ICIC Express Letters
Part B: Applications
Volume **8**, Number **6**, June **2017**

ICIC International ©2017 ISSN 2185-2766
pp. 911–918

ROBUST SOUND EVENT RECOGNITION WITH HIERARCHICAL ELM-SAE AND TWO-STAGE ENSEMBLE LEARNING

Jie Yin and Junjie Zhang

School of Communication and Information Engineering
Shanghai University
No. 149, Yanchang Road, Jing'an District, Shanghai 200072, P. R. China
yinjie_shu@shu.edu.cn; zjj@staff.shu.edu.cn

Received December 2016; accepted March 2017

Abstract. *The automatic sound event recognition, which can achieve human-like sound recognition performance on a variety of hearing tasks, has attracted considerable attention. The Spectrogram Image Feature is an effective feature extraction method in SER system. Meanwhile, ELM-AE is a significant feature representation learning algorithm with very high efficiency, but it suffers from non-effective performance on natural signals. In this work, a novel hierarchical ELM-based sparse auto-encoder (H-ELM-SAE) algorithm is proposed to improve the robust and effective feature representation of the original ELM-AE. Hierarchically encoded outputs are projected randomly in each layer and then the each layer fused representations are respectively fed to the ELM classifiers based two-stage ensemble learning (TsEL) algorithm to achieve the decision of the sound signals. The experimental results on the RWCP Sound Scene Database show that the proposed SER framework outperforms the state-of-the-art DNN algorithm, suggesting it is potential for the SER system especially in noisy condition.*
Keywords: Sound event recognition, Spectrogram image feature, Hierarchical ELM-based sparse auto-encoder, Two-stage ensemble learning

1. **Introduction.** Acoustic sound event recognition (SER), which aims at processing the continuous acoustic signal and converting it into symbolic descriptions of the corresponding sound events present at the auditory scene [1,2], is attracting considerable attention in recent years. SER can be utilized in a variety of applications, including context-based indexing and retrieval, unobtrusive monitoring, and acoustic surveillance. Furthermore, the detected events can be used as mid-level-representation in other research areas, e.g., audio context recognition, automatic tagging, and audio segmentation [1-3].

The SER system usually consists of three components, namely signal preprocessing, feature extraction and classification [2]. The commonly extracted features for SER are mostly hand-crafted descriptors, which are at a low semantic level, and also generic for different sound datasets without data-specificity [4]. In contrast to the hand-crafted features, the learning-based feature representation methods have gained their good reputation for SER in recent years, because they are data-specific and robust, and the learned features have a higher semantic level [5].

The typical feature learning methods for SER include bag of words [6], sparse coding [7], exemplar-based coding [8], and deep learning (DL) [9]. Specially, DL has achieved great success in SER and performs superiorly to the commonly used hand-crafted features. McLoughlin et al. [10] proposed to use deep neural network (DNN) classifier for representing the time-frequency features from the stabilized auditory image (SAI) and spectrogram image features (SIF), respectively, for SER. Notably, feature learning by multiple restricted boltzmann machine (RBM) networks is the key point that this DNN classifier can improve feature representation of original time-frequency features. Other DL algorithms, such as deep belief network (DBN), convolutional neural networks (CNN)

[11] and auto-encoder (AE) [12], have also been effectively used for SER. However, it is still time-costing to train a deep network by these DL algorithms for a large-scale dataset.

The extreme learning machine (ELM) is a supervised learning algorithm based on single layer feed-forward neural networks (SLFNs), which offers significant advantages, such as effective performance, least user intervention, real-time learning and ease of implementation [13]. However, the ELM algorithms generally suffer from the problem that the shallow architecture in ELM networks usually results in non-effective performance on the natural signals (e.g., images/videos), even with a large number of hidden nodes.

