Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

# Horizontal progressive and longitudinal leapfrogging fuzzy classification with feature activity adjustment

Wei Xue [a], Ta Zhou [a,b,*], Jing Cai [b]

[a] School of Electrical and Information Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212100, China
[b] Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, 999077, China

ARTICLE INFO

*Article history:*
Received 28 July 2021
Received in revised form 8 December 2021
Accepted 20 January 2022
Available online 29 January 2022

*Keywords:*
Distance correlation coefficient
EEG signals recognition
Feedback network
Fuzzy classification
Linear mapping

ABSTRACT

Classification accuracy and interpretability are crucial importance for recognizing seizures based on electroencephalogram (EEG) signals. This study presents a novel deep ladder-type Takagi–Sugeno–Kang (TSK) fuzzy classifier (D-LT-TSK) that alternately utilizes horizontal progressive learning and longitudinal leapfrogging learning styles. Based on the nonuniform probability distribution co-generated by the distance correlation (DC) coefficient and random bias matrix, a feature activity adjustment mechanism (DC-FAM) is adopted to adjust the activity of each feature to realize the evolution from full connection to partial connection between the input layer and rule layers of the TSK classifier. Feedforward and feedback neural networks are combined to learn consequent parameters in the Then-part of fuzzy rules, for the sake of strengthening the approximation performance and achieving fast converge capability. To take full advantage of valuable decision-making information, D-LT-TSK is learned in the horizontal progressive and longitudinal leapfrogging learning style by mapping the decision-making information of learning modules into the original input space. Experimental results demonstrated that (1) the highly interpretable D-LT-TSK be capable of yielding satisfactory classification performance by utilizing short fuzzy rules, and (2) the optimization algorithm in the Then-part enhanced the approximation performance and accelerate the convergence speed.

## 1. Introduction

Epilepsy is a transient brain dysfunction caused by abnormally firing neurons in the brain. Approximately 50 million people worldwide suffer from epilepsy, among which three-quarters do not have access to effective treatment [1]. The risk of premature death in people with epilepsy is up to three times higher than that in the general population [1]. Therefore, development of an affordable and reliable epilepsy recognition technology is of crucial importance for achieving population-wide early diagnosis and timely interventions for people with epilepsy, especially those from low-income and middle-income countries. At present, diagnosis of epilepsy mainly depends on patient's medical history and electroencephalogram (EEG) signal. Particularly, EEG signal recognition has played a vital role in the diagnosis of latent epilepsy and various forms of atypical epilepsy [2–9]. In recent years, pattern recognition technology has been caught in the spotlight of attention in epilepsy detection [2,6–8], in the hope of identifying potential at-risk candidates ahead of clinical manifestations for timely effective intervention.

In the development of automated epileptic EEG signal recognition systems, tremendous effort has been made on classification of epileptic signals using pattern recognition [4–7]. One typical example of signal recognition algorithms is the fuzzy system [2–7,10], which is an intelligent fuzzy inference technology that is based on fuzzy sets and fuzzy inference theory. Fuzzy systems not only provide strong approximation ability, but also are of high interpretability  [11]. Given this unique advantage, fuzzy systems have been widely employed in medical diagnosis [2–7]. Of note, fuzzy systems have been extensively applied in the development of automatic recognition classifiers, examples include adaptive neuro-fuzzy inference systems (ANFIS)[2,3], type-2 fuzzy systems [4,5] and Takagi–Sugeno–Kang (TSK) fuzzy systems [6,7,12]. In [3], based on the Granger Causality (GC), the ANFIS and the symplectic geometry embedding dimension, Farokhzadi proposed an adaptive neuro-fuzzy inference system Granger causality (ANFISGC) that can detect linear and nonlinear connections between EEG and magnetoencephalography (MEG) time series. In analyzing the brain-derived time series of epilepsy patients obtained from the MEG inverse problem, the performance of ANFISGC reached approximately at clinical level. Additionally, a multiview learning framework based on a TSK fuzzy classifier called MV-TSK-FS was proposed to recognize epileptic EEG signals in [7]. Through the application of Shannon's

* Corresponding author at: Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, 999077, China.
*E-mail address:* ta.zhou@polyu.edu.hk (T. Zhou).

https://doi.org/10.1016/j.asoc.2022.108511
1568-4946/© 2022 Elsevier B.V. All rights reserved.

entropy-based view-weighted mechanism and multiview collaborative learning mechanism, MV-TSK-FS is capable of presenting importance of each view extracted by different feature extraction methods according to its weight. The final decision can therefore be made based on the weighted output of different views.

However, these automatic epilepsy recognition classifiers are facing the following challenges: (1) Due to the time-varying and non-linear properties of the EEG signal, outstanding approximation and classification performance of a classifier can hardly be guaranteed. (2) Due to the lack of model interpretability, clinical utility of the prediction model is greatly restricted.

Confronted with these challenges, this study presents a deep ladder-type TSK fuzzy classifier (D-LT-TSK) to classify epileptic EEG signals. By integrating horizontal progressive learning and longitudinal leapfrogging learning styles, this study constructs a ladder-type learning structure that provides the possibility to avoid sacrificing the interpretability of classifiers by simply adding fuzzy rules when improving the classification performance. Our main contributions to D-LT-TSK can be divided into four parts:

(1) In the IF-part of the fuzzy rules, a feature activity adjustment mechanism (i.e., the DC-FAM) consisting of the distance correlation coefficient and random probability adjusts the activity of each feature participating in each rule. DC-FAM can effectively improve the applicability of the classifier in the face of complex and changeable systems by expanding the differences of the participating features in each rule.

(2) Under the initial consequent parameters obtained by the extreme learning machine (ELM) [13], an adaptive moment estimation (Adam) optimization algorithm [14] that is suitable for solving noisy and nonstationary targets, as in the case of an epileptic EEG signal, is adopted to efficiently solve consequent parameters in the Then-part of the fuzzy rules. In this study, experiments showed that the ELM-based Adam algorithm has strong approximation performance and the capability for fast converge.

(3) In each learning module, multiple zero-order TSK fuzzy classifiers are trained simultaneously in the same input space. In the meantime, the classifier with the most satisfactory classification performance is retained as the prediction function of the learning module. Because the deep structure is built with these high-performance learning modules, the D-LT-TSK can obtain a strong learning ability with a shallow structure.

(4) By mapping the decision-making information of the learning modules into the original input space, the proposed ladder-type structure is learned in the horizontal progressive and longitudinal leapfrogging learning styles, which can fully utilize the effective decision-making information. Additionally, the classification performance of the D-LT-TSK is improved due to the horizontal progressive learning guided by a single learning module in the current layer and the longitudinal leapfrogging learning co-guided by doubled learning modules in the front layer. Moreover, the input space of each learning module is consistent with the original input space. Therefore, the physical interpretation of the original space can be translated to each learning module. This indicates that the D-LT-TSK can enhance the model interpretability.

The remaining sections of this paper are organized as follows. In Sections 2 and 3, we briefly review the theoretical knowledge involved and details of D-LT-TSK construction, respectively. Experimental results are reported in Section 4. Finally, Section 5 concludes this study.

## 2. Related works

In recent years, machine learning (ML) techniques [15–19] and deep learning (DL) methods [20–24] have been caught in the spotlight of attention in automatic epilepsy detection. In addition to the fuzzy classification methods mentioned in Section 1, the current mainstream ML methods for epilepsy detection mainly include support vector machines (SVM) [15–17] and extreme learning machines (ELM) [18,19,25]. To overcome the shortcoming of multiscale entropy (MSE) and multiscale permutation entropy (MPE), Sukriti et al. [15] proposed an SVM-based seizure detection system for classifying EEG signals based on two new features, multiscale dispersion entropy (MDE) and refined composite multiscale dispersion entropy (RCMDE). Besides, Patidar et al. [16] firstly adopted least squares support vector machine (LS-SVM) with radial basis function (RBF) kernel function to explore the feasibility of Kraskov entropy based on adjustable tunable-Q wavelet transform (TQWT). Combining the Mahalanobis-similarity-based feature with the sample entropy, Song et al. [18] proposed a novel feature-fusion (MS-SE-FF) method to construct an ELM-based feature-fusion seizure detection method. The experimental results show that the proposed method has a good effect on epileptic seizure detection while maintaining efficiency and simplicity. With the increase of available data and the development of computers, substantial research has been done to detect epilepsy using DL models such as convolutional neural networks (CNNs) [20–22,25] and long short-term memory (LSTM) neural networks [23,24,26]. Based on CNNs and autoencoders, Wen et al. [21] proposed an autoencoders-based deep convolution network (AE-CDNN) to perform unsupervised feature learning from EEG in epilepsy. In the encoder part, features are extracted by CNN, which constantly iterates multiple convolution kernels' convolution and down-sampling to reduce the number of features. Besides, Chatzichristos et al. [26] innovatively introduced attention-gated U-nets into the application of EEG based seizure detection. To produce robust predictions, they fuse the multi-view attention-gated U-net outputs with a bidirectional LSTM with 4 hidden nodes.

In this study, experiments were performed on a publicly available EEG dataset collected from Bonn University [35]. Table 1 retrieves thirteen meaningful studies of this dataset from 2016 to 2021.

The following content focuses on the theoretical basis involved in this study, mainly including zero-order TSK fuzzy classifiers [12], extreme learning machines (ELMs) [13] and distance correlation (DC) coefficients [36].

A zero-order TSK fuzzy classifier [12] is adopted to build the hierarchical fuzzy classifier due to its high interpretability. The $k$th fuzzy rule can be described as follows:

$$\begin{aligned}&\text{IF}\quad x_1 \text{ is } A_1^k \wedge x_2 \text{ is } A_2^k \wedge \cdots \wedge x_d \text{ is } A_d^k\\&\text{Then } y^k = f^k(\boldsymbol{x}) = p_0^k,\ k = 1, 2, \ldots, K\end{aligned} \tag{1}$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^T$ is the input vector, $j = 1, 2, \ldots, d$, $A_j^k$ is a fuzzy set in the $j$th input domain for the $k$th fuzzy rule, $K$ is the number of rules and $\wedge$ is a fuzzy conjunction operator. Each fuzzy rule corresponds to the input vector $\boldsymbol{x}$, and maps the fuzzy subset $\boldsymbol{A}^k$ of the input space to the fuzzy set $f^k(\boldsymbol{x})$ of the output space, $\boldsymbol{A}^k = (A_1^k, A_2^k, \ldots, A_d^k)^T$. Let $\mu^k(\boldsymbol{x})$ be the fuzzy membership function corresponding to the fuzzy set $f^k(\boldsymbol{x})$. In Eq. (2), $\mu^k(\boldsymbol{x})$ can be obtained by the membership value corresponding to each dimension under the fuzzy conjunction operator $\wedge$.

$$\mu^k(\boldsymbol{x}) = \mu_1^k(x_1) \wedge \mu_2^k(x_2) \wedge \cdots \wedge \mu_d^k(x_d) \tag{2}$$

in which $\mu_j^k(x_j)$ is the fuzzy membership function and $j = 1, 2, \ldots, d$. In this study, the Gaussian membership function is

**Table 1**
Thirteen meaningful studies from 2013 to 2021.

| Studies | Reference | Feature extraction | Classifier | Accuracy (%) |
|---|---|---|---|---|
| Song et al. (2016) | [18] | Mahalanobis distance, DWT | ELM | 97.53 |
| Peker et al. (2016) | [27] | DTCWT | CVANN | 98.28 |
| Wang et al. (2017) | [28] | STFT, Energy | Random forest | 96.00 |
| Zhang et al. (2017) | [29] | VMD, AR | Random forest | 97.35 |
| Li et al. (2017) | [17] | DWT, EA | SVM | 94.67 |
| Patidar et al. (2017) | [16] | TQWT, Entropy | LS-SVM | 97.75 |
| Wen et al. (2018) | [21] | AE, DCNN | AdaBoost | 95.00 |
| Hassan et al. (2020) | [30] | GEEMDAN, Spectral | LPBoost | 97.60 |
| Abiyev et al. (2020) | [20] | CNN | Softmax | 98.67 |
| Akyol (2020) | [31] | SEA, DNN | Sigmoid | 97.17 |
| Hong et al. (2021) | [32] | Dictionary learning | DLWH | 99.50 |
| Eltrass et al. (2021) | [33] | QKLMS adaptive filters | – | 97.88 |
| Radman et al. (2021) | [34] | Temporal, Spectral | EDT | 100.0 |

adopted as the fuzzy membership function of the zero-order TSK fuzzy classifier. $\mu_j^k(x_j)$ can be obtained as follows

$$\mu_j^k(x_j) = \exp\left(-\frac{1}{2} \times \left(\frac{(x_j - c_j^k)}{\sigma_j^k}\right)^2\right) \tag{3}$$

where $c_j^k$ and $\sigma_j^k$ can be obtained by the fuzzy $c$-means (FCM) algorithm [7] in Eqs. (4) and (5).

$$c_j^k = \frac{\sum_{i=1}^{N} \mu_{ik} x_{ij}}{\sum_{i=1}^{N} \mu_{ik}} \tag{4}$$

$$\sigma_j^k = \frac{h.\sum_{i=1}^{N} \mu_{ik}(x_{ij} - c_j^k)^2}{\sum_{i=1}^{N} \mu_{ik}} \tag{5}$$

where the fuzzy membership degree $\mu_{ik}$ is obtained by FCM and $i = 1, 2, \ldots, N$. $h$ is a scale parameter that can be manually tuned or optimized by learning strategies, such as cross-validation strategy.

