-FS can insure its generalization capacity.

V. CONCLUSION

In this paper, a novel deep view-reduction TSK fuzzy system DVR-TSK-FS is proposed in which two effective mechanisms associating with the deep structure are developed to

insure its performance. In the first mechanism, a sample-distribution-dependent parameter is defined to control the learning of the view weight during the multi-view learning in each component. This parameter is user-free and set according to the feature space in advance such that the learnt weight of each view indeed reflects the amount of pattern information involved in each view. The second mechanism sets out a view reduction principle that weak views are automatically reduced by comparing the learnt weight with a fixed threshold which is automatically set according to the number of objects and the dimension of feature space. Based on the stacked generalization principle, all components are linked in a stacked way. The proposed algorithm DVR-TSK-FS is verified on a multi-view EEG dataset and its performance is compared with other benchmarking algorithms.

REFERENCES

[1] Z. Jiang, F.-L. Chung, and S. Wang, "Recognition of multiclass epileptic EEG signals based on knowledge and label space inductive transfer," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 630–642, Apr. 2019.

[2] T. Zhang and W. Z. Chen, "LMD based features for the automatic seizure detection of EEG signals using SVM," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 8, pp. 1100–1108, Aug. 2017.

[3] L. Xie, Z. Deng, P. Xu, K.-S. Choi, and S. Wang, "Generalized hidden-mapping transductive transfer learning for recognition of epileptic electroencephalogram signals," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2200–2214, Jun. 2019.

[4] Y. Jiang, Z. Deng, F.-L. Chung, G. Wang, P. Qian, K.-S. Choi, and S. Wang, "Recognition of epileptic EEG signals using a novel multiview TSK fuzzy system," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 1, pp. 3–20, Feb. 2017.

[5] Y. Jiang, D. Wu, Z. Deng, P. Qian, J. Wang, G. Wang, F.-L. Chung, K.-S. Choi, and S. Wang, "Seizure classification from EEG signals using transfer learning, semi-supervised learning and TSK fuzzy system," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2270–2284, Dec. 2017.

[6] Z. Deng, Y. Jiang, K.-S. Choi, F.-L. Chung, and S. Wang, "Knowledge-leverage-based TSK fuzzy system modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 8, pp. 1200–1212, Aug. 2013.

[7] C. Yang, Z. Deng, K.-S. Choi, and S. Wang, "Takagi-Sugeno-Kang transfer learning fuzzy logic system for the adaptive recognition of epileptic electroencephalogram signals," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1079–1094, Oct. 2016.

[8] L.-X. Wang, "Analysis and design of hierarchical fuzzy systems," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 5, pp. 617–624, Oct. 1999.

[9] A. Delgado and E. C. Reyes, "Applying Shannon entropy to select alternative plants as food for livestock: A case study in ecuador," in *Proc. CACIDI IEEE Conf. Comput. Sci.*, Buenos Aires, Argentina, Nov. 2016, pp. 1–5.

[10] H. Zhang and H. Zhang, "The analysis of the associated Shannon entropy on the populations with homologous genotypes mating and homologous phenotypes mating," in *Proc. 26th Chin. Control Decis. Conf. (CCDC)*, Changsha, China, May 2014, pp. 3885–3889.

[11] D. R. Cox and P. A. Lewis, *The Statistical Analysis of Series of Events*, London, U.K.: Methuen, 1966.

[12] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[13] T. Zhou, H. Ishibuchi, and S. Wang, "Stacked blockwise combination of interpretable TSK fuzzy classifiers by negative correlation learning," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 6, pp. 3327–3341, Dec. 2018.

[14] F. Zhao and Q. Tang, "A KNN learning algorithm for collusion-resistant spectrum auction in small cell networks," *IEEE Access*, vol. 6, pp. 45796–45803, 2018.

[15] S. Huang, Y. Lyu, Y. Peng, and M. Huang, "Analysis of factors influencing rockfall runout distance and prediction model based on an improved KNN algorithm," *IEEE Access*, vol. 7, pp. 66739–66752, 2019.

[16] M. Yang, Y. Zuo, M. Chen, and X. Yu, "Scalable distributed kNN processing on clustered data streams," *IEEE Access*, vol. 7, pp. 103198–103208, 2019.

[17] Q. Hu, X. Che, L. Zhang, D. Zhang, M. Guo, and D. Yu, "Rank entropy-based decision trees for monotonic classification," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 11, pp. 2052–2064, Nov. 2012.

[18] S. Tsang, B. Kao, K. Y. Yip, W. S. Ho, and S. D. Lee, "Decision trees for uncertain data," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 64–78, Jan. 2011.

[19] E. A. Vivaldi and A. Bassi, "Frequency domain analysis of Sleep EEG for visualization and automated state detection," in *Proc. 28th Annu Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2006, pp. 3740–3743.

[20] G. C. Y. Fong, P. U. Shah, M. N. Gee, J. M. Serratosa, I. P. Castroviejo, S. Khan, S. H. Ravat, J. Mani, Y. Huang, H. Z. Zhao, M. T. Medina, L. J. Treiman, G. Pineda, and A. V. Delgado-Escueta, "Childhood absence epilepsy with tonic-clonic seizures and electroencephalogram 3-4-Hz spike and multispike-slow wave complexes: Linkage to chromosome 8q24," *Amer. J. Human Genet.*, vol. 63, no. 4, pp. 1117–1129, Oct. 1998.

