CONVOLUTIONAL NEURAL NETWORKS WITH BINAURAL REPRESENTATIONS AND BACKGROUND SUBTRACTION FOR ACOUSTIC SCENE CLASSIFICATION

Yoonchang Han[1], *Jeongsoo Park*[1,2] *Kyogu Lee*[2]

[1] Cochlear.ai, Seoul, Korea
[2] Music and Audio Research Group, Seoul National University, Seoul, Korea
{ychan, jspark}@cochlear.ai, kglee@snu.ac.kr

ABSTRACT

In this paper, we demonstrate how we applied convolutional neural network for DCASE 2017 task 1, acoustic scene classification. We propose a variety of preprocessing methods that emphasise different acoustic characteristics such as binaural representations, harmonic-percussive source separation, and background subtraction. We also present a network structure designed for paired input to make the most of the spatial information contained in the stereo. The experimental results show that the proposed network structures and the preprocessing methods effectively learn acoustic characteristics from the audio recordings, and their ensemble model significantly reduces the error rate further, exhibiting an accuracy of 0.917 for 4-fold cross-validation on the development. The proposed system achieved second place in DCASE 2017 task 1 with an accuracy of 0.804 on the evaluation set.

Index Terms— DCASE 2017, acoustic scene classification, convolutional neural network, binaural representations, harmonic-percussive source separation, background subtraction

1. INTRODUCTION

Sounds contain a variety of information that humans use to understand the surroundings, and our behaviours and thoughts are heavily based on this auditory information along with information gathered from different sensory registers. Even if visual information is not given, humans can easily recognise the scene from the surrounding sounds because our expectations are well trained from experience. For instance, we know that bird chirping sound is likely recorded in the park, and cutlery sound is recorded in the restaurant. In addition, it is also possible to guess the size of the space from the sound, because cave-like environment such as metro station produce a lot of reverberations while outdoor scenes do not. However, creating an automated system that understands acoustic scenes is difficult, because it is a fairly high level of information.

Although acoustic scene classification (ASC) is one of the main objectives of machine listening research [1], the research community has lacked benchmark dataset so far [2]. Arguably, Detection and Classification of Acoustic Scenes and Events (DCASE) challenge organised by IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee is one of the first large-scale challenges for ASC research. A number of novel approaches have been proposed in DCASE 2013 [3] and DCASE 2016 [4],and performances of submitted systems are evaluated under the same experimental conditions. In DCASE 2013, most of the submissions are based on hand-made acoustic features along with classifier such as in [5, 6]. Some techniques that widely used for image processing such as a histogram of gradients (HOG) [7] and recurrence quantification analysis (RQA) [8] features also achieved top places. There was also an approach that utilises deep learning such as [9] using restricted Boltzmann machine, but it showed moderate classification accuracy, presumably due to small amounts of data.

DCASE 2016 task 1 is essentially an extended version of the previous DCASE 2013 ASC task, providing a larger amount of data for an increased number of scenes. Many of participants applied a deep learning approach such as a convolutional neural network (ConvNet) [10, 11, 12] and recurrent neural network (RNN) [13, 14]. Although deep learning approach has been successful, top ranks were achieved by i-Vector [15] and non-negative matrix factorization (NMF) [16], which are rather conventional dictionary learning methods. Also, about half of submitted algorithms in this challenge used mel-frequency cepstral coefficients (MFCCs), one of the most popular hand-made features. As can be seen from the results of the DCASE task in the past, the deep learning approach has shown promising results but clearly no better than the existing methods.

Deep learning technology is rapidly evolving everyday. Although DCASE 2017 [17] provides an increased amount of audio data compare to 2013, it is still not sufficient to take full advantage of the potential of deep learning approach. However, we believe that finding an appropriate way to utilise deep learning is one of the most important research topics in the audio processing field at the moment. This paper demonstrates our approach on ASC task using ConvNets and propose various audio domain specific preprocessing methods that emphasise the different aspects of the acoustic scene. The following sections describe the details of the proposed system and the experimental results and conclusions.