Through the center of gravity defuzzification [12], the whole output of the zero-order TSK fuzzy classifier can be computed as

$$y^0 = \sum_{k=1}^{K} \frac{\mu^k(\mathbf{x}) f^k(\mathbf{x})}{\sum_{k'}^{K} \mu^{k'}(\mathbf{x})} = \sum_{k=1}^{K} \left(\frac{\prod_{j=1}^{d} \mu_j^k(x_j)}{\sum_{k'=1}^{K} \prod_{j=1}^{d} \mu_j^{k'}(x_j)}\right) f^k(\mathbf{x}) \tag{6}$$

According to [37], the terse output of the zero-order TSK fuzzy classifier in Eq. (6) can be written as follows

$$y^0 = (p_0^1, p_0^2, \ldots, p_0^K)[(\tilde{\mu}^1(\mathbf{x})\mathbf{x}_e)^T, (\tilde{\mu}^2(\mathbf{x})\mathbf{x}_e)^T, \ldots, (\tilde{\mu}^K(\mathbf{x})\mathbf{x}_e)^T]^T \tag{7}$$

where $\tilde{\mu}^k(\mathbf{x}) = \frac{\prod_{j=1}^{d} \mu_j^k(x_j)}{\sum_{k'=1}^{K} \prod_{j=1}^{d} \mu_j^{k'}(x_j)}$ and $\mathbf{x}_e = (1, \mathbf{x}^T)^T$.

An extreme learning machine (ELM) is a type of single-hidden layer feedforward network [13]. In the ELM model, the input weights $\mathbf{w}_l$ and hidden layer biases $b_l$ are randomly assigned. $l = 1, 2, \ldots, L$, hidden neuron number $L$ is equivalent to rules number $K$. The output weight $\boldsymbol{\beta}$ [13] can be calculated by Eq. (8)

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{8}$$

where $\mathbf{T} = [t_1, t_2, \ldots, t_N]^T$ is the label matrix, and $\mathbf{H}^{\dagger}$ is the Moore–Penrose generalized inverse of matrix $\mathbf{H}$ which can be obtained by Eq. (9)

$$\mathbf{H} = \begin{pmatrix} g(\mathbf{w}_1 \cdot \mathbf{x}_1 + b_1) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1 \cdot \mathbf{x}_N + b_1) & \cdots & g(\mathbf{w}_L \cdot \mathbf{x}_N + b_L) \end{pmatrix}_{N \times L} \tag{9}$$

where $g(.)$ is the activation function.

To overcome the weakness of the Pearson coefficient [38], Székely et al. defined a distance correlation (DC) coefficient that can measure the degree of nonlinear correlation [36]. The non-negative DC coefficient $R_n(\mathbf{X}, \mathbf{Y})$ between $\mathbf{X}$ and $\mathbf{Y}$ is defined [36] as follows:

$$R_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\nu_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\nu_n^2(\mathbf{X}, \mathbf{X})\nu_n^2(\mathbf{Y}, \mathbf{Y})}}, & \nu_n^2(\mathbf{X}, \mathbf{X})\nu_n^2(\mathbf{Y}, \mathbf{Y}) > 0 \\ 0, & \nu_n^2(\mathbf{X}, \mathbf{X})\nu_n^2(\mathbf{Y}, \mathbf{Y}) = 0 \end{cases} \tag{10}$$

where $\mathbf{X} = [x_1, x_2, \ldots, x_n]^T$ and $\mathbf{Y} = [y_1, y_2, \ldots, y_n]^T$. The empirical distance covariance $\nu_n^2(\mathbf{X}, \mathbf{Y})$ [36] is defined by

$$\nu_n^2(\mathbf{X}, \mathbf{Y}) = \left(\hat{S}_1 - \hat{S}_3\right) + \left(\hat{S}_2 - \hat{S}_3\right) \tag{11}$$

$\hat{S}_1$, $\hat{S}_2$ and $\hat{S}_3$ [36] can be obtained as follows:

$$\hat{S}_1 = \frac{1}{n^2} \sum_{i}^{n} \sum_{l}^{n} \|x_i - x_l\|_{dx} \|y_i - y_l\|_{dy} \tag{12}$$

$$\hat{S}_2 = \frac{1}{n^2} \sum_{i}^{n} \sum_{l}^{n} \|x_i - x_l\|_{dx} \frac{1}{n^2} \sum_{i}^{n} \sum_{l}^{n} \|y_i - y_l\|_{dy} \tag{13}$$

$$\hat{S}_3 = \frac{1}{n^3} \sum_{i}^{n} \sum_{l}^{n} \sum_{j}^{n} \|x_i - x_l\|_{dx} \left\|y_i - y_j\right\|_{dy} \tag{14}$$

Similarly, $\nu_n^2(\mathbf{X}, \mathbf{X})$ and $\nu_n^2(\mathbf{Y}, \mathbf{Y})$ are defined as above.

## 3. Construction of epileptic EEG recognition algorithm

In the IF-part of the fuzzy rules, a feature activity adjustment mechanism is introduced to improve the ability to handle complex and variable classification tasks. Additionally, in the Then-part of the fuzzy rules, the consequent parameters solved by ELM serve as the initial value of the Adam optimization algorithm for the second approximation fitting. Finally, based on the principle of linear mapping [39], a novel ladder-type learning structure is built to further enhance the capability of approximation. Next, we describe the optimized methods and the novel learning structure.

### 3.1. Activity adjustment mechanism

As shown in Fig. 1, a DC-based feature activity adjustment mechanism (DC-FAM) is introduced to adjust the activity of each feature participating in each rule. Below, we construct an activity matrix ($\mathbf{Q} \in \mathbf{R}^{d \times K}$) based on the random bias matrix ($\mathbf{Z} \in \mathbf{R}^{d \times K}$) and the distance correlation coefficient matrix ($\mathbf{P} \in \mathbf{R}^{d \times K}$) to adjust the activity of each feature in each rule. $\mathbf{Q}$, $\mathbf{Z}$ **and** $\mathbf{P}$ are obtained in Eq. (15)–(17), respectively.

$$\mathbf{Z} = (z_{jk})_{d \times K}, \; z_{jk} \in \left[min\{R_n(\mathbf{x}_j, \mathbf{Y})\}, max\{R_n(\mathbf{x}_j, \mathbf{Y})\}\right] \tag{15}$$

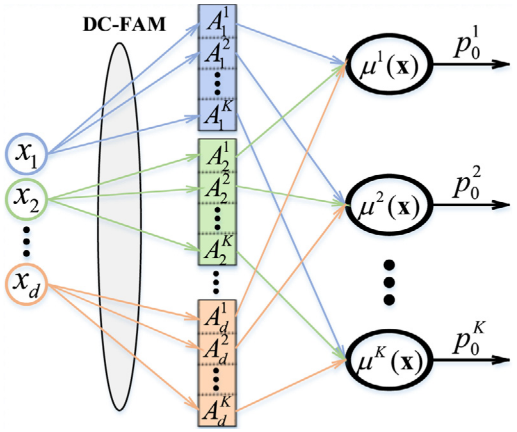


**Fig. 1.** Fuzzy inference framework of TSK fuzzy classifier.

$$\boldsymbol{P} = \begin{bmatrix} R_n(\boldsymbol{x}_1, \boldsymbol{Y}) & R_n(\boldsymbol{x}_1, \boldsymbol{Y}) & \cdots & R_n(\boldsymbol{x}_1, \boldsymbol{Y}) \\ R_n(\boldsymbol{x}_2, \boldsymbol{Y}) & R_n(\boldsymbol{x}_2, \boldsymbol{Y}) & \cdots & R_n(\boldsymbol{x}_2, \boldsymbol{Y}) \\ \vdots & \vdots & \ddots & \vdots \\ R_n(\boldsymbol{x}_d, \boldsymbol{Y}) & R_n(\boldsymbol{x}_d, \boldsymbol{Y}) & \cdots & R_n(\boldsymbol{x}_d, \boldsymbol{Y}) \end{bmatrix} \tag{16}$$

$$\mathbf{Q} = (q_{jk})_{d\times K} = \boldsymbol{Z} \odot \boldsymbol{P} \tag{17}$$

in which each element $z_{jk}$ is randomly generated, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d]$, $\boldsymbol{x}_j \in \boldsymbol{R}^N$, $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, d$, $K$ is the number of fuzzy rules, $d$ is the feature number, $N$ is the number of samples, $\odot$ represents a corresponding element multiplication and the DC coefficient $R_n(\boldsymbol{x}_j, \boldsymbol{Y})$ between $\boldsymbol{x}_j$ and $\boldsymbol{Y}$ is obtained in Eq. (10).

The adjustable threshold $\tau$ is given artificially as an activity criterion. When the activity degree $q_{jk}$ of the $j$th feature in the $k$th rule is higher than $\tau$, the $j$th feature will participate in the inference of the $k$th rule. In contrast, the $j$th feature is discarded by the $k$th fuzzy rule, and the corresponding logical interpretation is discarded. The activity matrix ($\mathbf{Q}$) in Eq. (17) is binarized to facilitate subsequent fuzzy inference in Eq. (18).

$$\Gamma_{jk} = \begin{cases} 1, & q_{jk} \geq \tau \\ 0, & q_{jk} < \tau \end{cases} \tag{18}$$

By reducing the number of features, it can alleviate the risk of the curse of dimensionality for TSK fuzzy classifiers. In some cases, we can also achieve better classification results by discarding redundant, noise-corrupted, or unimportant features [40].

According to the matrix representation of DC-FAM under five fuzzy rules in Fig. 2, the role of DC-FAM can be summarized into the following three points: (1) The distance correlation coefficient $R_n(\boldsymbol{x}_j, \boldsymbol{Y})$ can measure the degree of correlation between features $\boldsymbol{x}_j$ and real tags $\boldsymbol{Y}$. Thanks to the participation of $R_n(\boldsymbol{x}_j, \boldsymbol{Y})$, DC-FAM can eliminate features with a low degree of correlation to reduce the number of features involved in fuzzy inference without deteriorating its classification performance; (2) In a complex and changing environment, the activity matrix (i.e., $\mathbf{Q}$) with certain randomness can expand the difference in the participating features of each rule and improve the classification ability [40]; and (3) Through the adjustable threshold $\tau$, we can adjust the number of discarded features and improve the universality of DC-FAM. It is worth noting that fewer features are discarded to avoid excessive loss of classification performance.

IF DC-FAM is not considered, then output value $h_{ik}$ is equal to $\prod_{j=1}^{d} u(k, x_{ij})$, where $u(k, x_{ij})$ is the Gaussian membership function. However, when the $j$th feature of the $k$th rule is discarded by DC-FAM, the value of the membership function $u(k, x_{ij})$ will not participate in the fuzzy inference process. Therefore, the optimized output value can be obtained by Eq. (19). Finally, the prediction of the zero-order TSK fuzzy classifier can be computed with Eq. (20).

$$h_{ik} = \prod_{j=1}^{d} u\left(k, x_{ij}\right) = \prod_{j=1}^{d} \left(1 - \Gamma \times (1 - u(k, x_{ij}))\right) \tag{19}$$

$$\boldsymbol{Y} = \boldsymbol{H\beta},\ (\boldsymbol{H\beta})_{ir} = \sum_{k=1}^{K} \beta_{kr} h_{ik} \tag{20}$$

where $r$=1, 2, …, $C$, $i$ = 1, 2, …, $N$, $C$ is the number of classes and $N$ is the number of samples. The output weight $\boldsymbol{\beta}$ is obtained from the optimization scheme in Section 3.2.

