Automatic Modulation Classification Using Deep Learning Based on Sparse Autoencoders With Nonnegativity Constraints

Afan Ali and Fan Yangyu

***Abstract*—We demonstrate a novel method for the automatic modulation classification based on a deep learning autoencoder network, trained by a nonnegativity constraint algorithm. The learning algorithm aims to constrain the negative weights, learns features that amount to a part-based representation of data, and disentangles a more meaningful hidden structure. The performance of this algorithm is tested on the fourth-order cumulants of the modulated signals. The results indicate that the autoencoder with nonnegativity constraint (ANC) improves the sparsity and minimizes the reconstruction error in comparison with the conventional sparse autoencoder. The classification accuracy of an ANC based deep network shows improved accuracy under limited signal length and fading channel.**

***Index Terms*—Autoencoder, automatic modulation classification, cumulants, deep learning networks, nonnegativity constraints.**

I. Introduction

THE automatic modulation classification (AMC) module is tasked for the identification of the modulation scheme in the received signal, which is usually complex, and is applied before demodulation. It plays an important role in various military and civilian applications. A detailed overview on the AMC can be found in [1]. Generally, the two methods proposed for the problem of AMC are likelihood based and feature based (FB). FB methods are prominent in the practical implementations because of the less complexity involved.

High-order cumulants [2], wavelet transform [3], and cyclic statistics [4] are some of the widely used features in the FB methods. Over the years, machine learning algorithms, such as support vector machines, K-nearest neighbor, and artificial neural networks, have been widely studied as the classification algorithms in AMC [5]–[7]. More recently, studies have shown that deep neural networks (DNN) can learn from the complex data structures and achieve a superior classification accuracy [8], [9]. This makes them an obvious choice in the problem of AMC mainly because of the denser modulation schemes used in the modern communication system to achieve efficient high-capacity data transmission. The two well-known learning algorithms that produce good representation for initializing deep architectures are [10] restricted Boltzman machines and different types of autoencoders. In this letter, we focus on unsupervised feature extraction based on autoencoders.

Autoencoders-based DNN are composed by stacking pretrained autoencoders layer by layer, followed by a supervised fine-tuning algorithm [11]. To make this model more robust to noise and improve classification accuracy, sparsity is proposed when training the autoencoders [12]. Dai *et al.* employed sparse autoencoders based classifier to classify seven modulation schemes [13]. Although a good classification rate has been achieved under low SNR, computation of the ambiguity function for each modulation scheme makes this approach computationally complex. Moreover, only additive white Gaussian noise (AWGN) channel has been assumed in simulation, which is unrealistic in practical implementations. Likewise, in another work, Kim *et al.* used a deep learning model for the problem of AMC in AWGN and fading channels [14]. They computed 21 features from the received signal based on power spectrum density and cumulants. However, authors used random initialization of the network instead of pretraining with the greedy layer-wise algorithm [15]. This makes it difficult to apply the network to train a large deep network with several layers.

In this letter, we demonstrate how to achieve a more meaningful representation from the received modulated signals based on autoencoders in the DNN. Inspired by the idea of the sparse coding [8], [16], [17] and nonnegative matrix factorization [18]–[20], we strive to unravel the hidden structure of the data by decomposing data into parts by proposing autoencoders employing nonnegativity constraint (ANC). Hence, the highlights of our contributions in this letter can be summarized as follows.

1) We propose a new approach to the problem of AMC by introducing a nonnegativity constraint into the training of the autoencoder to learn a sparse, part-based representation of the input data. The training is then stretched to train a deep network with stacked autoencoders and a softmax classification layer, again constraining the weights of the overall network to be nonnegative. The part-based representation of the data helps to unravel the hidden structure of the data, thus producing a better reconstruction of the data. To the best of our knowledge, no work has been done in AMC, using nonnegativity constraint in training the autoencoders.
2) Simulation results indicate that the proposed ANC method performs better than the methods employing sparse autoencoder (SAE) in AMC under various scenarios.

Manuscript received July 6, 2017; revised September 11, 2017; accepted September 11, 2017. Date of publication September 14, 2017; date of current version September 27, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao Paulo Papa. *(Corresponding author: Afan Ali.)*

The authors are with the School of Electronics and Information Engineering, Northwestern Polytechnical University, Xian 710072, China (e-mail: afanali@mail.nwpu.edu.cn; fan_yangyu@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2752459

The letter is organized as follows. Section II explains the problem statement, including signal model adopted and the signal features used. Section III discusses the method proposed for the problem of AMC. Section IV is about the experiments and results. Conclusion is drawn in Section V.