2. SYSTEM ARCHITECTURE

This section introduces the proposed audio preprocessing methods.It also describes the details of the proposed ConvNet architecture and how we have configured the ensemble model from them.

2.1. Audio Preprocessing

In general, we used a full 44.1 kHz without downsampling and amplitude of audio clips was normalised first. Then, we extracted the spectrograms with 128 bin mel-scale following [10] which is a sufficient size to keep spectral characteristics while greatly reduce feature dimensions. The window size for short-time Fourier transform was 2,048 samples (46 ms) with a hop size of 1,024 samples (23 ms). The resulting mel-spectrogram was converted into logarithmic scale, and standardised by subtracting the mean value and dividing

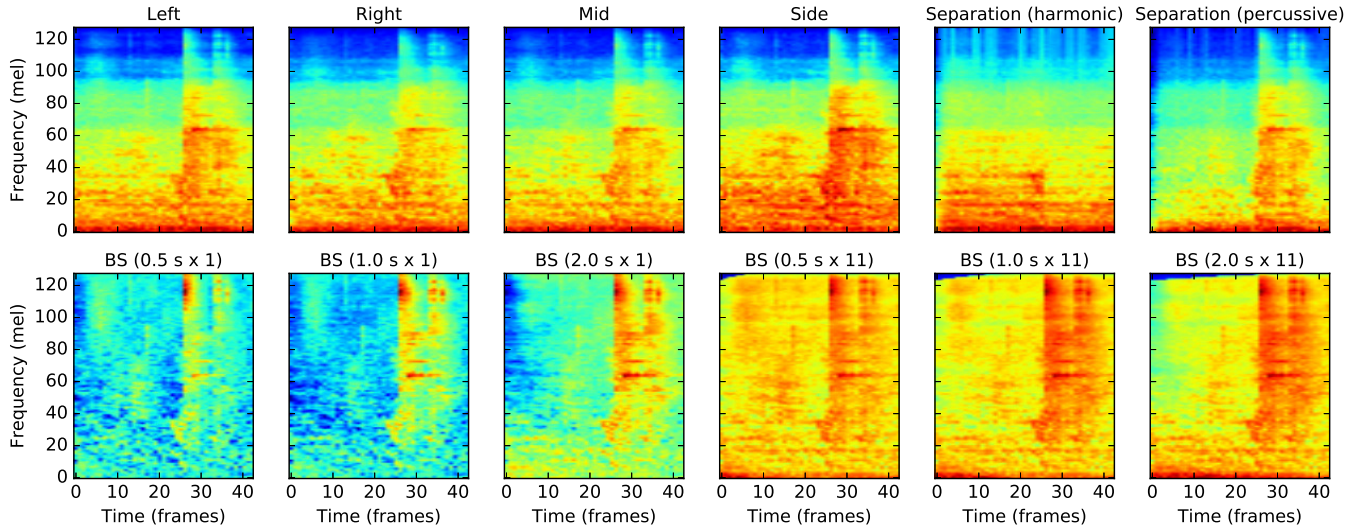

Figure 1: Extracted mel-spectrogram examples of proposed preprocessing methods applied to an audio clip for "café/restaurant" scene. "BS" is background subtraction method, and the numbers in the brackets are the median filtering kernel sizes for time and frequency axes.

by the standard deviation. Standardisation is performed feature-wise and parameters are obtained only from training data to scale both of training and testing data. Finally, we split 10 s audio clip into 1 s audio chunks without overlap for both of training and testing. We used multiple versions of mel-spectrogram which can be largely divided into three methods which are binaural representations, source separation, and background subtraction (BS). A detailed explanation of each method is presented below, and examples of extracted mel-spectrograms are illustrated in Fig.1.