### 3.2. ELM-based optimization algorithm

After obtaining the output matrix $\mathbf{H}$ of the rule layer, the least-squares (LS) method [41] can be used to optimize the consequent parameters $\boldsymbol{\beta}$. The objective function can be described as

$$\min_{\beta_{kr}} \quad J(\beta_{kr}) = \frac{1}{2} \sum_{r=1}^{C} \sum_{i=1}^{N} \left( \sum_{k=1}^{K} h_{ik}\beta_{kr} - t_{ir} \right)^2 \tag{21}$$

where $C$ is the number of classes and $\beta_{kr}$ is the consequent parameter of the $k$th rule for the $r$th class. When the $i$th sample is the $r$th class, $t_{ir} = 1$ and $\sum_{r=1}^{C} t_{ir} = 1$. To avoid model overfitting, L2-regularization [42] is introduced to optimize Eq. (21)

$$\min_{\beta_{kr}} \ J(\beta_{kr}) = \frac{1}{2} \sum_{r=1}^{C} \sum_{i=1}^{N} \left( \sum_{k=1}^{K} h_{ik}\beta_{kr} - t_{ir} \right)^2 + \frac{\lambda}{2} \sum_{r=1}^{C} \sum_{k=1}^{K} (\beta_{kr})^2 \tag{22}$$

where $\lambda$ is a regularization parameter. Eq. (23) is the matrix expression of Eq. (22) [42]. Besides, the gradient calculation of the objective function $J\,(\boldsymbol{\beta})$ can be seen in Eq. (24), where $\boldsymbol{I}$ is the identity matrix.

$$\min_{\boldsymbol{\beta} \in \boldsymbol{R}^{K\times C}} \ J\,(\boldsymbol{\beta}) = \frac{1}{2} \left\| \boldsymbol{H\beta} - \boldsymbol{T} \right\|^2 + \frac{\lambda}{2} \left\| \boldsymbol{\beta} \right\|^2 \tag{23}$$

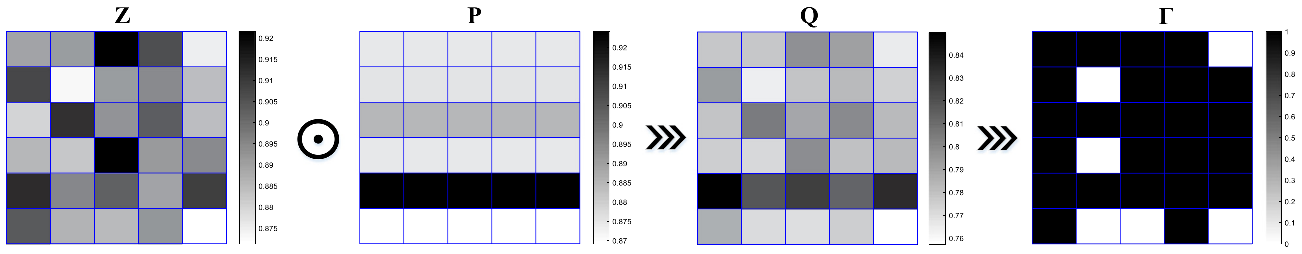


**Fig. 2.** Matrix representation of DC-FAM under five fuzzy rules.

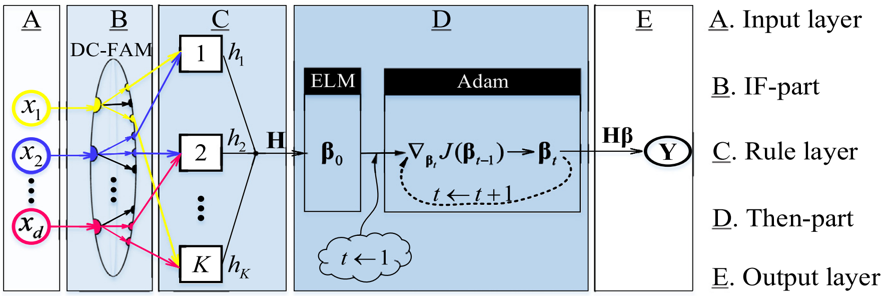


**Fig. 3.** Optimized structure of TSK fuzzy classifier.

$$f(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = \frac{1}{2} \times \left(\nabla_{\boldsymbol{\beta}} \|\boldsymbol{H\beta} - \boldsymbol{T}\|^2 + \lambda \cdot \nabla_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2\right)$$
$$= \boldsymbol{H}^T \boldsymbol{H\beta} + \lambda \boldsymbol{I\beta} - \boldsymbol{H}^T \boldsymbol{T} \qquad (24)$$

```
Initialization: According to the recommended parameter values in [30], we set the exponential decay rate $\gamma_1 = 0.9$ for the 1st moment estimate, the exponential decay rate $\gamma_2 = 0.9999$ for the 2nd moment estimate, $\zeta = 10^{-8}$, learning rate $log_{10}(\alpha) \in [-4, \ldots, -1]$. Initialize the 1st moment vector $\boldsymbol{m}_0$ and 2nd moment vector $\boldsymbol{v}_0$ with 0.
Feedback iteration: Set t=1
  while $\boldsymbol{\beta}_t$ not converged do
      Step1: Get the gradient at timestep $t$
          $f(\boldsymbol{\beta}_t) \leftarrow \nabla_{\boldsymbol{\beta}_t} J(\boldsymbol{\beta}_{t-1})$
      Step2: Update biased 1st moment estimate [30]
          $\boldsymbol{m}_t \leftarrow \gamma_1 \cdot \boldsymbol{m}_{t-1} + (1-\gamma_1) f(\boldsymbol{\beta}_t)$
      Step3: Update biased 2nd raw moment estimate [30]
          $\boldsymbol{v}_t \leftarrow \gamma_2 \cdot \boldsymbol{v}_{t-1} + (1-\gamma_2) \cdot (f(\boldsymbol{\beta}_t) \odot f(\boldsymbol{\beta}_t))$
      Step4: Adjust learning rate
          $\alpha_t \leftarrow \frac{\alpha \cdot \sqrt{1-(\gamma_2)^t}}{(1-(\gamma_1)^t)}$
      Step5: Update consequent parameters $\boldsymbol{\beta}_t$
          $\boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}_{t-1} - \frac{\alpha_t \cdot \boldsymbol{m}_t}{(\sqrt{\boldsymbol{v}_t} + \zeta)}$
  end while
Return $\boldsymbol{\beta}_t$
```

Based on the gradient in Eq. (24), an adaptive moment estimation (Adam) optimization algorithm [14] is adopted to obtain satisfactory consequent parameters. In 2015, Kingma and Ba proposed the Adam algorithm based on adaptive low-order moment estimation [14]. This algorithm designs independent adaptive learning rates for different parameters by calculating gradient first-order moment estimation and second-order moment estimation. Adam's implementation process is introduced above [14].

The initial value $\boldsymbol{\beta}_0$ has an important influence on the entire convergence process and the final convergence performance. This study adopts ELM to obtain $\boldsymbol{\beta}_0$, which makes the gradient considerably small at the initial iteration. More details about the derivation of the ELM can be seen in Section 2. Eq. (25) gives the final solution equation of the ELM [13].

$$\boldsymbol{\beta}_0 = \boldsymbol{H}^{\dagger} \boldsymbol{T} \qquad (25)$$

ELM can produce a small training error while ensuring the minimum norm of weights [13]. Bartlett stated on the generalization of a forward neural network that when the training error of the neural network is small, the smaller weighted norm will yield better generalization performance of the network [43].

This optimized structure of the TSK fuzzy classifier is shown in Fig. 3. In general, the ELM-based Adam algorithm for solving consequent parameters has the following four advantages: (1) Under the initial parameters of Eq. (25) with low complexity, the Adam algorithm only requires a few iterations to complete the convergence, and this performance can be proven in Fig. 6. (2) The high generalization performance of ELM can be absorbed by D-LT-TSK. (3) The Adam algorithm be capable of solving noisy and nonstationary targets [14] as in the case of the epileptic EEG signal studied in this paper. (4) The Adam algorithm can calculate the adaptive learning rate $\alpha_t$ through the first-order moment mean, and the initial troubles of $\boldsymbol{\beta}_0$ are solved by ELM. These optimizations can reduce the parameter burden of our proposed classifier.

### 3.3. Depth-ladder-type learning structure

Based on the linear mapping principle [39] and the stacked generalization principle [44], we design a ladder-type structure that is built with multiple zero-order TSK fuzzy classifiers.

The learning structure of the D-LT-TSK is illustrated in Fig. 4. In each learning module, these $M_{dp}$ zero-order TSK fuzzy classifiers are trained simultaneously under the same input space, $dp = 1, 2, \ldots, DP$. An optimal strategy is adopted to retain the classifier with the most satisfactory classification performance. Since the performance of the entire learning module depends on multiple zero-order TSK fuzzy classifiers, it can effectively reduce the threat of individual inferior TSK fuzzy classifiers on the classification performance and the stability of the entire classifier. The ladder-type structure is built with horizontal progressive and longitudinal leapfrogging learning, which can be divided into two parts: horizontal progressive learning style and longitudinal leapfrogging learning style.

**Horizontal progressive learning style:** Within each layer, the left learning module is first trained on the training set $\boldsymbol{X}_{dp}^{left}$ to obtain the prediction result $\boldsymbol{Y}_{dp}^{left}$. Then the linear mapping matrix $\mathbf{A}$ representation of $\boldsymbol{Y}_{dp}^{left}$ obtained from the Eq. (28) is generalized to the original input space $\boldsymbol{X}$ as the input space of the right learning module (i.e., $\boldsymbol{X}_{dp}^{right}$). Subsequently, we extract the prediction label $\boldsymbol{t} \in \boldsymbol{R}^N$ from $\boldsymbol{Y}_{dp}^{left} \in \boldsymbol{R}^{N \times C}$ and artificially give a linear space vector $\boldsymbol{\eta} \in \boldsymbol{R}^d$. Suppose that $V_N = V_N(F)$ and $V_d = V_d(F)$ are linear space with dimensions $N$ and $d$ on domain $F$, respectively. The mapping from linear space $V_N$ to $V_d$ can be expressed as:

$$\varphi : V_N \to V_d \qquad (26)$$

Suppose $\mathbf{t}$ and $\boldsymbol{\eta}$ are a vector basis of $V_N$ and $V_d$ respectively. The linear mapping $\varphi$ based on $\mathbf{t}$ and $\boldsymbol{\eta}$ can be expressed as Eq. (27), in which $\mathbf{t} = (t_1, t_2, \ldots, t_N)$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_d)$. $\mathbf{A}$ obtained in Eq. (28) is the matrix representation of the linear

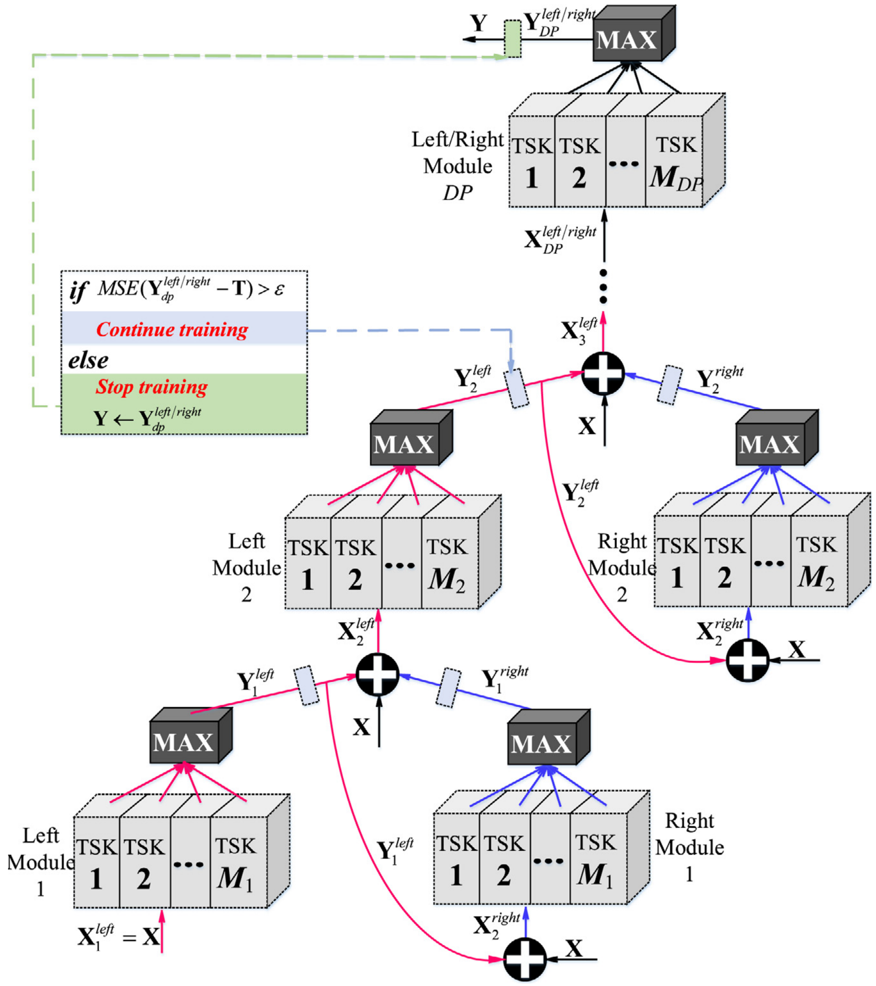


**Fig. 4.** Depth-ladder-type learning structure of D-LT-TSK.

mapping $\varphi$ under the bases $\mathbf{t}$ and $\boldsymbol{\eta}$. After the vector basis of the spaces $V_N$ and $V_d$ are specified (i.e., $\mathbf{t}$ and $\boldsymbol{\eta}$), matrix $\mathbf{A}$ is unique.

$$\varphi\left(t_i\right) = \sum_{j=1}^{d} a_{ji}\eta_j, \ \ i = 1, 2, \ldots, N \tag{27}$$

$$\begin{aligned}\varphi\left(t_1, t_2, \ldots, t_N\right) &= \left(\sum_{j=1}^{d} a_{j1}\eta_j, \sum_{j=1}^{d} a_{j2}\eta_j, \ldots, \sum_{j=1}^{d} a_{jN}\eta_j\right) \\ &= \left(\eta_1, \eta_2, \ldots, \eta_d\right)\boldsymbol{A}, \quad \mathbf{A} = \left(a_{ji}\right)_{d\times N}\end{aligned} \tag{28}$$

Based on the stacked generalization principle [44], the mapping matrix $\boldsymbol{A}_{dp}^{left}$ (i.e., matrix $\mathbf{A}$) on the left side of the $dp$th layer is stacked into the original input space $\mathbf{X}$ to obtain $\boldsymbol{X}_{dp}^{right}$(i.e., the training data of the learning module on the right side of the $dp$th layer).

$$\boldsymbol{X}_{dp}^{right} = \begin{bmatrix}1 & \vartheta\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \left(\boldsymbol{A}_{dp}^{left}\right)^T\end{bmatrix}^T \tag{29}$$

where $\vartheta$ is a small generalization coefficient. Multiple experiments showed that 0.01 to 0.07 is the suitable optimization interval for $\vartheta$.