[21] S. Blanco, S. Kochen, O. A. Rosso, and P. Salgado, "Applying timefrequency analysis to seizure EEG activity," *IEEE Eng. Med. Biol. Mag.*, vol. 16, no. 1, pp. 64–71, Jan./Feb. 1997.

[22] Z. Zhang, H. Kawabata, and Z.-Q. Liu, "EEG analysis using fast wavelet transform," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 4, Oct. 2000, pp. 2959–2964.

[23] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[24] Y. Jiang, J. Zheng, X. Gu, J. Xue, and P. Qian, "A novel synthetic CT generation method using multitask maximum entropy clustering," *IEEE Access*, vol. 7, pp. 119644–119653, 2019.

[25] P. Qian, Y. Jiang, Z. Deng, L. Hu, S. Sun, S. Wang, and R. F. Muzic, "Cluster prototypes and fuzzy memberships jointly leveraged cross-domain maximum entropy clustering," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 181–193, Jan. 2016.

[26] P. Qian, J. Zhou, Y. Jiang, F. Liang, K. Zhao, S. Wang, K. Su, and R. Muzic, "Multi-view maximum entropy clustering by jointly leveraging inter-view collaborations and intra-view-weighted attributes," *IEEE Access*, vol. 6, pp. 28594–28610, 2018.

[27] Z. Deng, Y. Jiang, F.-L. Chung, H. Ishibuchi, K. S. Choi, and S. Wang, "Transfer prototype-based fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1210–1232, Oct. 2016.

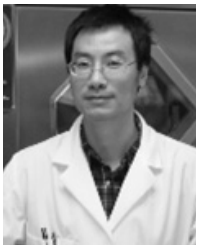
[28] P. Qian, Y. Jiang, S. Wang, K.-H. Su, J. Wang, L. Hu, and R. F. Muzic, "Affinity and penalty jointly constrained spectral clustering with all-compatibility, flexibility, and robustness," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1123–1138, May 2017.

[29] S. Wang, F.-L. Chung, H.-B. Shen, and D. Hu, "Cascaded centralized TSK fuzzy system: Universal approximator and high interpretation," *Appl. Soft Comput.*, vol. 5, no. 2, pp. 131–145, Jan. 2005.

[30] F.-L. Chung and J.-C. Duan, "On multistage fuzzy neural network modeling," *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 2, pp. 125–142, Apr. 2000.

YUANPENG ZHANG (M'17) received the Ph.D. degree in information engineering from the School of Computer Application Technology, Jiangnan University, in 2018. He is currently an Associate Professor with the Department of Medical Informatics, Nantong University. He is also a Postdoctoral Fellow with the Department of Health Information Technology, The Hong Kong Polytechnic University. He has published about 20 articles in international/national journals, including the IEEE TRANSACTIONS ON FUZZY SYSTEMS, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and *ACM Transactions on Multimedia Computing, Communications, and Applications*. His main research interests include pattern recognition and data mining.

XIANGZHE LI received the M.D. degree of rehabilitation medicine and physical therapy from Nanjing Medical University, in 2017. He is currently a Doctor of rehabilitation medicine and physical therapy with Suzhou Science and Technology Town Hospital. He has published about ten articles in international/national journals, including *Spinal Cord*, *Spinal*, and *Cell Death and Disease*. His main research interests include neuroelectrophysiology and neurological rehabilitation.



JUNQING ZHU received the bachelor's degree in physics from the Shanghai University of Science and Technology, Shanghai, China, in 1989. In 2006, he joined Case Western Reserve University, where he is currently a Senior Staff of radiology. In the past three decades, he has been focused on in vivo molecular imaging. He also puts special emphasis on molecular imaging in neurodegenerative diseases, such as multiple sclerosis (MS), Alzheimer's disease (AD), Parkinson's disease, Epilepsy disease, and DNA damage and repair in cancer. By continuously working with molecular imaging probe group, he pioneered imaging of myelin based on different imaging modalities, such as PET, MRI, and near-infrared fluorescence imaging.

CHUNYING WU received the Ph.D. degree in imaging medicine and nuclear medicine from the School of Medicine, Fudan University, Shanghai, China, in 2003. In 2004, she was a Postdoctoral Fellow with the University of Illinois at Chicago. In 2006, she joined Case Western Reserve University, where she is currently an Instructor of Radiology with the Division of Molecular Imaging Center, Case Center for Imaging Research, School of Medicine. She has extensive experience in radiopharmaceutical development for PET and SPECT imaging. Over the past ten years, her research has focused on the development of small-molecular probes for PET imaging in Alzheimer's disease, multiple sclerosis and DNA damage, and repair in cancer. Her research was selected for Molecular Imaging/CMIIT Basic Science and Neuroscience Summary Session highlight talk in the Society of Nuclear Medicine (SNM) annual meeting, in 2013 and 2016, respectively. Some of her work has also been selected twice as the second place in the poster competition at the annual SNM conferences.

QINFENG WU received the Ph.D. degree of rehabilitation medicine and physical therapy from Nanjing Medical University, in 2016. He is currently an Associate Professor with the Suzhou Hospital afflicted to Nanjing Medical University, and an Associate Chief Physician with Suzhou Science and Technology Town Hospital. He has published about ten articles in international/national journals, including *Cell Death and Diseases*, *PeerJ*, and *Spinal Cord*. His main research interests include neurological rehabilitation and brain networks.