2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS) | 979-8-3503-5271-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISPDS62779.2024.10667619

Ensembling Temporal Convolutional Networks for Heart Sound Classification

Zhongyu Luo
School of Electronic Information Engineering
Foshan University
Foshan, China
luozhongyu2799@163.com

Junjie He
School of Electronic Information Engineering
Foshan University
Foshan, China
graversama@outlook.com

Aiguo Wang*
School of Electronic Information Engineering
Foshan University
Foshan, China
wangaiguo2546@163.com

***Abstract*—Accurately classifying heart sound signals is crucial for the detection and diagnosis of cardiovascular diseases. Due to the inherent complexity of heart sounds and variations among subjects, capturing the latent high-order temporal and spatial dependencies in the signals remains challenging. To this end, we design an end-to-end ensembled temporal convolutional networks aimed at enhancing accuracy. Specifically, we first use temporal convolutional networks with different dilation bases to better analyze heart sound signals and to obtain individual classifiers. Afterwards, the classifiers and a combining strategy are jointly optimized with the mixture of experts. Finally, comparative experiments regarding accuracy and F1 on three public datasets are conducted. Results show that the proposed model outperforms its components and majority voting-based ensemble model.**

Keywords—heart sound classification, temporal convolutional network, mixture of experts

I. Introduction

Cardiovascular diseases are among the leading causes of death globally, and thus the design and development of precise detection and diagnostic tools are crucial for healthcare [1]. Traditional auscultation is an effective and convenient way to facilitate the early diagnosis of heart diseases, but it largely depends on the subjective interpretation of heart sounds, which is easily influenced by the individual professional knowledge and experience [2]. Therefore, the use of artificial intelligence techniques to automate the process is of great value and has attracted attention from researchers.

Time-series heart sound signals contain complex temporal and spatial dependencies, and thus how to encode the signals largely determines the performance of a heart sound classifier. Traditional heart sound signal analysis methods often use hand-crafted features and classical machine learning models, and their shallow structures prevents them from effectively learning high-order relationships among raw signals [3]. In contrast, deep learning has the end-to-end capacity to jointly learn features and optimizing a classifier [4]. Several studies have utilized deep learning to develop heart sound classifiers. For example, Ranipa et al. proposed a multimodal attention convolutional neural network with feature-level fusion to learn high-level features from Mel-cepstral domain as well as general frequency domain features. Experimental results show that their model obtained accuracy of 91.54% [5]. Shi et al. explored different long-short term memory (LSTM) networks, including LSTM and bi-directional LSTM, to segment heart sound signals into different physiological stages [6]. They then conducted comparative experiments and results demonstrated the effectiveness of their proposed model. To capture the spatiotemporal dependencies in heart sound signals, researchers have also explored the use of temporal convolutional network (TCN) in designing heart sound classifier [7]. For example, Dissanayake et al. studied the segmentation and anomaly localization of signals using multi-stage stacked temporal convolutional networks. Experimental results indicate that 91.75% accuracy was obtained [8].

Although temporal convolutional networks could capture long-term dependencies, support parallel computing, and obtain satisfactory performance, the choice of dilation base and the use of dilation rates largely determine their performance. Currently, one common solution is to set the values empirically, which lacks flexibility and robustness when applied to different cases. One viable solution is to use multiple dilation bases and integrate them under the ensemble learning framework, where we face the problem of how to combine and optimize multiple temporal convolutional networks jointly. To address this, this study proposes an end-to-end ensembled temporal convolutional networks under the mixture of experts (TCN-MoE) to better capture high-order temporal and spatial dependencies in the raw signals. The main contributions of this study are as follows: (1) TCNs ensembled by the mixture of experts are proposed. We first utilize TCNs with different dilation bases for individual heart sound classifier and then jointly optimize the classifiers and a combining strategy. This enables TCN-MoE to have the end-to-end capability. (2) Extensive comparative experiments are conducted on public datasets in terms of four performance metrics. Results demonstrate that rTCN-MoE outperforms its components and that the use of the mixture of experts performs better than the use of simple majority voting ensemble.