**Longitudinal leapfrogging learning style:** After completing the two learning modules in the $dp$th layer, the training set of the left learning module in the $(dp+1)$th layer can be constructed through Eq. (30).

$$\begin{cases} \boldsymbol{X}_1^{left} = \boldsymbol{X} \\ \boldsymbol{X}_{dp+1}^{left} = \begin{bmatrix}1 & \frac{\vartheta}{2} & \frac{\vartheta}{2}\end{bmatrix}\begin{bmatrix}\boldsymbol{X} & \left(\boldsymbol{A}_{dp}^{left}\right)^T & \left(\boldsymbol{A}_{dp}^{right}\right)^T\end{bmatrix}^T, \\ dp = 1, 2, \ldots, DP \end{cases} \tag{30}$$

If the mean square error (MSE) between the predicted value $\boldsymbol{Y}_{dp}^{left/right}$ and label vector $\boldsymbol{T}$ is smaller than the preset termination threshold $\varepsilon$ (i.e., $MSE(\boldsymbol{Y}_{dp}^{left/right} - \boldsymbol{T}) < \varepsilon$), then the training process is terminated and the predicted value is used as the whole classifier for output (i.e., $\boldsymbol{Y} \leftarrow \boldsymbol{Y}_{dp}^{left/right}$). $\boldsymbol{Y}_{dp}^{left/right}$ represents the predicted value of either the left learning module or the right learning module.

The advantages of constructing the training set in Eqs. (29) and (30) can be summarized as follows. (1) Mapping the prediction information represented by matrix $\mathbf{A}$ into the original input space, the learning process of the current module is guided by the effective prediction information of the previous learning modules. Therefore, the classification performance can be improved to a certain extent. (2) Obviously, the training space of each learning module is consistent with the original space $\mathbf{X}$, so that the physical interpretation of the original space can be well translated to each training layer. This indicates that D-LT-TSK can enhance model interpretability. (3) The coefficient $\vartheta$ is relatively small, so that the original input space is slightly adjusted rather than largely distorted.

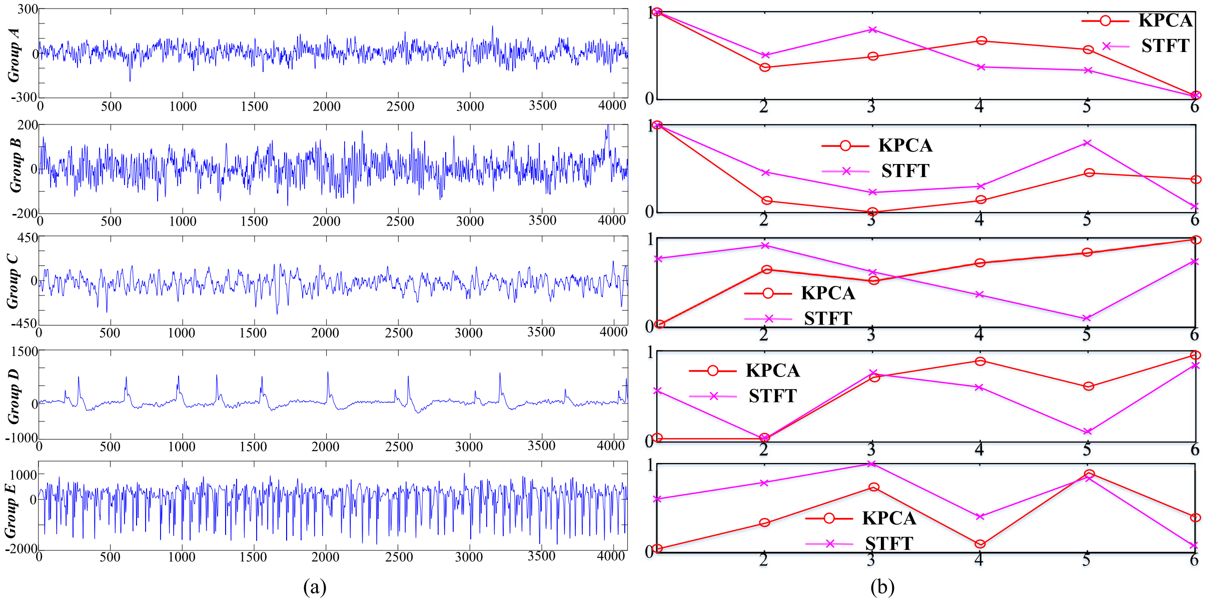


**Fig. 5.** Typical epileptic EEG signals for Groups A to E. (a) Original epileptic EEG signals for groups A to E. (b) Extracted features using KPCA and STFT methods for Groups A to E.

From Eq. (29), we observe that the effective prediction of the left learning module facilitates the learning process of the right module. In other words, horizontal progressive learning is the process of guiding the learning of the right learning module by the left learning module. Additionally, the left and right modules' prediction results obtained by the horizontal progressive learning are stacked together to the next layer in Eq. (30). That is, the valuable prediction information from the $dp$th training layer co-guides the left learning module of the $(dp + 1)$th layer. More importantly, the classification performance of horizontal progressive learning guided by a single prediction information is relatively weaker than that of the longitudinal leapfrogging learning co-guided by two prediction information. Therefore, alternating training between horizontal progressive learning and longitudinal leapfrogging learning helps D-LT-TSK terminate training at an appropriate fitting degree and reduces the risk of overfitting caused by single longitudinal leapfrogging learning.

## 4. Experiments and results

In this section, we apply the D-LT-TSK to an epileptic EEG classification task and verify the superiority of our model in epileptic EEG classification. Additionally, the high interpretability of the D-LT-TSK is demonstrated by the concise interpretation of each fuzzy rule. The epileptic EEG signals [35] adopted in this study can be downloaded from https://www.ukbonn.de/epileptologie. They consist of five groups: *A*, *B*, *C*, *D*, and *E*. *Group A* and *Group B* were collected from healthy volunteers, while *Group C*, *Group D*, and *Group E* were obtained from epileptic volunteers. Each group contains 100 single-channel EEG segments. Each segment is sampled continuously for 23.6 s. The sampling rate is 173.6 Hz [35]. More details can be found on the above-mentioned website. Details on feature extraction and setup are presented in Sections 4.1 and 4.2. Additionally, a verification of the ELM-based Adam algorithm, comparative analysis, parameter sensitivity analysis and demonstration of the interpretability of the proposed method are presented in Sections 4.3, 4.4, 4.5 and 4.6, respectively.

### 4.1. Feature extraction

We use a time-domain feature extraction method (i.e., kernel principal component analysis (KPCA) [45]) and a frequency-domain method (i.e., short-time Fourier transform (STFT) [46]) to extract valuable features from the original EEG signals. Fig. 5 shows the first original epileptic signal from the five groups (i.e., From *Group A* to *Group E*) and the corresponding features extracted by KPCA and STFT. Besides, we provide details of the two feature extraction methods used in the experiment, as shown below.

(1) KPCA is a non-linear algorithm that extends the traditional PCA [47] algorithm using a kernel method, which can realize complex non-linear mapping and mine the non-linear information contained in the dataset. In this study, Gauss kernel function $K(x, y) = e^{-\|x-y\|/\sigma^2}$ are used as the mapping kernels in KPCA. In this experiment, we take the first six features as the top six eigenvalues. Fig. 5(b) shows the six feature values obtained by the KPCA.

(2) STFT is a method that uses a fixed-length window function to intercept time-domain signals and adopts Fourier transform to analyze the intercepted signals. By translating the window function on the time axis, a series of local spectra in each time period can be obtained effectively. In this study, we calculate the spectrogram with a 1s Hamming window for every 0.5s. The steps of feature extraction using STFT are as follows: (1) The original EEG signal is divided into different local stationary signal segments by STFT. (2) Fourier transform is used to analyze the time-varying characteristics of the signal. Subsequently, we can obtain the corresponding local signal spectrum. (3) The energy of the epileptic EEG signals is divided into six bands, i.e., $\delta_1$ 0–2 Hz, $\delta_2$ 2–4 Hz, $\theta$ 4–8 Hz, $\alpha$ 8–15 Hz, $\beta$ 15–30 Hz, and $\gamma$ 30–60 Hz) by a Hamming window.

### 4.2. Setup

Considering the integrity and rigor of the experiment, three combinations of groups (i.e., AE, ABCD, and ABCDE) are adopted to construct training sets with sample sizes of 200 for group AE, 400

**Table 2**
Hyperparameter settings for six classifiers.

| Classifiers | Parameter Descriptions | Hyperparameter Settings |
|---|---|---|
| Zero/first-order<br>TSK fuzzy classifier | ● The number of clusters $\kappa$<br>● The scale parameter $\vartheta$<br>● The regularization parameter $\gamma$<br>● The number of fuzzy rules in the six datasets $R_1$, $R_2$, $R_3$, $R_4$, $R_5$ and $R_6$ | ● $\kappa = R$<br>● $\vartheta$ is adjusted from 0.01 to 100 in the steps of 0.01<br>● $\gamma$ is adjusted from 0.01 to 100 in the steps of 0.01<br>● $R_1$, $R_2$, $R_3$, $R_4$, $R_5$ and $R_6$ are optimized by the CV strategy with an interval of 1, and the ranges of optimization are [8,14], [11,19], [16,25], [7,15], [11,19] and [16,25], respectively. |
| LIBSVM (Linear) | ● The kernel width $\upsilon$<br>● Other parameters adopt default values | ● $\upsilon$ is selected from $\{10^{-1}$, 1.5, 10, 20, 50, 100, 150$\}$<br>● – |
| ELM | ● The number of hidden neurons in six datasets $L_1$, $L_2$, $L_3$, $L_4$, $L_5$ and $L_6$ | ● $L_1$, $L_2$, $L_3$, $L_4$, $L_5$ and $L_6$ are optimized by the CV strategy with an interval of 1, and the ranges of optimization are [7,12], [10,17], [16,22], [7,12], [10,17] and [16,22], respectively. |
| DBN | ● weight cost $\xi$, initial momentum $\varsigma$, epsilonw $\rho$, epsilonvb $\iota$, epsilonhb $\varpi$ and final momentum $\varphi$<br>● The hidden nodes in the six datasets $l_1$, $l_2$, $l_3$, $l_4$, $l_5$ and $l_6$<br>● 3-layer restricted Boltzmann machines (RBM) | ● $\xi = 0.0003$, $\varsigma = 0.5$, $\rho = 0.1$, $\iota = 0.1$, $\varpi = 0.1$, $\varphi = 0.9$<br>● $l_1$, $l_2$, $l_3$, $l_4$, $l_5$ and $l_6$ are optimized by the CV strategy with an interval of 1, and the ranges of optimization are [5,15], [7,22], [10,25], [4,12], [8,22] and [10,25], respectively.<br>● – |
| D-LT-TSK | ● The number of fuzzy rules in the six datasets $K_1$, $K_2$, $K_3$, $K_4$, $K_5$ and $K_6$<br>● With 2 or 3 layers; | ● $K_1$, $K_2$, $K_3$, $K_4$, $K_5$ and $K_6$ are optimized with an interval of 1, and the ranges of optimization are [4,8], [4,10], [8,16], [4,8], [6,13] and [6,15], respectively.<br>● – |