To this end, the ELM based auto-encoder (ELM-AE) network is proposed in [14] and its variants have also been proposed for different applications [14-16]. Kasun et al. attempt to develop a novel multi-layer learning architecture with ELM-AE simply stacked layer by layer for unsupervised representational learning from large-scale data, which is several orders of magnitude faster than other DL algorithms. Furthermore, Tang et al. [15] develop a novel ELM-based hierarchical learning framework for multilayer perceptron, which achieves more robust and better feature representation and generalization by unsupervised multilayer encoding learning followed by supervised classification. Recently, Tissera and McDonnell [16] present a method for synthesizing deep neural networks using ELMs as a stack of supervised auto-encoders, which enhances classification rates and runtime complexity.

In this study, we propose a feature learning and classification framework for SER with H-ELM-SAE and TsEL, in which H-ELM-SAE algorithm learns the robust feature representations of sound segments layer by layer, and then a two-stage ensemble learning algorithm is used to layer-wise fuse feature representations from H-ELM-SAE and classify sound events taking full advantage of the representations of each hidden layer. The main contributions are threefold: (1) An ELM-SAE algorithm is proposed to achieve the robust feature representation for sound signals; (2) An H-ELM-SAE algorithm is proposed to capture the correlations among multiple ELM-SAE layers for further improving feature representation; (3) A TsEL framework is proposed as a classifier to fuse the decisions of H-ELM-SAE to improve classification performance and robustness.

2. H-ELM-SAE and TsEL Based Robust SER Framework.

As shown in Figure 1, the proposed SER framework consists of three components: feature extraction from sound frame, H-ELM-SAE and TsEL. Firstly, the spectrogram image features (SIF) extraction operation performs on a sound file to generate features for each segmented frame, which have the same length for analysis [10]. The H-ELM-SAE algorithm is then implemented on the SIF features for each frame to learn more effective feature representation, which will generate multiple-layer mixed features. Next, in the first stage of the TsEL component, the features mixed each layer representation and raw input data are first respectively fed to the ELM classifiers to generate probability output values, and then the multiple probability values are fused by the weighted voting based ensemble learning algorithm to achieve the decision of the current frame. Finally, the second stage ensemble learning algorithm is conducted on all the frames belonging to a sound file to fuse their decisions and yield the final classification result for SER. The H-ELM-SAE and TsEL components are introduced in the following sections. Since the SIF feature extraction is not the key point of this work, please refer to [10] for details.

ELM is an effective SLFNs-based learning algorithm with randomly generated hidden nodes as shown in Figure 2(a) [13]. The ELM theory can also be applied to building a multi-layer AE, which performs layer-by-layer unsupervised learning [14,15].

For a training set $\{(x_i, y_i), i = 1, \ldots, n\}$, the input data $\boldsymbol{x}$ is mapped to the ELM random feature space with the network output by

$$f_L(\boldsymbol{x}) = \sum_{i=1}^{L} \beta_i h_i(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})\boldsymbol{\beta} \tag{1}$$

$$h_i(\boldsymbol{x}) = g(\boldsymbol{a}_i \cdot \boldsymbol{x} + b_i) \tag{2}$$

where (a_i, b_i) represents randomly generated input weights and bias of hidden layer of ELM. $g(x)$ is the activation function, and β_i is the output weights of the ELM network.

By adding a regularization term to improve the generalization performance and make the solution more robust, the resulting solution $\boldsymbol{\beta}$ is given by [13]

$$\boldsymbol{\beta} = \left(\boldsymbol{H}^T\boldsymbol{H} + \frac{\boldsymbol{I}}{C}\right)^{-1} \boldsymbol{H}^T\boldsymbol{Y} \tag{3}$$

As a variant of AE, ELM-SAE significantly improves the training speed [14] and also achieves excellent representation performance by building multi-layer networks [15]. Figure 2(b) shows the input data is first transformed into an ELM random feature space, and then a multi-layer unsupervised learning is conducted to achieve high-level feature representation. The equation of the output of the ith hidden layer is written as

$$\boldsymbol{H}_i = g\left(\boldsymbol{H}_{i-1} \cdot \boldsymbol{\beta}\right) \tag{4}$$