II. The Proposed Heart Sound Classification Model

Fig. 1 presents the proposed heart sound classification model TCN-MoE that takes the temporal convolutional networks as building blocks and utilizes mixture of experts to combine weak classifiers. Specifically, to facilitate feature learning, we first extract Mel-frequency cepstral coefficients (MFCC) as well as its first-order difference (ΔMFCC) and second-order difference (Δ^2MFCC) from the raw heart sound signals. Then, to better capture the rich information of heart sounds, temporal convolutional networks with a trunk branch and a mask branch is utilized, as shown in Fig. 1. The main branch is a temporal convolutional network that consists of a one-dimensional input,

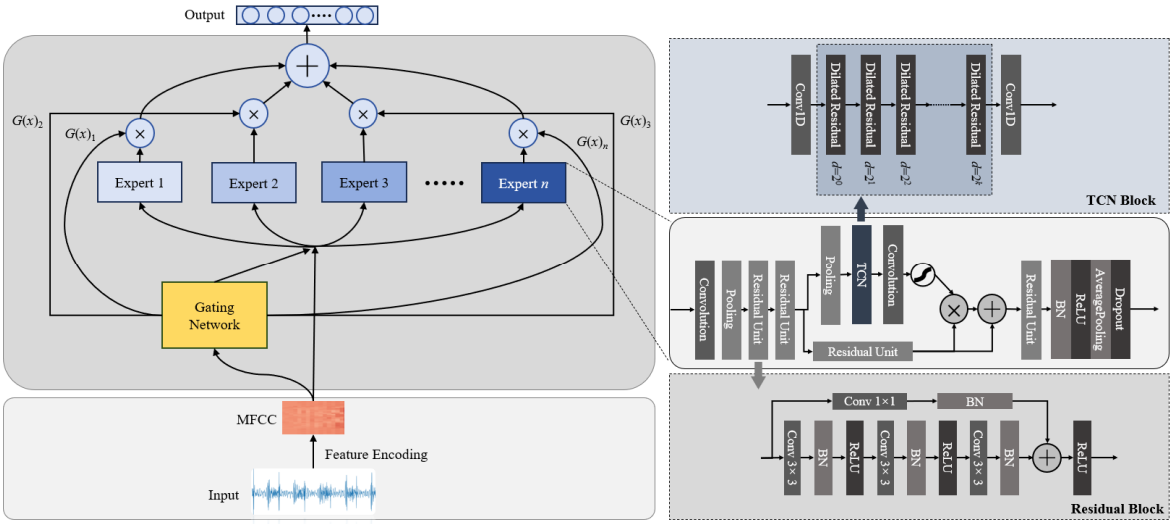


Figure 1. The proposed heart sound classification model TCN-MoE

a one-dimensional convolution output layer, and several dilated convolutional layers with different dilation rates. Suppose the dilation base is d, the dilation rates increase for consecutive layers and take the values of d^1, d^2, … and d^k (k is a predefined number). A residual unit rather than an identify function is used in the mask branch [9].

The use of different dilation bases d would generate different individual heart sound classifiers. Ensemble learning is a common strategy to combine them into a strong classifier. Herein, the mixture of experts (MoE) is then utilized to combine them into a strong classifier [10]. MoE is a machine learning technique that can combine multiple expert models into one larger model, and it aims to improve the accuracy and capability of a prediction system. The MoE has a gating network and n experts (individual model). The former evaluates the importance of each expert and assigns weights to them. Specifically, given a training sample c, assuming the output of the i^{th} expert is $\boldsymbol{o}_i^c$ and its true label is $\boldsymbol{d}_i^c$, the loss E^c is calculated using Eq. (1).

$$E^c = \sum_{i=1}^{n} p_i^c \left\| \boldsymbol{d}^c - \boldsymbol{o}_i^c \right\|^2 \tag{1}$$

where p_i^c is the weight of the i^{th} expert for sample c. For a training set D with m samples, the total loss is:

$$E = \frac{1}{|D|} \sum_{c \in D} E^c \tag{2}$$

Clearly, compared with simple ensemble learning models such as majority voting, the use of MoE helps jointly optimize the classifiers and the combining strategy. Hence, TCN-MoE is an end-to-end learning model that would greatly relieve users from the tedious tasks of feature learning and classifier training.