II. Problem Statement

A. Signal Model

We assume that baseband I and Q components of $r(n)$ are extracted in a coherent, synchronous environment with a signal tone signaling. In our approach, we treat channel as a flat fading in which phase and frequency offsets are added separately. Thus, the baseband samples of $r(n)$ after matched filtering can be written as follows:

$$r(n) = \boldsymbol{a} e^{j(2\pi f_0 n + \theta_0)} s(n) + g(n) \tag{1}$$

where $\boldsymbol{a}$ is the attenuation factor, n is the symbol index, transmitted symbol $s(n)$ is generated from one of M modulations, f_0 is the frequency offset, θ_0 is the phase offset, and $g(n)$ is the complex Gaussian noise with power σ_g^2. Moreover, to compensate for the unknown channel gain and to bring consistency on all signal models, we normalize I and Q component of the received samples as $(\Re(r(n)) - \mu_x)/\sigma_x$ and $(\Im(r(n)) - \mu_y)/\sigma_y$, respectively, where μ is the estimated signal mean and σ is the estimated standard deviation.

B. Signal Features

Although various features can be used for AMC, in this letter we consider the fourth-order cumulants of $r(n)$ as features for classification [21]. They present a very distinct and robust shape of the distribution of noisy signal constellation. Given N samples, $\{r(n)_{n=1}^N\}$, the fourth-order cumulants of $r(n)$ are computed as follows:

$$\begin{aligned}
\hat{C}_{40} &= \frac{1}{N}\sum_{n=1}^{N} r^4(n) - 3\hat{C}_{20}^2 \\
\hat{C}_{41} &= \frac{1}{N}\sum_{n=1}^{N} r^3(n) r^*(n) - 3\hat{C}_{20}\hat{C}_{21} \\
\hat{C}_{42} &= \frac{1}{N}\sum_{n=1}^{N} |r(n)|^4 - |\hat{C}_{20}|^2 - 2\hat{C}_{21}^2
\end{aligned} \tag{2}$$

where $r^*(n)$ is the conjugate of $r(n)$, $\hat{C}_{20} = \frac{1}{N}\sum_{n=1}^{N} r^2(n)$, and $\hat{C}_{21} = \frac{1}{N}\sum_{n=1}^{N} |r(n)|^2$. Assume we use $\boldsymbol{\psi} = g(r(N))$ to extract the features of $r(N)$, e.g., $\hat{C}_{40}$, $\hat{C}_{41}$, and $\hat{C}_{42}$ in (2), where $\boldsymbol{\psi}$ describes the input feature vector, which contains the concatenated features from each modulation scheme.

III. Proposed Method

We propose a three layer fully connected DNN based on autoencoders. It works on the principle of unsupervised learning and endeavors to recompute its input at the output. As shown in Fig. 1, it strives to learn a function

$$\hat{\boldsymbol{\psi}} = f_{\mathbf{W},\mathbf{b}}(\boldsymbol{\psi}) \approx \boldsymbol{\psi} \tag{3}$$

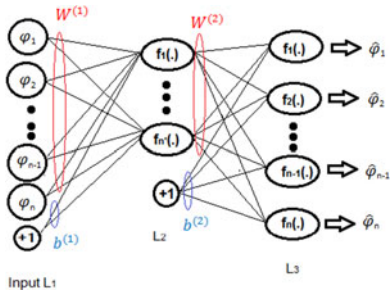

Fig. 1. Schematic diagram of the proposed autoenocoder-based DNN.

where $\mathbf{W} = \{W_1, W_2\}$ and $\mathbf{b} = \{b_1, b_2\}$ represents weight and biases of both layers, respectively. Given our input vector $\boldsymbol{\psi} \in [0,1]^n$, the first deterministic mapping is parametrized by $\theta_1 = \{W_1, b_1\}$ and is achieved as

$$\mathbf{h} = g_{\theta_1}(\boldsymbol{\psi}) = f(W_1\boldsymbol{\psi} + b_1) \tag{4}$$

where $\mathbf{h} \in [0,1]^{n'}$, $W_1 \in R^{n' \times n}$, $b_1 \in R^{n' \times 1}$, and $f(.)$ is a sigmoid function defined as $f(x) = \frac{1}{1+e^{-x}}$. The hidden representation, $\mathbf{h}$ in (4) is mapped back to reconstructed vector, $\hat{\boldsymbol{\psi}} \in [0,1]^n$, parametrized by $\theta_2 = \{W_2, b_2\}$ and given by