**Table 3**
A table summarizes classification and generalization performance among TSK-ELM, TSK-ELM-Adam in the six EEG datasets (*KPCA-AE*, *KPCA-ABCD*, *KPCA-ABCDE*, *STFT-AE*, *STFT-ABCD*, *STFT-ABCDE*).

| | TSK-ELM | | | TSK-Adam | | | TSK-ELM-Adam (D-LT-TSK) | | |
|---|---|---|---|---|---|---|---|---|---|
| | TrAcc$^1$ (Var$^5$)<br>TrTime$^2$ (Var$^5$) | TeAcc$^3$<br>TeTime$^4$ | Rules$^6$ | TrAcc$^1$ (Var$^5$)<br>TrTime$^2$ (Var$^5$) | TeAcc$^3$<br>TeTime$^4$ | Rules$^6$ | TrAcc$^1$ (Var$^5$)<br>TrTime$^2$ (Var$^5$) | TeAcc$^3$<br>TeTime$^4$ | Rules$^6$ |
| *KPCA-AE* | 87.44(9.7E−4)<br>0.000(0.0E−0) | 83.66<br>0.000 | 5 | 89.33(**7.1E−4**)<br>0.005(6.3E−5) | 85.95<br>0.000 | 5 | **89.44**(1.3E−3)<br>0.000(1.7E−7) | **86.27**<br>0.000 | 5 |
| *KPCA-ABCD* | 90.22(**3.3E−4**)<br>0.000(0.0E−0) | 81.51<br>0.000 | 7 | 90.17(3.5E−4)<br>0.025(3.8E−4) | 80.23<br>0.000 | 7 | **91.33**(6.9E−4)<br>0.001(2.3E−7) | **82.01**<br>0.000 | 7 |
| *KPCA-ABCDE* | 83.53(6.3E−4)<br>0.013(1.0E−3) | 77.78<br>0.000 | 12 | 84.23(4.3E−4)<br>0.042(1.6E−3) | **80.33**<br>0.000 | 12 | **85.77**(**6.1E−5**)<br>0.016(9.7E−4) | 80.25<br>0.000 | 12 |
| *STFT-AE* | 97.51(1.9E−4)<br>0.000(0.0E−0) | 98.37<br>0.000 | 5 | 97.17(2.8E−4)<br>0.017(7.9E−6) | 98.03<br>0.000 | 5 | **98.84**(**5.1E−5**)<br>0.000(1.4E−7) | **98.87**<br>0.000 | 5 |
| *STFT-ABCD* | 94.06(6.4E−4)<br>0.000(0.0E−0) | 93.38<br>0.000 | 8 | 95.28(4.1E−4)<br>0.026(1.6E−4) | 94.56<br>0.000 | 8 | **96.50**(**2.0E−4**)<br>0.001(7.6E−8) | **95.05**<br>0.000 | 8 |
| *STFT-ABCDE* | 90.38(1.3E−3)<br>0.000(0.0E−0) | 91.64<br>0.000 | 11 | 93.50(**1.4E−4**)<br>0.052(7.5E−4) | 91.22<br>0.000 | 11 | **94.24**(1.5E−3)<br>0.001(1.4E−7) | **91.51**<br>0.000 | 11 |

TrAcc$^1$: average Training accuracy, TrTime$^2$: Training time, TeAcc$^3$: Testing accuracy, TeTime$^4$: Testing time, Var$^5$: Variance, Rules$^6$: number of fuzzy rules.

for group ABCD, and 500 for group ABCDE. Through the application of two feature extraction methods, we generate six datasets (i.e., *KPCA-AE, KPCA-ABCD, KPCA-ABCDE, STFT-AE, STFT-ABCD*, and *STFT-ABCDE*) with six features.

D-LT-TSK is a rule-based fuzzy classifier with deep learning specificity. It is imperative to compare and analyze the D-LT-TSK from the perspective of fuzzy classifiers and deep neural networks. Therefore, two typical fuzzy classifiers (i.e., zero-order TSK fuzzy classifier and first-order TSK fuzzy classifier) and one state-of-the-art deep learning classifier (i.e., Deep Belief Network, DBN [48]) are adopted for comparison with the D-LT-TSK. In addition, an integrated software A Library for Support Vector Machines (LIBSVM) [49], which is used for support vector classification and regression, with a linear kernel to form a comparative classifier [i.e., LIBSVM (Linear)], is also adopted for comparison to further reflect the classification performance of our classifier. Apart from this, the classic ELM algorithm, which is involved in the proposed D-LT-TSK classifier, is also incorporated to the list of comparing algorithms. The hyperparameter settings for the six classifiers are given in Table 2.

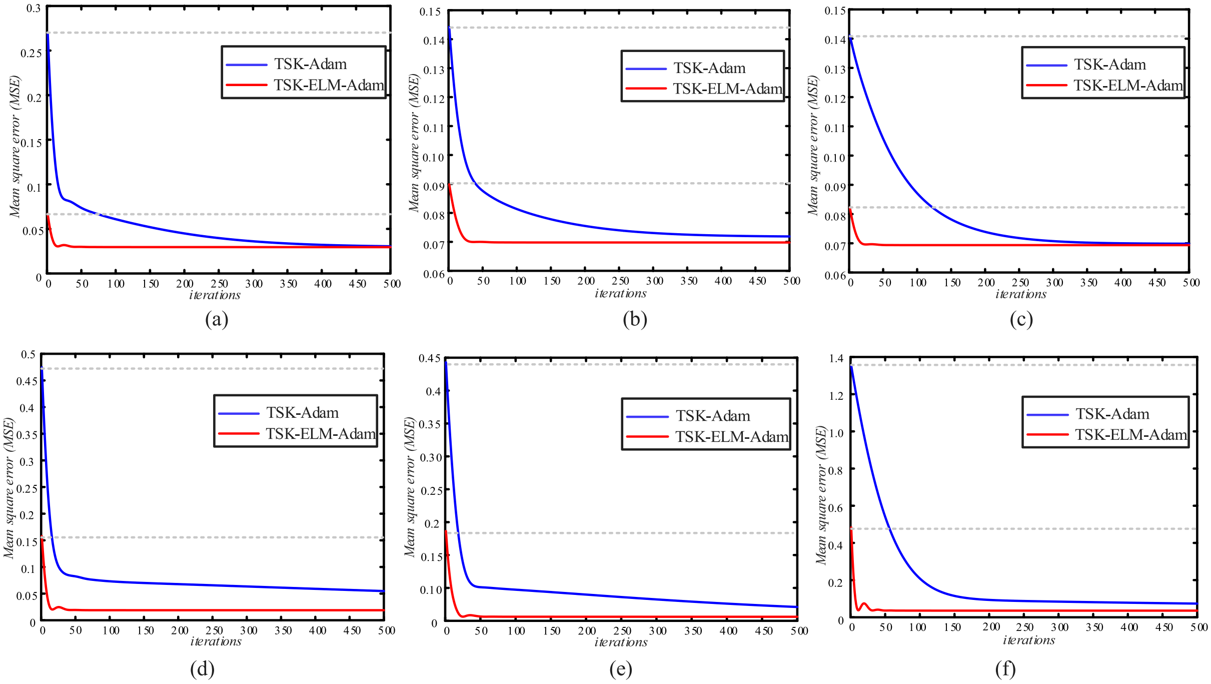
All experiments are carried out on a computer with an Intel Core i7-7700HQ 2.8 GHz processor and 8 GB of RAM. In this study, we randomly stratified 80% of each dataset for training, and 20% for testing.

### 4.3. Verification: ELM-based optimization algorithm

In this subsection, three methods (i.e., ELM, Adam and ELM-based Adam) for solving consequent parameters are taken in the comparative experiments to verify the effectiveness and feasibility of the method proposed in Section 3.2, we denote them as TSK-ELM, TSK-Adam, and TSK-ELM-Adam, respectively.

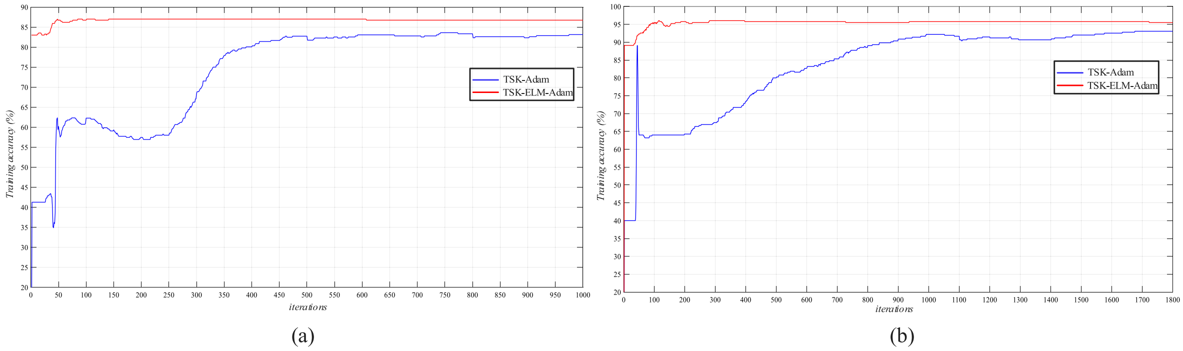
From Table 3, it can be observed that incorporations of both ELM and Adam algorithms into the TSK classifier (i.e., the TSK-ELM-Adam) outperformed the other two comparing methods in majority of the studied EEG datasets. For example, without integration with the Adam algorithm, TSK-ELM provided better performance than the TSK-Adam only in minority of the datasets. By contrast, the TSK-ELM-Adam, after integrating the Adam algorithms into the TSK-ELM, outperformed in majority of the studied dataset, achieving the highest scores compared to the other two comparing methods (Table 3).

Apart from this, Fig. 6 visualizes the impact of number of iterations on the mean square error of the TSK-Adam and TSK-ELM-Adam methods in the six EEG datasets. It can be seen in Fig. 6 that the TSK-ELM-Adam generally resulted in lower mean square error throughout the studied range of iterations for all the datasets. Particularly, the TSK-ELM-Adam be capable of attaining



(a) (b) (c) (d) (e) (f)

**Fig. 6.** Plots visualize the impact of number of iterations on the mean square error (MSE) curves of two gradient descent methods (i.e., TSK-Adam and TSK-ELM-Adam) on the six EEG datasets: (a) *KPCA-AE*; (b) *KPCA-ABCD*; (c) *KPCA-ABCDE*; (d) *STFT-AE*; (e) *STFT-ABCD*; (f) *STFT-ABCDE*.



(a) (b)

**Fig. 7.** Plots demonstrate the influence of iteration steps on the training accuracy of two gradient descent methods (i.e., TSK-Adam and TSK-ELM-Adam) on the two EEG datasets: (a) *KPCA-ABCDE*; (b) *STFT-ABCDE*.

the minimum mean square error in fewer iterations. This capability is of vital importance when it comes to handling large sample data.

Furthermore, two datasets of the largest sample size (i.e., *KPCA-ABCDE* and *STFT-ABCDE*) were employed to demonstrate the influence of iteration steps on the training accuracy of the TSK-Adam and TSK-ELM-Adam classifiers, as shown in Fig. 7. It is indicated that the TSK-ELM-Adam required few iteration steps to reach its maximum classification performance and resulted in superior classification capability throughout the entire range of iteration steps studied, compared with the TSK-Adam.

### 4.4. Comparative analysis

In this subsection, we thoroughly analyze performance of the proposed D-LT-TSK classifier, in terms of accuracy of the training and testing sets, stability of the training process, and computational efficiency, in comparison to the other five comparing algorithms (as listed in Table 2) among the studied datasets.

Fig. 8 illuminates average training accuracy with corresponding variance of the six comparing algorithms in the six EEG datasets. Among the six comparing algorithms, the proposed D-LT-TSK classifier generally yielded the best classification performance in the studied datasets, particularly in *KPCA-AE*, *KPCA-ABCD*, and *KPCA-ABCDE* datasets. Besides, the training stability was, in general, better for the D-LT-TSK classifier than other comparing algorithms in most studied datasets, as reflected by the smaller variance of the training accuracy (Fig. 8).

Fig. 9 displays the average testing accuracy of the six comparing algorithms in the six EEG datasets. It can be observed that the proposed D-LT-TSK classifier outperformed majority of the comparing algorithms in all the six datasets, in terms of the average testing accuracy. This also indicates that the D-LT-TSK classifier presented an outstanding generalization performance between training and testing sets (Figs. 8 and 9).