III. Experimental Setup and Results

A. Experimental Dataset

To evaluate the effectiveness of the proposed model, we conduct experiments on three public available heart sound datasets, including the PhysioNet/CinC Challenge dataset (PCCD), Kaggle Heartbeat Sounds dataset (KHSD), and Yaseen dataset (YSD).

The PhysioNet/CinC Challenge dataset consists of six sub-datasets (labeled A to F) that were collected from healthy individuals and patients with various cardiac conditions (e.g., coronary artery disease and heart valve defects) in non-clinical and clinical settings. There are total 3240 samples, which were recorded at a sampling rate of 2000Hz, with durations ranging from 5 to 120 seconds. The Kaggle Heartbeat Sounds dataset includes heartbeats and metadata that were collected from hospital clinical trials using the DigiScope digital stethoscope and from the public via the iStethoscope Pro iPhone app. The samples were recorded at a sampling rate of 16000 Hz, with durations ranging from 1 to 30 seconds. The Yaseen dataset has 800 abnormal samples and 200 healthy samples, which were recorded at a sampling rate of 8000 Hz, with durations ranging from 1 to 4 seconds. The abnormal samples include 200 aortic stenosis (AS), 200 mitral stenosis (MS), 200 mitral regurgitation (MR), and 200 mitral valve prolapse (MVP) cases.

B. Experimental Setup

A five-fold cross-validation is used to generate independent training and test sets, where each of the five folds serves as a test set and the remaining are training set. The following steps are performed to preprocess the heart sound signals. Frist, a 2s sliding window is used to divide the raw signals into segments. Each segment is processed with a fifth-order Butterworth filter with a frequency range of 25 to 400 Hz with an aim to smooth the signals. Then, MFCC as well as the first-order difference and second-order difference of MFCC are extracted from each segment to form a feature vector. Next, TCN-MoE is trained on the training set and validated on the test data. The above procedures are repeated five times and their average concerning accuracy (acc), precision (prec), recall (rec), and F1 are reported, where F1 is the harmonic mean of precision and recall,

$$F1 = \frac{2 \times precision \times recall}{precision \times recall} \tag{3}$$

To evaluate the performance of TCN-MoE, we compare it with its individual components and the weighted majority voting based ensemble, where the final result is assigned to the corresponding class by using a majority voting mechanism. In our study, we utilize three temporal neural networks with the dilation bases of 1, 2, and 3, respectively. We train the models on a server equipped with one NVIDIA GeForce RTX 4090 GPU and one Intel(R) Core(TM) i7-13700KF 3.42GHz CPU. The Adam optimizer is used to update the network parameters, which are initialized by the Xavier normal initializer. A batch size of 32 and an initial learning rate of 0.001 are empirically set. The learning rate is reduced by a factor of 10. The networks are trained for 30 epochs from scratch in the PyTorch framework. Particularly, we first train individual classifiers and then train the whole network (i.e., the individual temporal network and the gating network).

C. Experimental Results

Besides the dilation base d, the value of k also has an impact on the performance on TCN. We here consider the candidate values of 2, 3, and 4. Table 1 presents the experimental results, where the best results are shown in bold. $\text{TCN}_{s,\,t}$ in the first column indicates that the TCN's dilation base is s and its dilation rates are s^0, s^1,… and s^t. TCN-$\text{MoE}_{s,\,t}$ denotes that it consists of multiple temporal convolutional networks with dilation bases of 0, 1, 2,..., s, respectively, and for each TCN, its dilation rates are s^0, s^1,… and s^t. Similarly, TCN-$\text{MV}_{s,\,t}$ are the results of majority voting. Such a parameter configuration helps investigate the role of dilation base and dilation rate. For better comparison, Fig. 2 plots the results.