Notably, each hidden layer of ELM-SAE works as an independent and separated feature extractor. The H-ELM-SAE training architecture is structurally divided into two separate

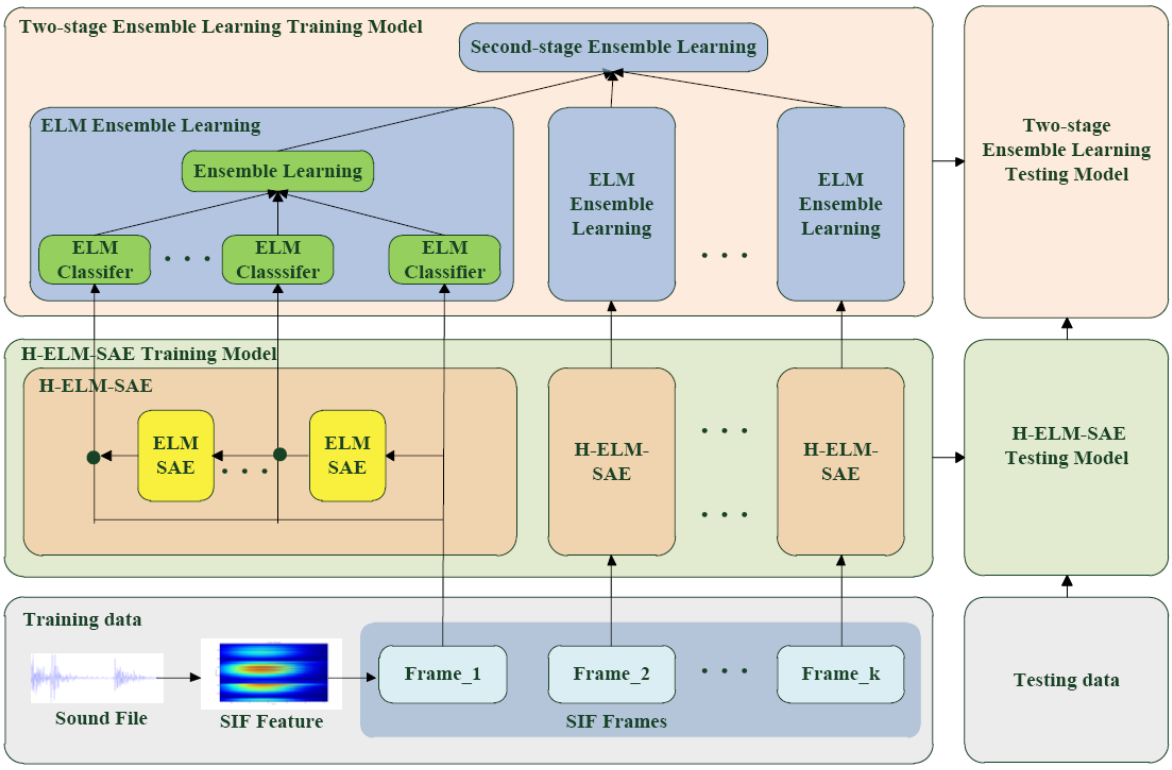

FIGURE 1. Flowchart of the proposed H-ELM-SAE and TsEL framework for SER

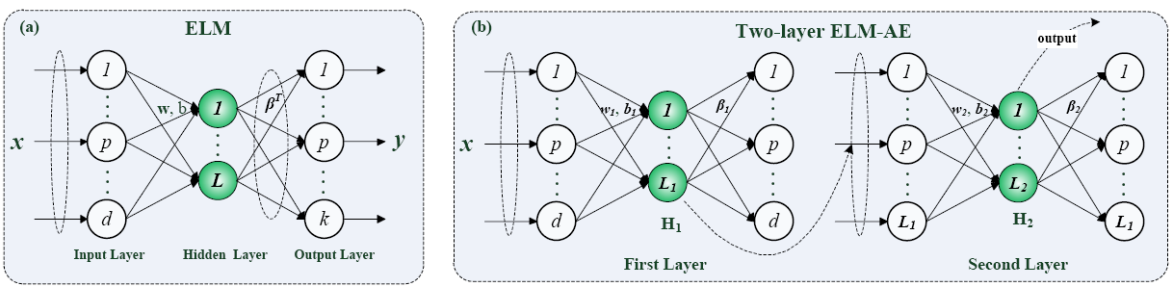

FIGURE 2. The network architecture of ELM and ELM-AE in (a) and (b), respectively

phases in a multilayer manner: 1) unsupervised hierarchical feature representation and 2) supervised feature classification. For the former phase, several ELM-SAEs with ℓ_1 penalty are utilized to extract multilayer sparse concatenated features which are mixed with the input data layer by layer; while for the latter classification, several ELM-based classifiers are conducted on layer-wise mixed features and the first-stage Ensemble Learning of these classifiers is performed for final decision making.

The two-stage ensemble learning (TsEL) framework is proposed for SER on the mixed ELM-SAE features, as shown in Figure 1. Specifically the first-stage ensemble learning [17] is implemented on the base ELM-based classifiers of the layer-wise ELM-SAE features generated from multi-layer ELM-SAE for each sound frame. And then, the second-stage ensemble learning is conducted to fuse the decisions of multiple sound frames from the first stage, and then generates the final classification decision for the current sound. It is worth noting that various ensemble learning algorithms, such as majority voting and weighted voting [10], can be used in this TsEL framework.

Weighted voting based first-stage ensemble algorithm: Suppose that p_{ij}^k and w_{ij}^k $(i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K)$ donate the probability and weight of the ith base classifier on the jth class for kth frame, respectively.

$$w_{ij}^k = \frac{\log\left(p_{ij}^k / 1 - p_{ij}^k\right)}{\sum_{i=1}^{I} \log\left(p_{ij}^k / 1 - p_{ij}^k\right)}, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J \tag{5}$$