Fig. 10 illustrates boxplots with beeswarm [50] to further display the average training accuracy of the six comparing classifiers over ten experiments. Notably, the proposed D-LT-TSK yielded the highest average training accuracy concentrated at approximately 88%, in comparison to the other five comparing classifiers. Apart from this, the D-LT-TSK was the second-best stable classifier following the ELM, as reflected by the variance of the training accuracy. It is worth mentioning that the there was

**Table 4**
A table summarizes the efficiency of the six comparing classifiers in each of the six EEG datasets, in terms of training and testing durations, with durations of less than 0.0001 recorded as 0. Complexity of the six classifiers, in aspects of number of rules or hidden neurons (Hns), is also displayed.

| | Zero-order TSK fuzzy classifier | | First-order TSK fuzzy classifier | | LIBSVM (Linear) | | ELM | | DBN | | D-LT-TSK | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | TrTime[1]<br>TeTime[2] | Rules[3] | TrTime[1]<br>TeTime[2] | Rules[3] | TrTime[1]<br>TeTime[2] | Rules[3] | TrTime[1]<br>TeTime[2] | Hns[4] | TrTime[1]<br>TeTime[2] | Hns[4] | TrTime[1]<br>TeTime[2] | Rules[3] |
| *KPCA-AE* | 0.0141<br>**0.0000** | 9 | 0.0359<br>**0.0000** | 8 | 0.0047<br>0.0047 | – | **0.0011**<br>**0.0000** | 9 | 1.1016<br>0.1547 | 9-8-6 | 0.0040<br>**0.0000** | **5-5** |
| *KPCA-ABCD* | **0.0035**<br>**0.0000** | 14 | 0.0338<br>**0.0000** | 15 | 0.0126<br>**0.0000** | – | 0.0188<br>**0.0000** | 15 | 1.3938<br>0.1706 | 16-14-11 | 0.0095<br>0.0003 | **6-5** |
| *KPCA-ABCDE* | **0.0244**<br>**0.0000** | 19 | 0.0737<br>**0.0000** | 17 | 0.0529<br>**0.0000** | – | 0.0265<br>**0.0000** | 18 | 2.6582<br>0.3362 | 19-17-15 | 0.0700<br>**0.0000** | **11-10** |
| *STFT-AE* | 0.1250<br>**0.0000** | 8 | 0.0578<br>0.0094 | 8 | **0.0000**<br>**0.0000** | – | 0.0034<br>**0.0000** | 8 | 0.7163<br>0.0906 | 9-8-7 | 0.0065<br>**0.0000** | **5-5** |
| *STFT-ABCD* | 0.0928<br>0.0031 | 14 | 0.1953<br>0.0031 | 16 | **0.0115**<br>**0.0000** | – | 0.0203<br>**0.0000** | 15 | 1.7063<br>0.1609 | 16-13-11 | 0.0162<br>**0.0000** | **9-8** |
| *STFT-ABCDE* | 0.1291<br>**0.0000** | 19 | 0.2694<br>0.0016 | 17 | 0.0448<br>0.0078 | – | 0.0188<br>0.0094 | 20 | 2.6563<br>0.3406 | 20-18-15 | **0.0106**<br>**0.0000** | **12-10** |

TrTime[1]: Training time, TeTime[2]: Testing time, Rules[3]: number of fuzzy rules, Hns[4]: number of hidden neurons.

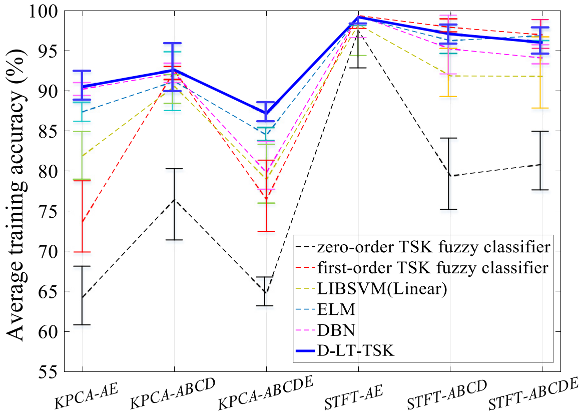


**Fig. 8.** Average training accuracy of the six comparing classifiers in six EEG datasets.

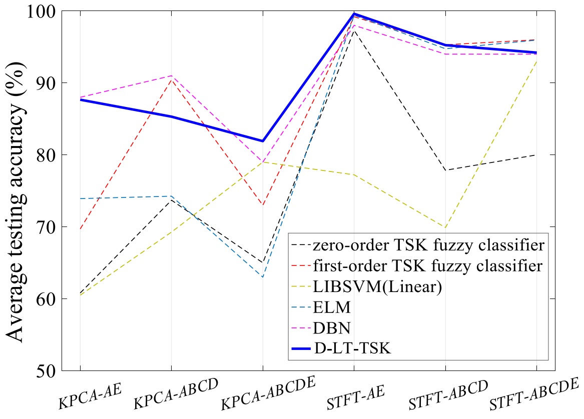


**Fig. 9.** Average testing accuracy of the six comparing classifiers in the six EEG datasets.

an outlier when using the ELM, which may weaken the stability of the ELM classifier. Several experiments showed that the ELM has had individual outliers under the *KPCA-ABCDE* dataset.

Table 4 summarizes the efficiency of the six comparing classifiers in each of the six EEG datasets, in terms of training and testing durations with durations of less than 0.0001s are recorded as 0. Complexity of the six classifiers, in aspects of number of rules or hidden neurons (Hns), is also displayed in Table 4. Firstly, it can be seen that the D-LT-TSK completed the training and testing processes in considerable short period of time in all the datasets, achieving the second-best and top-ranked efficiency in training and testing process, respectively, compared to the other five comparing classifiers (Table 4). This indicates that the D-LT-TSK is capable of fast approximation. Moreover, these results also demonstrate the validity of the proposed approximation structure and show the advantages of the ELM-based Adam method for determining consequent parameters. Secondly, the D-LT-TSK deployed the least number of fuzzy rules among all the comparing classifiers. The number of hidden neurons (i.e., Hns) used by the ELM and DBN is consistently higher than that by the D-LT-TSK (Table 4). Remarkable, this renders higher model interpretability by using the D-LT-TSK, while maintaining satisfactory classification performance.

Automatic recognition of epileptic EEG signals can have two practical applications: a wearable device [51], and a reference clinical diagnosis tool [52]. The proposed D-LT-TSK contains the following particularities to make it better adapt to the above practical applications: (a) The superior learning capability of the D-LT-TSK provides a solid foundation for its practical application (Figs. 8–10); (b) The ELM-based Adam optimization algorithm ensures that the D-LT-TSK offers fast convergence rate, which provides the possibility for achieving real-time epilepsy recognition (Figs. 6–7, Tables 3–4); (c) The feature activity adjustment mechanism (DC-FAM) is capable of expanding the difference in the participating features of each rule and enhance D-LT-TSK's ability to handle complex and variable tasks. Therefore, D-LT-TSK can potentially achieve promising applications for different EEG-based recognition tasks; and (d) The highly interpretable D-LT-TSK is preferable for medical diagnosis applications than the traditional deep learning models that lack model interpretability.

### 4.5. Parameter sensitivity analysis

In this subsection, impacts of five key parameters (i.e., $M$, $\tau$, $K$, $\vartheta$ and $DP$) on the proposed D-LT-TSK classifier are analyzed. $M$, $K$ and $DP$ are the number of TSK fuzzy classifiers in a learning module, fuzzy rules and layers of D-LT-TSK, respectively. Besides, $\tau$ and $\vartheta$ are the activity criterion in Eq. (18) and the small generalization coefficient in Eq. (29), respectively.

The D-LT-TSK is constructed by multiple learning modules through the horizontal progressive learning style and the longitudinal leapfrogging learning style. Three parameters (i.e., $M$, $\tau$ and $K$) that directly influent the performance of learning modules are firstly optimized to ensure outstanding performance of the

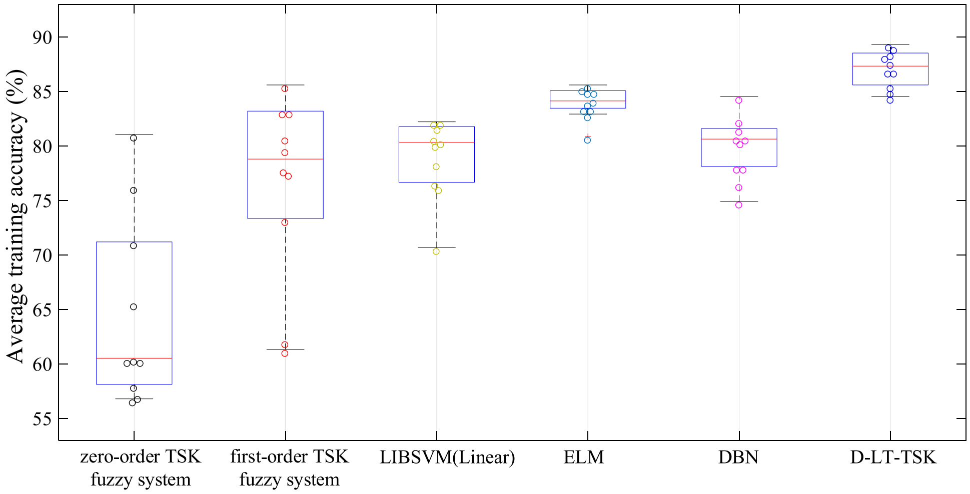
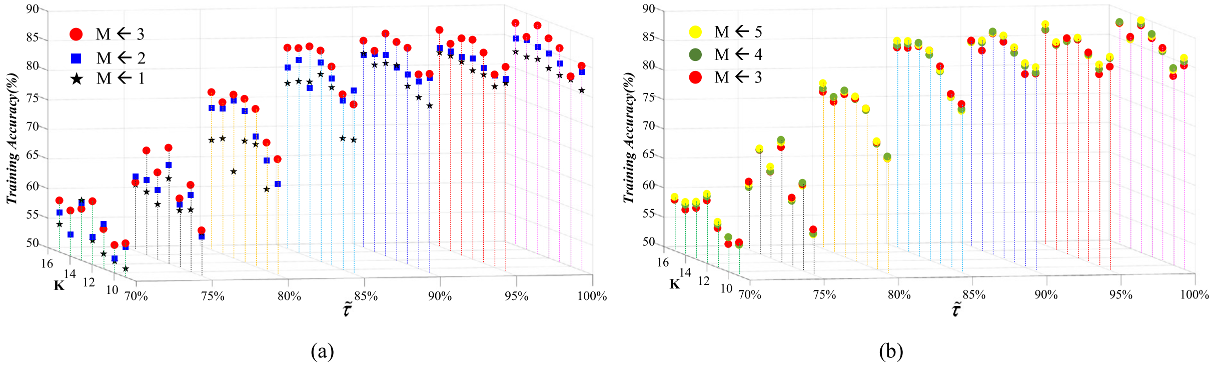


**Fig. 10.** A boxplot illustrates variations of the average training accuracy of the six comparing classifiers across ten experiments on the *KPCA-ABCDE* dataset.



(a)

(b)

**Fig. 11.** Sensitivity law of learning modules integrated by $M$ zero-order TSK fuzzy classifiers under two adjustable parameters (i.e., $K$ and $\tilde{\tau}$) (a) $M \in 1, 2, 3$; (b) $M \in 3, 4, 5$.

single learning module. According to Section 3.1, $\tau$ can be used to control the DC-FAM for adjusting the number of retained features. Therefore, studying the percentage of retained features (i.e., $\tilde{\tau}$) instead of $\tau$ would allow us to better appreciate the law of $\tau$. In this subsection, the *KPCA-ABCDE* dataset, which is the largest dataset using the KPCA method, was employed to demonstrate sensitivity of the D-LT-TSK classifier against the five parameters.

Fig. 11 visualizes the sensitivity law of learning module integrated by $M$ zero-order TSK fuzzy classifiers under two adjustable parameters (i.e., $K$ and $\tilde{\tau}$) (a) $M \in \{1, 2, 3\}$; (b) $M \in \{3, 4, 5\}$. In Fig. 11(a), it can be clearly observed that when $K$ and $\tilde{\tau}$ change in a certain range, the training accuracy is greatly improved in most cases with an increase of $M$ from 1 to 3. While when $M$ is increased from 3 to 5, the training accuracy is only slightly improved, or even locally decreased, as shown in Fig. 11(b). In addition, $M$ is usually taken as 3, which is a trade-off between the classification performance and the interpretability of the classifier in this experiment.