$$\hat{\boldsymbol{\psi}} = g_{\theta_2}(\mathbf{h}) = f(W_2\mathbf{h} + b_2) \tag{5}$$

where $W_2 \in R^{n \times n'}$ and $b_2 \in R^{n \times 1}$. To optimize the parameters of the model in (3), i.e., $\{\theta_1, \theta_2\}$, the average reconstruction error is used as the cost function

$$J(\mathbf{W}, \mathbf{b}) = \frac{1}{m}\sum_{z=1}^{m} \frac{1}{2} \|\hat{\boldsymbol{\psi}^{(z)}} - \boldsymbol{\psi}^{(z)}\|^2 \tag{6}$$

where m is the total number of training examples.

In conventional networks, two restrictions are imposed to the cost function in (6) to produce a more meaningful hidden representation of the input. The first restriction added to (6) is known as the sparsity constraint, which results in a rapid convergence of training using the back propagation algorithm. Several algorithms have been proposed in literature to learn a sparse representation using autoencoders [22]. Here, we impose sparsity by imposing a limitation to the activation of hidden units $\mathbf{h}$ using a Kullback–Leibler (KL) divergence function [23]. Assume, $h_j(\boldsymbol{\psi}^{(z)})$ denote the activation of the hidden unit j with respect to the input $\boldsymbol{\psi}^{(z)}$. Then, average activation of the hidden units is given by

$$\hat{p_j} = \frac{1}{m}\sum_{z=1}^{m} h_j(\boldsymbol{\psi}^{(z)}). \tag{7}$$

To impose sparsity, we limit $\hat{p_j} = p$, where p is the positive sparsity parameter chosen close to 0. Hence, we try to minimize the KL divergence between $\hat{p_j}$ and p as follows:

$$J_{\mathrm{KL}}(p \| \hat{\boldsymbol{p_j}}) = \sum_{j=1}^{n'} p \log \frac{p}{\hat{p_j}} + (1-p) \log \frac{1-p}{1-\hat{p_j}} \tag{8}$$

where $\hat{\boldsymbol{p_j}}$ is the vector of average hidden activation. The second restriction term added to (6) is known as weight decay constraint which helps to avoid overfitting. Hence, the cost function for

SAE is given by

$$J_{\text{SAE}} = (\mathbf{W}, \mathbf{b}) = J(\mathbf{W}, \mathbf{b}) + \beta J_{\text{KL}}(p \| \hat{\boldsymbol{p}}_{\boldsymbol{j}}) + \frac{\lambda}{2} \sum_{l=1}^{2} \sum_{i=1}^{s_l} \sum_{j=1}^{s_l+1} (w_{ij}^{(l)})^2 \tag{9}$$

where β is the sparsity penalty, λ is weight decay parameter, and s_l and s_{l+1} are the sizes of adjacent layers.

Almost all the published literature has employed J_{SAE} computed in (9) in the training of autoencoders-based DNN for the problem of AMC [13], [14], [24]. Although it achieves a significant improvement in performance from the conventional AMC methods, its performance deteriorates under various multifading channels and limited received signal length. Therefore, we propose to use a part-based representation of input using a nonnegativity constraint in autoencoding. We strive to show that this results in more meaningful representation of the data under various multipath fading channels by disentangling the hidden structure of data. To the best of our knowledge, this method has not been used in the problem of AMC yet.

A. Autoencoders With Nonnegativity Constraint

The part-based representation of data is achieved by breaking the input data into parts, which when combined additively, reconstruct the original input [15]. Therefore, input can be decomposed in each layer of an autoencoder while weights in $\mathbf{W}$ are constrained to be positive [15]. To achieve a better performance in reconstruction of the input data, the autoencoder decompose the input, $\boldsymbol{\psi}$ into parts in the encoding phase given by (4), and then combine them in an additive manner in the decoding phase given by (5). This is achieved by replacing the weight decay constraint in (9) with a quadratic function [20]. We represent this new cost function as J_{ANC}, which is represented as

$$J_{\text{ANC}}(\mathbf{W}, \mathbf{b}) = J(\mathbf{W}, \mathbf{b}) + \beta J_{\text{KL}}(p \| \hat{\boldsymbol{p}}_{\boldsymbol{j}}) + \frac{\alpha}{2} \sum_{l=1}^{2} \sum_{i=1}^{s_l} \sum_{j=1}^{s_l+1} g(w_{ij}^{(l)}) \tag{10}$$

where

$$g(w_{ij}) = \begin{cases} w_{ij}^2, & \text{if } w_{ij} < 0 \\ 0, & \text{otherwise} \end{cases}$$

and $\alpha \geq 0$. We aim to minimize (10), which reduces the number of nonnegative weights of each layer and the overall average reconstruction error. Further, we update the weight and biases using the gradient of (10) as follows:

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \eta \frac{\partial}{\partial w_{ij}^{(l)}} J_{\text{ANC}}(\mathbf{W}, \mathbf{b}) \tag{11}$$

$$b_{i}^{(l)} = b_{i}^{(l)} - \eta \frac{\partial}{\partial b_{i}^{(l)}} J_{\text{ANC}}(\mathbf{W}, \mathbf{b}) \tag{12}$$

TABLE I
DESIGN PARAMETERS FOR SAE AND ANC METHODS

Parameters	SAE	ANC
Sparsity constraint (β)	3	3
Weight decay constraint (λ)	0.004	–
Nonnegativity constraint (α)	–	0.003
Sparsity parameter (p)	0.06	0.06
Max. iterations	400	400

where $\eta > 0$ is the learning rate. The derivative of (10) with respect to the weights can be computed as

$$\frac{\partial}{\partial w_{ij}^{(l)}} J_{\text{ANC}}(\mathbf{W}, \mathbf{b}) = \frac{\partial}{\partial w_{ij}^{(l)}} J(\mathbf{W}, \mathbf{b}) + \beta \frac{\partial}{\partial w_{ij}^{(l)}} J_{\text{KL}}(p \| \hat{\boldsymbol{p}}_{\boldsymbol{j}}) + \alpha r(w_{ij}^{(l)}) \tag{13}$$

where

$$r(x) = \begin{cases} w_{ij}, & \text{if } w_{ij} < 0 \\ 0, & \text{otherwise.} \end{cases}$$

We use back-propagation algorithm to calculate the partial derivatives in (12) and the first two terms in (13).

B. Deep Network Using ANC

In this letter, a deep ANC network is pretrained, i.e., step-by-step training of the three layers of the autoencoders is done using a greedy layer-wise approach [11] and the output of the previous autoencoder is used as the input to the next autoencoder. At the last autoencoder layer, output is used as an input to the softmax classification layer, which is trained in a supervised manner. The last step of the greedy layer-wise training is to stack the trained ANC and softmax layer, followed by fine-tuning the deep network in supervised manner to achieve a superior classification accuracy. We constrained the nonnegative weights of softmax classifer in the same way as was done for ANC in Section III-A.

IV. EXPERIMENT RESULTS

In this section, we report the performance tests of the proposed technique for the classification of five modulation classes, i.e, BPSK, 8-PSK (phase-shift keying), 4-QAM (quadratic-amplitude modulation), 16-QAM, 64-QAM, in two sets of experiments. All the simulations are performed in MATLAB environment. In the first phase of the experiments, we assume that frequency and phase offsets are known at the receiver. The number of samples used are 512, 1024, 2048, and 4096, respectively, and the SNRs used are 0 dB, 5 dB, 10 dB, and 15 dB. For each value of the number of samples and SNR, 10 000 realizations are generated for each modulation class to train the deep network. The results of ANC method is compared by a three-layered SAE of (9). To tune the hyperparameters, each algorithm is tested with a range of values for each regularization parameter to minimize the cost in (9) and (10). Table I shows the values for each parameter. For the test data, same settings are used. Table II summarizes the results for a particular number of samples and SNR. It is clear from these

TABLE II
CLASSIFICATION ACCURACY USING SAE AND ANC METHODS FOR DIFFERENT SNRS AND NUMBER OF SAMPLES

Method	SNRs	512	1024	2048	4096
SAE	0	0.82	0.87	0.92	0.95
	5	0.84	0.89	0.93	0.98
	10	0.9	0.92	0.97	0.99
	15	0.91	0.96	0.99	1.0
ANC	0	0.86 ± 0.03	0.9 ± 0.02	0.95 ± 0.04	0.98 ± 0.02
	5	0.9 ± 0.04	0.93 ± 0.02	0.99 ± 0.04	1.0
	10	0.95 ± 0.03	0.98 ± 0.03	1.0	1.0
	15	0.98 ± 0.04	1.0	1.0	1.0