The final output of this multi-classifier ensemble learning is given by

$$s_j^k = f_j\left(w_{ij}^k, \hat{Y}_{ij}^k\right) = \sum\nolimits_{i=1}^{I} w_{ij}^k \cdot \hat{Y}_{ij}^k, \quad j = 1, \ldots, J \tag{6}$$

The final classification result of kth frame is decided by

$$label^k = \arg\max_j \left(s_1^k, s_2^k, \ldots, s_J^k\right) \tag{7}$$

Second-stage ensemble algorithm: Based on [10], the final decision of multiple frames corresponding to original sound is decided by the following three ensemble methods:

(1) the baseline method that considers the maximum class score from the mean of all predictive probability values of classifiers as the final decision (denoted as -b):

$$label\_b = \arg\max_j \sum_k s_j^k \tag{8}$$

(2) majority voting based ensemble learning (denoted as -v):

$$label\_v = \arg\max_k \left(\arg\max_j \left(s_j^k\right)\right) \tag{9}$$

(3) weighted-voting based ensemble learning that weighs the votes of individual classifiers by calculating the context energy e^k (denoted as -e):

$$label\_e = \arg\max_j \sum_k e^k \cdot s_j^k \tag{10}$$

3. Experiments and Results.

3.1. Dataset and data preprocessing.

The performance of the proposed H-ELM-TsEL algorithms for SER was evaluated on the Real World Computing Partnership (RWCP) Sound Scene Database in Real Acoustic Environments [18]. The noise-corrupted data use four background noise environments selected from the NOISEX-92 database, namely 'Destroyer Control Room', 'Speech Babble', 'Factory Floor1' and 'Jet Cockpit 1' [4]. McLoughlin et al. have achieved the state-of-the-art performance on this RWCP dataset with the DNN algorithm [10] and have provided the processed SIF feature data to

reproduce the experiments. Thus we directly use this dataset with extracted SIF features in this study. The details about SIF features and data processing can be found in [4,10]. The proposed H-ELM-TsEL algorithm was then implemented to learn feature representation from the extracted SIF features, and the learned features were further fed to the TsEL framework for SER.