From Table 1 and Fig. 2, we can observe that TCNs with different dilation bases obtain mixed classification results across the three datasets. For example, on KHSD, $\text{TCN}_{2,\,2}$ obtains 0.9113 F1 score, compared to the 0.8817 of $\text{TCN}_{1,\,2}$ and 0.9008 of $\text{TCN}_{3,\,2}$; on YSD, the F1 of the three models are 0.9855, 0.9844, and 0.9824, respectively. Second, we can observe that the value of k has an impact on the recognition performance. For example, on KHSD, $\text{TCN}_{1,\,2}$ achieves accuracy of 83.41% compared to the 86.83 % of $\text{TCN}_{1,\,3}$ and 86.70% of $\text{TCN}_{1,\,4}$; $\text{TCN}_{2,\,2}$ achieves accuracy of 94.05% compared to the 94.01% accuracy of $\text{TCN}_{2,\,3}$ and 93.85% accuracy of $\text{TCN}_{2,\,4}$ on PCCD. Third, we can observe that ensemble learning models generally obtains better performance than its individual component. For instance, the accuracy of TCN-$\text{MoE}_{3,\,2}$ and TCN-$\text{MV}_{3,\,2}$ on YSD are 98.80% and 98.30%, respectively, which are higher than those of $\text{TCN}_{1,\,2}$ (97.60%), $\text{TCN}_{2,\,2}$ (97.90%), and $\text{TCN}_{3,\,2}$ (97.75%).

TABLE I. RESULTS OF DIFFERENT HEART SOUND CLASSIFICATION MODELS

	PCCD				KHSD				YSD			
	acc	***prec***	***rec***	***F1***	***acc***	***prec***	***rec***	***F1***	***acc***	***prec***	***rec***	***F1***
$\text{TCN}_{1,\,2}$	93.81%	95.84%	96.09%	0.9596	83.41%	85.81%	91.23%	0.8817	97.60%	98.85%	98.30%	0.9855
$\text{TCN}_{1,\,3}$	93.91%	96.03%	95.74%	0.9589	86.83%	91.02%	89.31%	0.9008	97.38%	98.72%	97.44%	0.9805
$\text{TCN}_{1,\,4}$	93.90%	95.81%	96.24%	0.9602	86.70%	90.10%	90.80%	0.9017	97.30%	98.72%	97.40%	0.9801
$\text{TCN}_{2,\,2}$	94.05%	96.23%	95.99%	0.9611	87.84%	89.45%	93.15%	0.9113	97.90%	98.97%	97.95%	0.9844
$\text{TCN}_{2,\,3}$	94.01%	95.70%	96.51%	0.9610	86.83%	89.19%	92.10%	0.9046	97.33%	98.70%	97.83%	0.9824
$\text{TCN}_{2,\,4}$	93.85%	95.68%	96.31%	0.9599	86.55%	91.20%	89.94%	0.9011	97.10%	98.58%	97.05%	0.9778
$\text{TCN}_{3,\,2}$	93.88%	95.82%	96.21%	0.9601	86.98%	92.05%	88.25%	0.9008	97.75%	98.92%	97.63%	0.9824
$\text{TCN}_{3,\,3}$	93.23%	94.76%	96.50%	0.9562	85.27%	87.32%	92.30%	0.8938	97.10%	98.60%	97.30%	0.9791
$\text{TCN}_{3,\,4}$	93.09%	94.63%	96.45%	0.9553	84.70%	87.23%	91.23%	0.8894	96.90%	98.49%	97.20%	0.9781
TCN-$\text{MV}_{3,\,2}$	94.12%	96.04%	96.29%	0.9617	87.70%	92.09%	89.73%	0.9072	98.30%	99.19%	99.15%	0.9915
TCN-$\text{MV}_{3,\,3}$	93.94%	95.82%	96.29%	0.9605	87.13%	90.64%	90.38%	0.9038	97.70%	98.89%	98.10%	0.9847
TCN-$\text{MV}_{3,\,2}$	93.83%	95.58%	96.41%	0.9599	86.83%	89.19%	92.10%	0.9046	97.25%	98.69%	97.06%	0.9782
TCN-$\text{MoE}_{3,\,2}$	**94.76%**	96.38%	96.79%	**0.9658**	**89.27%**	92.48%	91.46%	**0.9195**	**98.80%**	99.42%	99.40%	**0.9940**
TCN-$\text{MoE}_{3,\,3}$	94.52%	96.32%	96.73%	0.9652	89.13%	92.65%	91.03%	0.9181	98.60%	99.32%	99.30%	0.9930
TCN-$\text{MoE}_{3,\,4}$	94.16%	95.79%	96.62%	0.9620	88.98%	91.55%	92.10%	0.9178	98.38%	99.22%	98.56%	0.9887