Fig. 12 is derived from projection of the three-dimensional map of Fig. 11 along the $K$-axis direction, the three colored shadows (Dark gray for $M = 1$, Blue for $M = 2$, and Red for $M = 3$) indicate the range of training accuracy (from maximum to minimum) against different settings of $\tilde{\tau}$. Of note, the impact of $\tilde{\tau}$ on the training accuracy indicates a bi-phasic trend for the settings of M (from $M = 1$ to $M = 3$), demonstrating higher training accuracy with increasing value $\tilde{\tau}$ from 70% to 90%, beyond which the accuracy becomes saturated or worsened. Fig. 13 demonstrates the sensitivity law of the learning module integrated by three zero-order TSK fuzzy classifiers. In Fig. 13, the trend of training accuracy with varying $\tilde{\tau}$ under different rule numbers is consistent with Fig. 11. Considering the above analysis, it is recommended for FA-FAM to retain 85% of the features.

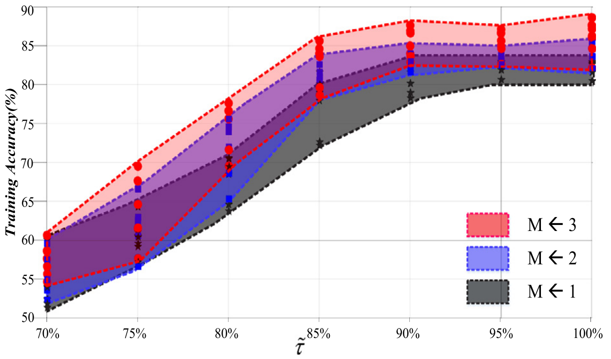


**Fig. 12.** Projection of Fig. 11(a) along the $K$-axis direction with the three colored shadows (Dark gray for $M = 1$, Blue for $M = 2$, and Red for $M = 3$) indicating the range of training accuracy against different setting of $\tilde{\tau}$.

Fig. 14 reflects the following properties: (1) An increase in fuzzy rules $K$ enhances the classification performance at the expense of sacrificing interpretability of D-LT-TSK. After balancing trade-off between classification performance and interpretability, a $K$ value of 12 is recommended (from Fig. 14(a)); (2) The range

**Table 5**
A table summarizes effects of the five studied parameters (i.e., M, $\tau$, $K$, $\vartheta$ and $DP$) on the D-LT-TSK in two phases.

| Parameters | Phase 1 | | Phase 2 | | Recommended value |
|---|---|---|---|---|---|
| | Interval | Trend | Interval | Trend | |
| $M$ | {1, 2, 3} | ↑ | {4, 5} | ⋇ | 3 |
| $\tau$ | {0.70, 0.75, 0.80, 0.85} | ↑ | {0.90, 0.95, 1} | ⋇ | 0.85 |
| $K$ | {8, 9, 10, 11, 12} | ↑ | {13, 14, 15, 16} | ⋇ | 12 |
| $\vartheta$ | {0, 0.005, 0.01, 0.02, 0.03} | ↑ | {0.04, 0.05, 0.06, 0.07} | ↓ | 0.03 |
| $DP$ | {1, 2, 3} | ↑ | {4, 5} | ↓ | 3 |

Note: ↑, ⋇ and ↓ respectively represent three trends: a significant increase, a slight increase or a local decrease, and a significant decrease.

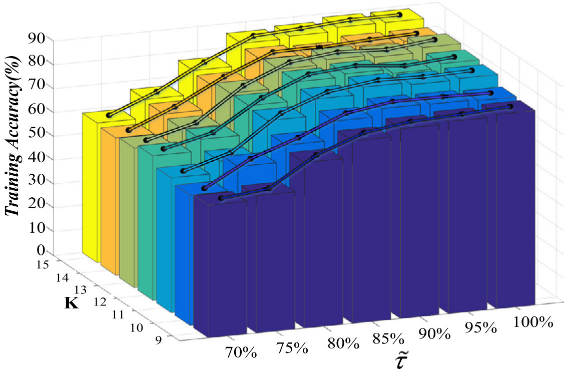


**Fig. 13.** Sensitivity law of learning module integrated by three zero-order TSK fuzzy classifiers.

from 0.01 to 0.03 for $\vartheta$ is recommended (from Fig. 14(b)); (3) Satisfactory classification performance can be obtained by D-LT-TSK after two horizontal progressive learnings and one longitudinal leapfrogging learning (from Fig. 14(c)).

Table 5 summarizes effects of the five studied parameters on the D-LT-TSK in two phases. In the first phase, the classification performance of D-LT-TSK is improved with the increase in the respective parameters. Besides, the effect of $\tau$, $K$, $\vartheta$ and $DP$ on the classifier is more prominent compared to that of parameter $M$. In the second phase, there are two main trends regarding the classification performance. On one hand, the classification performance of D-LT-TSK is slightly improved, or even locally decreased as the parameters $M$, $\tau$ and $K$ are increased. On the other hand, the classification performance is markedly dropped with increasing values of $\vartheta$ and $DP$. In general, recommended values for $M$, $\tau$, $K$, $\vartheta$ and $DP$ are 3, 0.85, 12, 0.03 and 3, respectively, for the *KPCA-ABCDE* dataset.

### 4.6. Interpretability analysis

The interpretability of a classifier is of crucial importance for assisting clinicians in diagnosing the etiology. In this subsection, the high interpretability of the D-LT-TSK is demonstrated clearly through a concise logical interpretation of each fuzzy rule. Five fuzzy rules are adopted to train the *KPCA-AE* dataset; the antecedent and consequent parameters corresponding to each fuzzy rule are summarized in Table 6.

To clearly illustrate model interpretability, we introduce an appropriate logical interpretation method. $c_j^k$ is normalized between 0 and 1. The interval is divided into five equally spaced intervals (i.e., [0.00, 0.20), [0.20, 0.40), [0.40, 0.60), [0.60, 0.80) and [0.80, 1.00]), and their corresponding logical interpretations are *very low*, *low*, *medium*, *high* and *very high*, respectively. When the center $c_j^k$ falls into one of the intervals, its interpretation is that of the interval. In addition, the logical interpretation of the features filtered by the DC-FAM is labeled "*discarded*".

We enumerate the first fuzzy rules in the form of "IF-Then" as an example. Of note, the interpretation of each fuzzy rule is concise and straightforward, which is conducive to enhancing the interpretability of the D-LT-TSK and promoting its application to other domains of medical classification.

***The first fuzzy rule:***

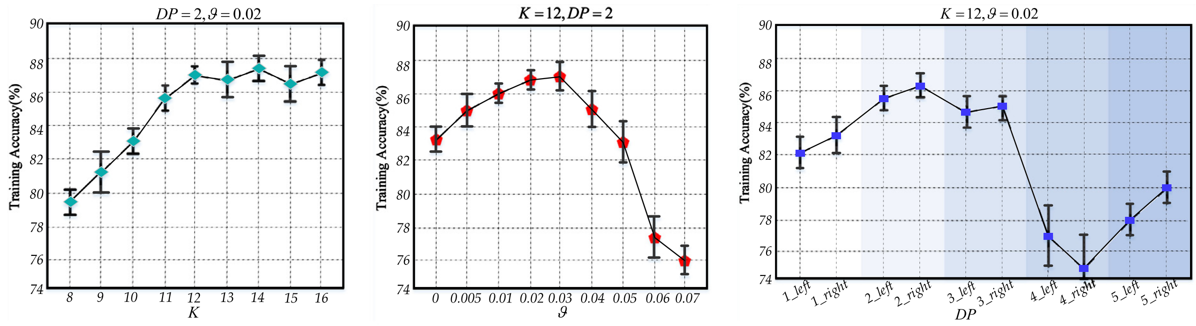
***IF*** *the energy of the EEG signal in Band 1 is* ***low****,*
*and the energy of the EEG signal in Band 2 is* ***discarded****,*
*and the energy of the EEG signal in Band 3 is* ***very high****,*
*and the energy of the EEG signal in Band 4 is* ***very high****,*
*and the energy of the EEG signal in Band 5 is* ***medium****,*
*and the energy of the EEG signal in Band 6 is* ***very low****,*
***Then*** $f^1(\boldsymbol{x}) = 0.589$.

## 5. Conclusions

This study presents a depth ladder-type learning structure that ensures the D-LT-TSK has outstanding classification performance and interpretability. In the IF-part of the fuzzy rules, the DC-FAM is proposed to improve the applicability of the classifier in the face of complex and varying systems by expanding the differences of the participating features in each rule. Besides, the ELM-based Adam algorithm with strong approximation performance and fast convergence is adopted to solve the subsequent parameters in the Then-part of the fuzzy rules. Based on the linear mapping principle and the stacked generalization principle, the depth ladder-type structure is built with concise learning modules by alternately using the horizontal progressive learning and longitudinal leapfrogging learning.

The core advantages of the D-LT-TSK are highlighted in two aspects: (1) The D-LT-TSK only requires fewer fuzzy rules to obtain satisfactory classification performance, and this capability is mainly attributed to the optimization of the TSK fuzzy classifier (i.e., DC-FAM and ELM-based Adam algorithm) and the depth ladder-type learning structure. As such, the superior learning capability is guaranteed along with the concise and high interpretability of each learning module. (2) The deep structure of D-LT-TSK is interpretable compared to traditional deep learning models. In Eqs. (29) and (30), the input space of each learning module is consistent with the original input space. Therefore, the physical interpretation of the original space can be translated to each learning module, implicating that the D-LT-TSK can potentially get rid of the so-called "black box" issue in traditional deep learning models.

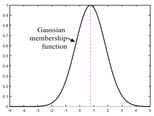
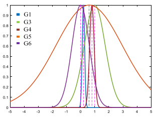
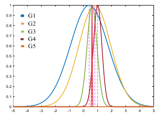
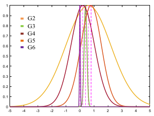
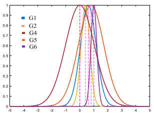
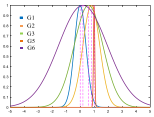
It is worth noting that the optimal values of parameter $\tilde{\tau}$ remains essentially unchanged across the datasets. Hence, the value of $\tilde{\tau}$ depends on characteristics of the dataset itself and is less influenced by other factors. In this study, a certain range is set for parameter $\tilde{\tau}$ to reveal the optimal value, which increases the parameter burden and computational effort of the model. In



(a)　(b)　(c)

**Fig. 14.** Sensitivity law of D-LT-TSK under three settings of adjustable parameters (i.e., $K$, $\vartheta$ and $DP$) (a) $K$; (b) $\vartheta$; (c) $DP$.

**Table 6**
A table summarizes the antecedent and consequent parameters, as well as the corresponding Gaussian membership functions, of the five rules used for D-LT-TSK training on the *KPCA-AE* dataset.

| | | Rules $k$ | Rule *1* | Rule *2* | Rule *3* | Rule *4* | Rule *5* |
|---|---|---|---|---|---|---|---|
| Antecedent parameters | F1 | $e^{-\frac{1}{2}\left(\frac{x-c_1^k}{\sigma_1^k}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.22}{0.09}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.42}{0.96}\right)^2}$ | $q_{13} < \tau$ | $e^{-\frac{1}{2}\left(\frac{x-0.58}{0.56}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.04}{0.42}\right)^2}$ |
| | F2 | $e^{-\frac{1}{2}\left(\frac{x-c_2^k}{\sigma_2^k}\right)^2}$ | $q_{21} < \tau$ | $e^{-\frac{1}{2}\left(\frac{x-0.77}{0.99}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.61}{0.91}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.42}{0.31}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.83}{0.60}\right)^2}$ |
| | F3 | $e^{-\frac{1}{2}\left(\frac{x-c_3^k}{\sigma_3^k}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.98}{0.95}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.64}{0.35}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.38}{0.05}\right)^2}$ | $q_{34} < \tau$ | $e^{-\frac{1}{2}\left(\frac{x-0.58}{0.96}\right)^2}$ |
| | F4 | $e^{-\frac{1}{2}\left(\frac{x-c_4^k}{\sigma_4^k}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.82}{0.18}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-1.00}{0.29}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.21}{0.84}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.00}{0.97}\right)^2}$ | $q_{45} < \tau$ |
| | F5 | $e^{-\frac{1}{2}\left(\frac{x-c_5^k}{\sigma_5^k}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.59}{0.98}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.58}{0.02}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.79}{0.73}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.81}{0.91}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-1.00}{0.05}\right)^2}$ |
| | F6 | $e^{-\frac{1}{2}\left(\frac{x-c_6^k}{\sigma_6^k}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.05}{0.62}\right)^2}$ | $q_{62} < \tau$ | $e^{-\frac{1}{2}\left(\frac{x-0.00}{0.03}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-1.00}{0.21}\right)^2}$ | $e^{-\frac{1}{2}\left(\frac{x-0.22}{0.98}\right)^2}$ |
| Consequent | $p$ | $p_0^k$ | $p_0^1 = 0.589$ | $p_0^2 = 0.115$ | $p_0^3 = 0.448$ | $p_0^4 = 0.195$ | $p_0^5 = 0.859$ |
| Six Gaussian membership functions in each rule | |  |  |  |  |  |  |

future research, we will consider the use of other algorithms such as Principal Component Analysis (PCA) [47] to analyze the dataset for determining the appropriate value of $\tilde{\tau}$.