3.2. **Experimental settings.** We conducted two same experiments as those in [10] to evaluate our proposed SER framework. In the first mismatched condition experiment, the data in training set were exclusively clean sounds without noise, but the data in testing set were corrupted by additive background noise at levels of 20, 10 and 0dB SNR. The second experiment was the multi-condition evaluation, in which both the data in training set and testing set comprised a variety of clean and noise-corrupted sounds. The 10-fold cross-validation strategy is performed for all algorithms, and the result of classification accuracy is given by the form of mean ± SD (standard deviation).

All the compared algorithms are listed as follows. (a) DNN: the results of DNN by McLoughlin et al. in [10] were selected as the baseline. (b) ELM: the original SIF features of each sound segment were directly fed to the ELM classifier. (c) ELM-AE: the two-layer ELM-AE was implemented on SIF features for each sound segment with the ELM classifier. (d) H-ELM-TsEL: the proposed H-ELM-SAE and TsEL based algorithm was conducted on SIF features as shown in Figure 1. Here the two-layer structure of ELM-SAE was used for the representational learning model and the first-stage ensemble learning of three ELM-based classifiers was performed. It is worth noting that three ensemble methods (-b, -v, -e) were used in the second stage of TsEL.

3.3. **Results of the mismatched condition experiment.** Tables 1 to 3 show the classification results of different algorithms for the first mismatched condition experiment with the -b, -v and -e ensemble learning methods in the second stage of TsEL, respectively.

Table 1. Classification results of different algorithms with the mean probability value ensemble learning (-b) at the second stage of TsEL (Unit: %)

	Clean	20dB	10dB	0dB	Mean
DNN	96.73	94.60	90.27	76.47	89.52
ELM	97.08±0.45	90.31±0.92	85.75±1.16	72.39±1.18	86.38±0.93
ELM-AE	97.68±1.31	94.25±1.53	92.28±2.04	86.63±1.43	92.71±1.58
H-ELM-TsEL	**98.40±0.55**	**96.55±0.74**	**95.95±0.54**	**90.43±2.25**	**95.33±1.02**

Table 2. Classification results of different algorithms with the context voting ensemble learning (-v) at the second stage of TsEL (Unit: %)

	Clean	20dB	10dB	0dB	Mean
DNN	**98.87**	95.33	92.40	78.87	91.37
ELM	93.76±0.61	91.35±0.62	86.99±1.06	72.81±1.29	86.23±0.90
ELM-AE	95.53±2.35	92.28±1.67	89.40±1.70	81.93±1.86	89.78±1.90
H-ELM-TsEL	97.23±0.73	**95.45±0.74**	**94.63±0.76**	**87.50±2.32**	**93.70±1.14**

3.4. **Results of the multi-condition evaluation experiment.** Tables 4 to 5 give the results of different algorithms on the second multi-condition evaluation experiment with the -v and -e ensemble learning methods in the second stage of TsEL, respectively.

TABLE 3. Classification results of different algorithms with e-scaled weight ensemble learning (-e) at the second stage of TsEL (Unit: %)

	Clean	20dB	10dB	0dB	Mean
DNN	96.00	94.37	93.53	85.13	92.26
ELM	95.85±0.55	93.81±0.77	92.82±0.66	87.93±0.86	92.60±0.71
ELM-AE	95.03±2.43	94.43±1.09	93.33±1.82	**90.33±1.43**	93.28±1.69
H-ELM-TsEL	**98.10±0.69**	**96.25±1.00**	**95.63±0.98**	90.08±2.20	**95.01±1.22**