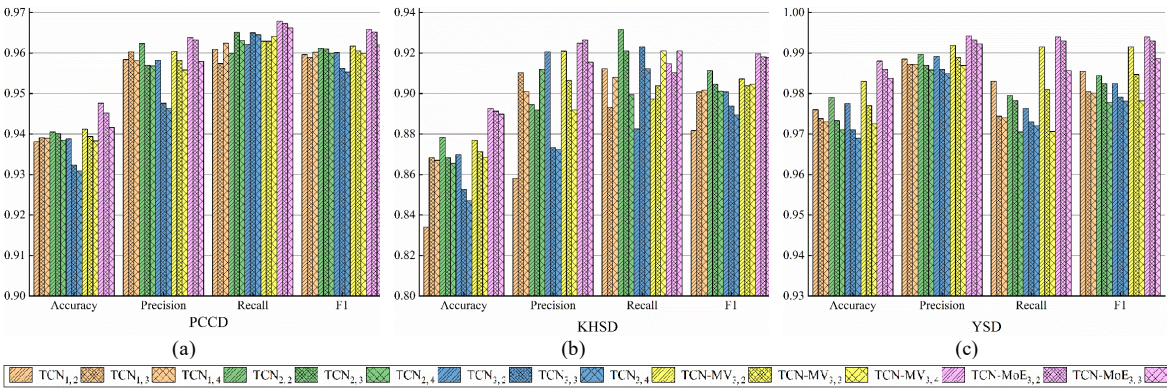
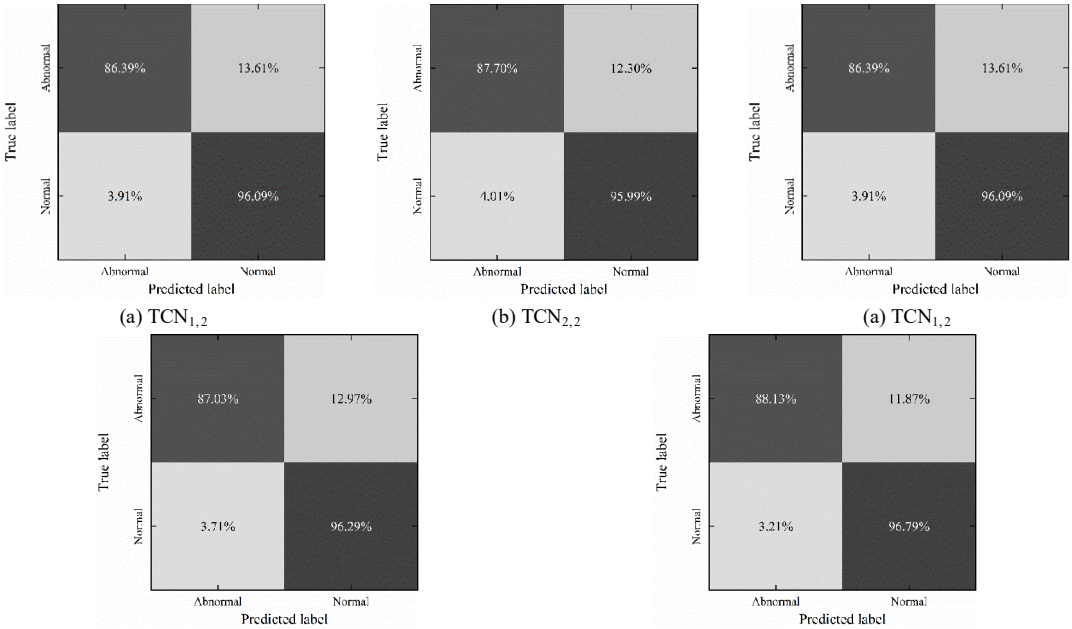


Figure 2. Results of different heart sound classification models on different datasets. (a) PCCD; (b) KHSD; (c) YSD.



(a) $TCN_{1,2}$ (b) $TCN_{2,2}$ (a) $TCN_{1,2}$

(d) $TCN\text{-}MV_{3,2}$ (e) $TCN\text{-}MoE_{3,2}$

Figure 3. Confusion matrix of different heart sound classification models.