## CRediT authorship contribution statement

**Wei Xue:** Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Ta Zhou:** Conceptualization, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Jing Cai:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Project administration.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Natural Science Foundation of Jiangsu Province, China [grant number BK20191200]; and the Natural science foundation of universities in Jiangsu, China province [grant number 19KJD520003].

## References

[1] [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/epilepsy.

[2] H. Komijani, M.R. Parsaei, E. Khajeh, M.J. Golkar, Zarrabi. H, EEG classification using recurrent adaptive neuro-fuzzy network based on time-series prediction, Neural Comput. Appl. 31 (2017) 2551–2562, http://dx.doi.org/10.1007/s00521-017-3213-3.

[3] M. Farokhzadi, G.-A. Hossein-Zadeh, H. Soltanian-Zadeh, Nonlinear effective connectivity measure based on adaptive Neuro Fuzzy Inference System and Granger Causality, NeuroImage 181 (2018) 382–394, http://dx.doi.org/10.1016/j.neuroimage.2018.07.024.

[4] U. Ekong, H.K. Lam, B. Xiao, G.X. Ouyang, H.B. Liu, K.Y. Chan, S.H. Ling, Classification of epilepsy seizure phase using interval type-2 fuzzy support vector machine, Neurocomputing 199 (2016) 66–76, http://dx.doi.org/10.1016/j.neucom.2016.03.033.

[5] M. Montazeri-Gh, S. Yazdani, Application of interval type-2 fuzzy logic systems to gas turbine fault diagnosis, Appl. Soft Comput. J. 96 (2020) http://dx.doi.org/10.1016/j.asoc.2020.106703.

[6] Y. Zhang, J. Dong, J. Zhu, C. Wu, Common and special knowledge-driven TSK fuzzy system and its modeling and application for epileptic EEG signals recognition, IEEE Access 7 (2019) 127600-127614, http://dx.doi.org/10.1109/ACCESS.2019.2937657.

[7] Y. Jiang, Z. Deng, F.-L. Chung, G. Wang, P. Qian, K.-S. Choi, S. Wang, Recognition of epileptic EEG signals using a novel MultiView TSK fuzzy system, IEEE Trans. Fuzzy Syst. 25 (1) (2017) 3–20, http://dx.doi.org/10.1109/TFUZZ.2016.2637405.

[8] U.R. Acharya, Y. Hagiwara, H. Adeli, Automated seizure prediction, Epilepsy Behav. 88 (2018) 251–261, http://dx.doi.org/10.1016/%20j.yebeh.2018.09.030.

[9] A. Abdelhaneed, M. Bayoumi, Semi-supervised EEG signals classification system for epileptic seizure detection, IEEE Signal Process. Lett. 26 (12) (2019) 1922–1926, http://dx.doi.org/10.1109/LSP.2019.2953870.

[10] H. Yu, S. Changyin, X. Yang, S. Zheng, H. Zou, Fuzzy support vector machine with relative density information for classifying imbalanced data, IEEE Trans. Fuzzy Syst. (12) (2019) 2353–2367, http://dx.doi.org/10.1109/TFUZZ.2019.2898371.

[11] L.I. Kuncheva, How good are fuzzy if-then classifiers? IEEE Trans. Syst. Man Cybern. B 30 (4) (2000) 501–509, http://dx.doi.org/10.1109/3477.865167.

[12] S. Feng, C.L.P. Chen, Fuzzy broad learning system: A novel neuro-fuzzy model for regression and classification, IEEE Trans. Cybern. 50 (2) (2018) 414–424, https://ieeexplore.ieee.org/document/8432091.

[13] G.-B. Huang, Q.-Y. Zhou, C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomputing 70 (489) (2006) 489–501, http://dx.doi.org/10.1016/j.neucom.2005.12.126.

[14] D.P. Kingma, J.L. Ba, Adam: A Method for Stochastic Optimization, in: Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.

[15] M. Chakraborty, D. Mitra, Automated detection of epileptic seizures using multiscale and refined composite multiscale dispersion entropy, Chaos Solitons Fractals 146 (2021) 110939, http://dx.doi.org/10.1016/j.chaos.2021.110939.

[16] S. Patidar, T. Panigrahi, Detection of epileptic seizure using kraskov entropy applied on tunable-Q wavelet transform of EEG signals, Biomed. Signal Process. Control 34 (2017) 74–80, http://dx.doi.org/10.1016/j.bspc.2017.01.001.

[17] M. Li, W. Chen, T. Zhang, Classification of epilepsy EEG signals using DWT-based envelope analysis and neural network ensemble, Biomed. Signal Process Control 31 (2017) 357–365, http://dx.doi.org/10.1016/j.bspc.2016.09.008.

[18] J.L.T. Song, W. Hu, R. Zhang, Automated detection of epileptic EEGs using a novel fusion feature and extreme learning machine, Neurocomputing 175 (2016) 383–391, http://dx.doi.org/10.1016/j.neucom.2015.10.070.

[19] D. Li, Q. Xie, Q. Jin, et al., A sequential method using multiplicative extreme learning machine for epileptic seizure detection, Neurocomputing 214 (2016) 692–707, http://dx.doi.org/10.1016/j.neucom.2016.06.056.

[20] R. Abiyev, M. Arslan, J. Bush Idoko, et al., Identification of epileptic EEG signals using convolutional neural networks, Appl. Sci. 10 (12) (2020) 4089, http://dx.doi.org/10.3390/app10124089.

[21] T. Wen, Z. Zhang, Deep convolution neural network and autoencoders-based unsupervised feature learning of EEG signals, IEEE Access 6 (2018) 25399–25410, https://ieeexplore.ieee.org/abstract/document/8355473.

[22] Y. Liu, Y.X. Huang, X. Zhang, et al., Deep C-LSTM neural network for epileptic seizure and tumor detection using high-dimension EEG signals, IEEE Access 8 (2020) 37495–37504, https://ieeexplore.ieee.org/abstract/document/9007764.

[23] R. Hussein, H. Palangi, R.K. Ward, et al., Optimized deep neural network architecture for robust detection of epileptic seizures using EEG signals, Clin. Neurophysiol. 130 (1) (2019) 25–37, http://dx.doi.org/10.1016/j.clinph.2018.10.010.

[24] H. RaviPrakash, M. Korostenskaja, E.M. Castillo, et al., Deep learning provides exceptional accuracy to ECoG-based Functional Language Mapping for epilepsy surgery, Front. Neurosci. 14 (2020) 409, http://dx.doi.org/10.3389/%20fnins.2020.00409.

[25] J. Cao, J. Zhu, W. Hu, et al., Epileptic signal classification with deep EEG features by stacked CNNs, IEEE Trans. Cogn. Dev. Syst. 12 (4) (2019) 709–722, https://ieeexplore.ieee.org/abstract/document/8807291.

[26] C. Chatzichristos, J. Dan, A.M. Narayanan, et al., Epileptic seizure detection in EEG via fusion of multi-view attention-gated U-net deep neural networks, in: 2020 IEEE Signal Process Med and Biol Symp (SPMB), 2020, pp. 1–7, https://ieeexplore.ieee.org/abstract/document/9353630.

[27] M. Peker, B. Sen, D. Delen, A novel method for automated diagnosis of epilepsy using complex-valued classifiers, IEEE J. Biomed. Health Inform. 20 (1) (2015) 108–118, https://ieeexplore.ieee.org/abstract/document/7001559.

[28] G. Wang, Z. Deng, K.S. Choi, Detection of epilepsy with electroencephalogram using rule-based classifiers, Neurocomputing 228 (2017) 283–290, http://dx.doi.org/10.1016/j.neucom.2016.09.080.

[29] T. Zhang, W. Chen, M. Li, AR based quadratic feature extraction in the VMD domain for the automated seizure detection of EEG using random forest classifier, Biomed. Signal Process. Control 31 (2017) 550–559, http://dx.doi.org/10.1016/j.bspc.2016.10.001.

[30] A.R. Hassan, A. Subasi, Y. Zhang, Epilepsy seizure detection using complete ensemble empirical mode decomposition with adaptive noise, Knowl. Based Syst. 191 (2020) 105333, http://dx.doi.org/10.1016/j.knosys.2019.105333.

[31] K. Akyol, Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection, Expert Syst. Appl. 148 (2020) 113239, http://dx.doi.org/10.1016/j.eswa.2020.113239.

[32] H. Peng, C. Li, J. Chao, et al., A novel automatic classification detection for epileptic seizure based on dictionary learning and sparse representation, Neurocomputing 424 (2021) 179–192, http://dx.doi.org/10.1016/%20j.neucom.2019.12.010.

[33] A.S. Eltrass, M.B. Tayel, A.F. EL-qady, Automatic epileptic seizure detection approach based on multi-stage Quantized Kernel Least Mean Square filters, Biomed. Signal Process. Control 70 (2021) 103031, http://dx.doi.org/10.1016/%20j.bspc.2021.103031.

[34] M. Radman, M. Moradi, A. Chaibakhsh, et al., Multi-feature fusion approach for epileptic seizure detection from EEG signals, IEEE Sens. J. 21 (3) (2020) 3533–3543, https://ieeexplore.ieee.org/abstract/document/9204737.

[35] R.G. Andrzejak, K. Lehnertz, F. Mormann, et al., Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E 64 (6) (2001) 061907, http://dx.doi.org/10.1103/PhysRevE.64.061907.

[36] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, Ann. Statist. 35 (6) (2007) 2769–2794.

[37] Z. Deng, Y. Jiang, F.-L. Chung, H. Ishibuchi, S. Wang, Knowledge-leverage based fuzzy system and its modeling, IEEE Trans. Fuzzy Syst. 21 (4) (2013) 597–609, http://dx.doi.org/10.1109/TFUZZ.2012.2212444.

[38] W.H. Pearson, Estimation of a correlation coefficient from an uncertainty measure, Psychometrika 31 (3) (1966) 421–433.

[39] A. Dmytryshyn, C.M. Da Fonseca T. Rybalkina, Classification of pairs of linear mappings between two vector spaces and between their quotient space and subspace, Linear Algebra Appl. 509 (2016) 228–246, http://dx.doi.org/10.1016/j.laa.2016.07.016.

[40] Hahn-Ming Lee, Chih-Ming Chen, Jyh-Ming Chen, Yu-Lu Jou, An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE Trans. Syst. Man Cybern. B 31 (3) (2001) 426–432.

[41] Y. Ren, J. Yang, L. Zhao, et al., A global weighted least-squares optimization framework for speckle filtering of PolSAR imagery, IEEE Trans. Geosci. Remote Sens. 57 (3) (2019) 1265–1277, http://dx.doi.org/10.1109/TGRS.2018.2865507.

[42] A. Wang, Y. Chen, N. An, J. Yang, L. Li, L. Jiang, Microarray missing value imputation: A regularized local learning method, IEEE/ACM Trans. Comput. Biol. Bioinform. 16 (3) (2019) 980–993, https://ieeexplore.ieee.org/document/8303223.

[43] P.L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inform. Theory 44 (2) (1998) 525–536, http://dx.doi.org/10.1109/18.661502.

[44] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259, http://dx.doi.org/10.1016/S0893-6080(05)80023-1.

[45] J. Yang, A.F. Frangi, J.-Y. Yang, Z. David, J. Zhong, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 230–244, http://dx.doi.org/10.1109/TPAMI.2005.33.

[46] S. Chikkerur, A.N. Cartwright, V. Govindaraju, Fingerprint enhancement using STFT analysis, Pattern Recognit. 40 (1) (2007) 198–211, http://dx.doi.org/10.1016/j.patcog.2006.05.036.

[47] X. Zheng, D. Levine, J. Shen, S.M. Gogarten, C. Laurie, B.S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data, Bioinformatics 28 (24) (2012) 3326–3328, http://dx.doi.org/10.1093/bioinformatics/bts606.

[48] G.E. Hinton, S. Osindero, Y.-W. Teh, A faster learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1544, http://dx.doi.org/10.1162/neco.2006.18.7.1527.

[49] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, 2005, [Online], Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[50] A. Eklund, Beeswarm: The bee swarm plot, an alternative to stripchart, version R package version 0.2.0, 2015, https://cran.r-project.org/web/packages/beeswarm/index.html.

[51] J. Yang, M. Sawan, From seizure detection to smart and fully embedded seizure prediction engine: A review, IEEE Trans. Biomed. Circuits Syst. 14 (5) (2020) 1008–1023, https://ieeexplore.ieee.org/document/9173735.

[52] Abbasi Bardia, Daniel M. Goldenholz, Machine learning applications in epilepsy, Epilepsia 60 (10) (2019) 2037–2047, http://dx.doi.org/10.1111/epi.16333.