TABLE 4. Multi-condition (MC) classification results of different algorithms with the context voting ensemble learning (-v) at the second stage of TsEL (Unit: %)

	Clean	20dB	10dB	0dB	Mean
DNN	96.90	96.90	93.20	80.40	91.85
ELM	93.34±0.60	91.55±0.57	89.22±0.91	79.78±1.62	88.47±0.93
ELM-AE	94.23±0.76	93.38±1.49	91.65±0.76	86.88±1.60	91.54±1.15
H-ELM-TsEL	**97.13±1.25**	**96.30±1.17**	**94.88±1.55**	**87.85±1.40**	**94.04±1.34**

TABLE 5. Multi-condition (MC) classification results of different algorithms with e-scaled weight ensemble learning (-e) at the second stage of TsEL (Unit: %)

	Clean	20dB	10dB	0dB	Mean
DNN	94.70	95.80	92.10	87.70	92.58
ELM	93.91±0.71	94.34±0.61	93.90±0.86	91.51±0.91	93.42±0.77
ELM-AE	94.50±1.04	94.48±1.04	94.50±0.77	93.68±0.94	94.29±0.95
H-ELM-TsEL	**97.43±0.46**	**97.28±0.61**	**96.93±0.51**	**96.40±0.61**	**97.03±0.55**

4. **Discussion.** In the experiment, three levels of noise are added to clean sound data, namely 20dB, 10dB and 0dB SNR conditions. With the increase of noise, the recognition performance for sound events decreases for all algorithms used in this study. However, in the 0dB SNR condition, the baseline DNN algorithm degenerates most rapidly compared with all the ELM-AE based algorithms as shown in Tables 1 to 5, while the proposed H-ELM-TsEL still achieves good performance with a more than 90% mean accuracy only except the result in the -v ensemble learning method. Therefore, our H-ELM-TsEL algorithm is effective and robust for SER, especially in the noisy environment.

There are three findings from the experimental results on RWCP Sound Scene Database: 1) ELM-AE outperforms the original ELM algorithm, which indicates that ELM-AE model can improve the performance of ELM; 2) The proposed H-ELM-TsEL algorithm by combining the ELM-SAE and TsEL models can further improve the representation performance and robustness of original ELM-AE for sound data especially in high noise; 3) The proposed H-ELM-TsEL framework is superior to the state-of-the-art DNN algorithm in [10] for SER.

In the current H-ELM-TsEL framework, the ELM-SAE features are integrated by a classifier- or decision-level fusion method, that is to say, ensemble learning can be applied to all the classification results of individual ELM-SAE. Specifically, a weighted-voting based ensemble learning algorithm is used in the first stage learning, which has shown its effectiveness. It should be noted that other ensemble learning algorithms, such as the margin distribution optimization method and the Adaboost-based method, also can be applied in this framework. On the other hand, instead of the classifier- or decision-level fusion, another way to properly integrate these ELM-SAE features is the feature-level

fusion. For example, the multiple kernel learning (MKL) method can effectively combine multiple channel features and then make a decision, since the multiple kernels in MKL can naturally correspond to features from different views [19]. This feature-level fusion method will be studied for our H-ELM-TsEL framework in the future.

5. **Conclusions.** This paper has proposed the H-ELM-SAE and two-stage ensemble learning based feature learning and classification framework for robust SER, which classify and synthesize on the layer-wise ELM-SAE representation mixed with original input. The standard evaluation task has also revealed that the proposed H-ELM-SAE formulation on the smoothed and de-noised SIF features achieves excellent classification accuracy and anti-noise robustness, especially for the challenging 0dB SNR noise condition.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (61132004, 61275073, 61420106011) and Shanghai Science and Technology Development Funds (15511105400, 15530500600, 16511104100, 14dz1104800). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange and M. D. Plumbley, Detection and classification of acoustic scenes and events, *IEEE Trans. Multimedia*, vol.17, no.10, pp.1733-1746, 2015.