This partially demonstrates the effectiveness of ensemble learning techniques. Fourth, we can observe that MoE based ensemble learning generally performs better than majority voting-based TCNs. For example, the accuracy of $TCN\text{-}MoE_{3,2}$, $TCN\text{-}MoE_{3,3}$, and $TCN\text{-}MoE_{3,4}$ are 94.76%, 94.52%, and 94.16%, respectively, which are higher than that of $TCN\text{-}MV_{3,2}$ (94.12%), $TCN\text{-}MV_{3,3}$ (93.94%), and $TCN\text{-}MV_{3,4}$ (93.83%) on PCCD. This is mainly because TCN-MoE jointly optimizes its individual classifiers and combination weights. Furthermore, we present the confusion matrix to evaluate the effectiveness of different heart sound classification models. Due to limited space, we herein only present the results on PCCD. The columns (rows) denote predicted (true) labels. From Fig. 3, we can observe that TCN-MoE generally achieves better accuracy.

IV. Conclusion

To better capture high-level spatial-temporal dependencies embedded in the raw heart sound signals, we in this study explore the use of temporal convolutional networks under the ensemble learning framework to design heart sound analysis models, named TCN-MoE. First, TCNs with different dilation bases are constructed as individual experts. We then use MoE technique to jointly optimize the experts and their combining weights. Finally, we conduct comparative experiments on three publicly available datasets and compare TCN-MoE with its components and the majority voting based model TCN-WV in terms of accuracy, precision, recall, and F1. Results show that the use of ensemble learning helps obtain enhanced accuracy and that TCN-MoE generally outperforms its competitors in the majority of cases. For the future work, we would explore more efficient and lightweight networks such as using deep separable convolution to increase processing speed. Second, it is difficult to collect massive data in medical settings, and one feasible way is to utilize data augmentation techniques. Hence, we would explore the effective technique in handling heart sound signals. Considering that traditional signal processing techniques can extract discriminant features, the combination of deep learning features and traditional features remains another research topic.

References

[1] A.N. Netto, L. Abraham, S. Philip, and H. Care, "HBNET: A blended ensemble model for the detection of cardiovascular anomalies using phonocardiogram," Technology and Health Care, 2024, pp. 1-21.

[2] Y. Zheng, X. Guo, Y. Wang, J. Qin, and F. Lv, "A multi-scale and multi-domain heart sound feature-based machine learning model for ACC/AHA heart failure stage classification, " Physiological Measurement, 2022, vol. 43, no. 6, p. 065002.

[3] K. Qian, Z. Bao, Z. Zhao, T. Koike, F. Dong, M. Schmitt, Q. Dong, J. Shen, W. Jiang, Y. Jiang, and B. Systems, "Learning representations from heart sound: A comparative study on shallow and deep models," Cyborg and Bionic Systems, 2024, vol. 5, p. 0075.

[4] A. Wang, S. Zhao, C. Zheng, J. Yang, G. Chen, and C. Chang, "Activities of daily living recognition with binary environment sensors using deep learning: A comparative study," IEEE Sensors Journal, 2021, vol. 21, no. 4, pp. 5423-5433.

[5] K. Ranipa, W. Zhu, and M.N.S. Swamy, "A novel feature-level fusion scheme with multimodal attention CNN for heart sound classification," Computer Methods and Programs in Biomedicine, 2024, vol. 248, p. 108122.

[6] K. Shi, S. Schellenberger, L. Weber, J. Wiedemann, F. Michler, T. Steigleder, A. Alessa, F. Lurz, C. Ostgathe, R. Weigel, A. Koelpin, "Segmentation of radar-recorded heart sound signals using bidirectional lstm networks," in Proc. EMBS, Berlin, Germany, 2019, pp. 6677-6680.

[7] Y. Yin, K. Ma, and M. Liu, "Temporal convolutional network connected with an anti-arrhythmia hidden semi-Markov model for heart sound segmentation," Applied Sciences, 2020, vol. 10, p. 7049.

[8] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Multi-stage stacked temporal convolution neural networks (MS-S-TCNs) for biosignal segmentation and anomaly localization," Pattern Recognition, 2023, vol. 139, p. 109440.

[9] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in Proc. CVPR, Hawaii, USA, 2017, pp. 3156-3164.

[10] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton, "Adaptive mixtures of local experts," Neural Computation, 1991, vol. 3, no. 1, pp. 79-87.