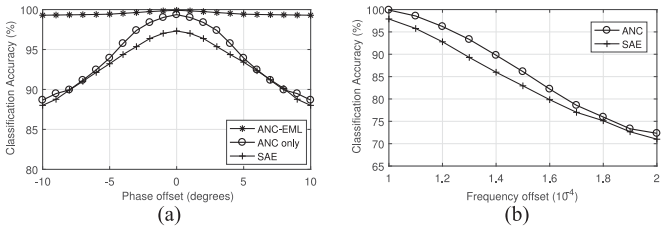

Fig. 2. Probability of correct classification under multipath fading channel. (a) Average classification accuracy versus phase offset. (b) Average classification accuracy versus frequency offset.

results that the proposed ANC method gives better results when compared to conventional SAE methods, e.g., when the SNR is 0 dB and number of samples are 512, ANC shows a performance advantage of about 4%. This shows the capability of the ANC method to extract robust and distinct features under lower SNR and limited signal length. As mentioned in the previous sections, this is justified because nonnegativity constraint added to the cost function of the ANC method disentangles the hidden nature of noisy input data to give better classification of targets.

In the second phase of experiments, we compare the performance of the ANC and SAE methods under the flat-fading channel. Both the SNR and signal length are fixed at 8 dB and 1024, respectively. The parameter settings for ANC and SAE are kept unchanged. The phase and frequency offsets are considered separately. First, we only considered carrier phase offset and presume that the frequency is perfectly matched at the receiver. Hence, our signal model for this experiment is given by

$$r(n) = \boldsymbol{a} e^{j\theta_0} s(n) + g(n). \tag{14}$$

The range of phase offset is considered from -10° to 10° with a step of 1°. For each phase offset, 1000 realizations of signals are generated using (14) and extended maximum likelihood (EML) estimator is used for preprocessing the signal to recover the phase offset as proposed in [25]. Fig. 2(a) illustrates the classification performance with the phase offset. As expected, the ANC method, involving phase offset estimation and recovery scheme, gives consistent classification accuracy of about 99% since mismatching is limited within a reasonable amount. On the contrary, the performance of the ANC and SAE starts to degrade as phase offset increases. Nevertheless, the ANC method, without the EML estimator, gives a performance advantage of about 1.5% till a phase offset of $\pm 5^\circ$. With the phase offset greater than $\pm 5^\circ$, phase mismatch results in a huge channel distortion and performance of both the methods falls considerably, converging to around 70%.

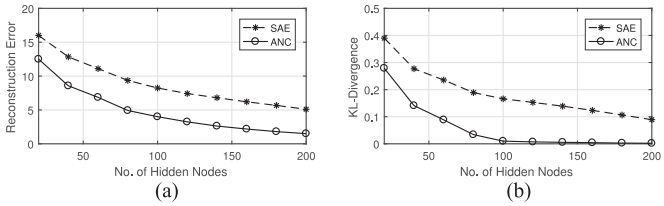

Fig. 3. Performance comparison, (a) reconstruction error computed by (6), (b) sparsity of hidden units measured by the KL divergence computed by (8).

Next, we consider the frequency offset while presuming a perfectly matched carrier phase. To model the receiver in the fading channel with frequency offset, we derive the following equation from (1):

$$r(n) = \boldsymbol{a} e^{j\frac{2\pi \cdot n \cdot f_0}{f}} s(n) + g(n) \tag{15}$$

where f_0 is the frequency offset. We used a relative frequency f_0/f, which is computed as the ratio between the actual frequency offset and the symbol sampling frequency to depict different level of frequency offset in the simulation. The rest of the experiment setup is same as was done in the previous part. The value of the frequency offset is limited to 1×10^{-4} and 2×10^{-4}. The result is shown in Fig. 2(b). The proposed ANC method has performance advantage of around $\pm 3\%$ till a frequency offset of 1.6×10^{-4}. The gap between the two curves decreases between the frequency offset of 1.6×10^{-4} and 2×10^{-4}. This reduced performance can be associated to the dense signal constellations used, e.g., 16-QAM and 64-QAM. There is a limited margin for any frequency offset with dense constellation scatter.

Moreover, we further evaluate the performance of the ANC classifier in terms of reconstruction error with respect to number of hidden nodes. The reconstruction error of the data set computed using (6) is depicted in Fig. 3(a). The results show that the ANC outperforms SAE method for different number of hidden nodes. This further proves that the nonnegativity constraint forces the autoencoders to learn part-based representation of $\boldsymbol{\psi}$, achieving more accuracy in reconstruction from their encoding. Next, to better understand the hidden activities, results in Fig. 3(b) depict the KL divergence computed using (8) for different number of hidden nodes. Greater sparsity achieved in ANC shows decreased value of $J_{\mathrm{KL}}(p\|\hat{\boldsymbol{p}}_{\boldsymbol{j}})$ in (8). This justifies that the hidden nodes in ANC have a less activation value as compared to SAE when it is averaged over full training data.