[2] R. V. Sharan and T. J. Moir, An overview of applications and advancements in automatic sound recognition, *Neurocomputing*, vol.200, pp.22-34, 2016.

[3] T. Lu, G. Y. Wang and F. Su, Context-based environmental audio event recognition for scene understanding, *Multimedia Systems*, vol.21, no.5, pp.507-524, 2015.

[4] J. Dennis, H. Tran and E. S. Chng, Image feature representation of the subband power distribution for robust sound event classification, *IEEE/ACM Trans. Audio Speech Language Processing*, vol.21, no.2, pp.367-377, 2013.

[5] H. Phan, L. Hertel, M. Maass, R. Mazur and A. Mertins, Learning representations for nonspeech audio events through their similarities to speech patterns, *IEEE/ACM Trans. Audio Speech Language Processing*, vol.24, no.4, pp.807-822, 2016.

[6] A. Plinge, R. Grzeszick and G. A. Fink, A bag-of-features approach to acoustic event detection, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp.3704-3708, 2014.

[7] X. Lu, Y. Tsao, S. Matsuda and C. Hori, Sparse representation based on a bag of spectral exemplars for acoustic event detection, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp.6255-6259, 2014.

[8] J. F. Gemmeke, L. Vuegen, B. Vanrumste and H. V. hamme, An exemplar-based NMF approach for audio event detection, *Proc. of the IEEE International Conference on Application of Signal Processing to Audio and Acoustics*, New Paltz, USA, pp.1-4, 2013.

[9] Z. Kons and O. Toledo-Ronen, Audio event classification using deep neural networks, *Proc. of INTERSPEECH*, Tel-Aviv, Israel, pp.1482-1486, 2013.

[10] I. McLoughlin, H. M. Zhang, Z. P. Xie, Y. Song and W. Xiao, Robust sound event classification using deep neural networks, *IEEE/ACM Trans. Audio Speech Language Processing*, vol.23, no.3, pp.540-552, 2015.

[11] H. M. Zhang, I. McLoughlin and Y. Song, Robust sound event recognition using convolutional neural networks, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, pp.559-563, 2015.

[12] E. Marchi, F. Vesperini, F. Eyben, S. Squartini and B. Schuller, A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp.1996-2000, 2014.

[13] G. B. Huang, H. Zhou, X. Ding and R. Zhang, Extreme learning machine for regression and multiclass classification, *IEEE Trans. System Man Cybernetics – Part B: Cybernetics*, vol.42, no.2, pp.513-529, 2012.

[14] L. L. C. Kasun, H. M. Zhou, G. B. Huang and C. M. Vong, Representational learning with extreme learning machine for big data, *IEEE Intelligence System*, vol.28, no.6, pp.31-34, 2013.

[15] J. X. Tang, C. W. Deng and G. B. Huang, Extreme learning machine for multilayer perceptron, *IEEE Trans. Neural Network Learning System*, vol.27, no.4, pp.809-821, 2015.
[16] M. D. Tissera and M. D. McDonnell, Deep extreme learning machines supervised autoencoding architecture for classification, *Neurocomputing*, vol.174, pp.42-49, 2016.
[17] L. J. Dang, F. C. Tian, L. Zhang, C. B. Kadri, X. W. Peng, X. Yin and S. Q. Liu, A novel classifier ensemble for recognition of multiple indoor air contaminants by an electronic nose, *Sensors Actuators A: Physical*, vol.207, pp.67-74, 2014.
[18] S. Nakamura, K. Hiyane, F. Asano, T. Yamada and T. Endo, Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition, *Proc. of the 6th European Conference on Speech Communication and Technology*, pp.2255-2258, 1999.
[19] M. Gonen and E. Alpaydm, Multiple kernel learning algorithms, *Journal of Machine Learning Research*, vol.12, pp.1835-1849, 2016.