V. CONCLUSION

In this letter, a novel AMC method was proposed based on a deep autoencoder-based network with nonnegative weight constraint. Nonnegativity encourages a part-based representation of data and disentangles more meaningful hidden features from it. The performance of the proposed method was compared to the conventional SAE based classifier in terms of classification accuracy. The results reveal that a deep network trained by employing nonnegative weights in the autoencoding achieves superior performance in limited signal length and fading channel. This is due to decomposing data into parts, which helps to discriminate between classes.

REFERENCES

[1] O. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, no. 2, pp. 137–156, Apr. 2007.

[2] S. Huang, Y. Yao, Y. Xiao, and Z. Feng, "Cumulant based maximum likelihood classification for overlapped signals," *Electron. Lett.*, vol. 52, no. 21, pp. 1761–1763, 2016.

[3] K. C. Ho, W. Prokopiw, and Y. T. Chan, "Modulation identification of digital signals by the wavelet transform," *IEE Proc.-Radar, Sonar Navigat.*, vol. 147, no. 4, pp. 169–176, 2000.

[4] O. A. Dobre, M. Oner, S. Rajan, and R. Inkol, "Cyclostationarity-based robust algorithms for QAM signal identification," *IEEE Commun. Lett.*, vol. 16, no. 1, pp. 12–15, Jan. 2012.

[5] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Automatic modulation classification using combination of genetic programming and KNN," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2742–2750, Aug. 2012.

[6] A. K. Nandi and E. E. Azzouz, "Modulation recognition using artificial neural networks," *Signal Process.*, vol. 56, no. 2, pp. 165–175, 1997.

[7] H. Mustafa and M. Doroslovacki, "Digital modulation recognition using support vector machine classifier," *in Proc. Conf. Record 38th Asilomar Conf. Signals, Syst. Comput.*, 2004, vol. 2, pp. 2238–2242.

[8] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[9] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal Inf. Process.*, vol. 3, pp. 1–29, 2014.

[10] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[11] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 153–161, 2007.

[12] M. Ranzato, Y. L. Boureau, and Y. LeCun, "Sparse feature learning for deep belief networks," *Adv. Neural Inf. Process. Syst.*, vol. 20, pp. 1185–1192, 2007.

[13] A. Dai, H. Zhang, and H. Sun, "Automatic modulation classification using stacked sparse auto-encoders," in *Proc. 13th Int. Conf. Signal Process.*, 2016, pp. 248–252.

[14] B. Kim, J. Kim, H. Chae, D. Yoon, and J. W. Choi, "Deep neural network-based automatic modulation classification technique," in *Proc. Int. Conf. Inf. Commun. Technol. Convergence*, 2016, pp. 579–582.

[15] J. Chorowski and J. M. Zurada, "Learning understandable neural networks with nonnegative weight constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 62–69, Jan. 2015.

[16] A. Makhzani and B. Frey, "*k-Sparse Autoencoders*," arXiv:1312.5663v2 [cs.LG], 2014.

[17] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] D. Chen and R. J. Plemmons, "Nonnegativity Constraints in Numerical Analysis," in *Proc. Symp. Birth Numer. Anal.*, 2008, pp. 541–544.

[20] E. Hosseini-Asl, J. M. Zurada, and O. Nasraoui, "Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2486–2498, Dec. 2016.

[21] A. Swami and B. M. Sadler, "Hierarchical digital modulation classification using cumulants," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 48, no. 3, pp. 416–429, Mar. 2000.

[22] M. A. Ranzato, C. Poultney, S. Chopra, and Y. Lecun, "Efficient learning of sparse representations with an energy-based model," *Adv. Neural Inf. Process. Syst.*, pp. 1137–1144, 2006.

[23] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[24] A. B. Ambaw, M. Bari, and M. Doroslovacki, "A case for stacked autoencoder based order recognition of continuous-phase FSK," in *Proc. 51st Annu. Conf. Inf. Sci. Syst.*, 2017, pp. 1–6.

[25] V. Zarzoso and A. Nandi, "Blind separation of independent sources for virtually any source probability density function," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2419–2432, Sep. 1999.