2.1.1. Binaural representations

Although it is common to record audios in stereo, it is usual to make it monaural first by averaging signals prior to processing, as in our previous work [10]. However, we decided to use left-right (LR) and mid-side (MS) pairs in this work, because these contain richer spatial information than mono. For instance, if a car passes in front of a microphone, the sound moves from L to R or R to L, while it is just amplitude change in mono. In addition, the MS representation emphasises the time difference between the sounds reaching each side of the stereo microphone. Use of binaural information have shown superior results in the previous DCASE challenge as in [15] as well. The Mid channel is defined as $L + R$ and the side channel is defined as $L - R$ which is a difference between two channels. For LR and MS, we used 2-conv. model for the analysis explained in the Section 2.2.

2.1.2. Harmonic-percussive source separation

Sound can be generally be divided into two types: harmonic and percussive. In conventional research efforts, harmonic-percussive sound separation (HPSS) algorithms were presented in the context of music signal processing aimed to separate drum sounds from the mixture as in [18]. Here, we separated the audio clips in the dataset into two using the NMF-based HPSS algorithm [19] which enables to separately exploit harmonic and percussive aspects of a sound. Prior to the separation, the stereo sounds are converted to mono. The experimental parameters used for the separation are 0.7, 1.05, 1.05, and 0.95 for α, β, γ, and δ, respectively, and frame size and hop size are 4,096 and 1,024 samples, respectively. The total number of bases is set as 200, consisting of 100 flat-initialised percussive bases and 100 randomly-initialized harmonic bases. Wiener filtering was not used for the post-processing of NMF, however, we have made the last 30 iterations out of 100 total iterations not to include prior imposition to reduce any artifacts that may be generated in the separation process.

2.1.3. Background subtraction

Typically, median filtering is used for removal of noise in scanned images. Moore Jr. and Jorgenson [20] used this technique for object extraction by subtracting median filtered data from original data. Although this technique is more commonly used in the image processing fields, we think that it can be useful to eliminate the "steady" noise from the environment or recording devices. By doing so, we expect the spectral characteristics of acoustic events in the mel-spectrogram to be emphasised and to be more robust against overfitting. Similar to object detection technology, we applied median filtering on the mel-spectrogram and subtracted it from the original version. We first converted stereo audio into mono prior to the process. The filter sizes used for median filtering are 21, 43, 87 for the time axis (approximately 0.5 s, 1.0 s, and 2.0 s), and 1, 11 for the frequency axis, which are chosen empirically by the experiment. Note that using a kernel size of 1 for on the frequency axis is virtually 1-D median filtering over time. As shown in the bottom row of Fig.1, the background subtraction process emphasizes different spectral characteristics from neighboring regions, which makes it easier to detect acoustic events.

2.2. Network Architecture

We used ConvNet consisting of 8 convolution layers using 3×3 receptive fields inspired by VGGNet [21]. In recent years, it has become common to use extremely deep (100 <) network and residual

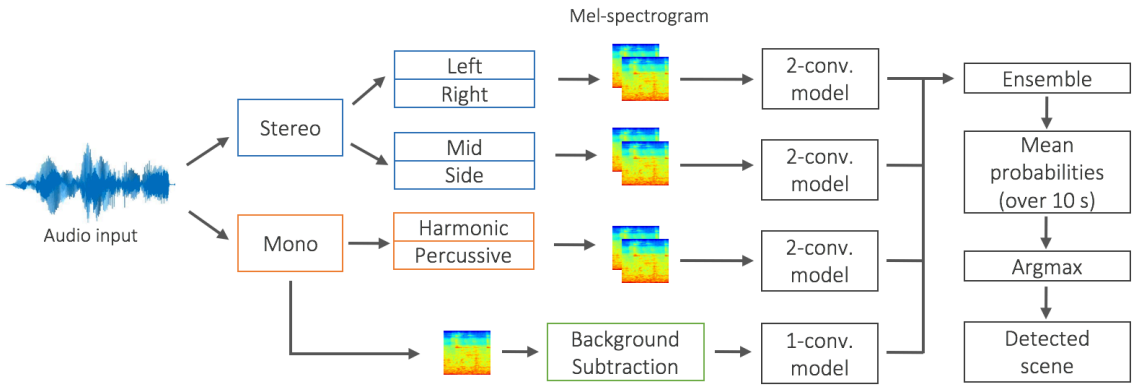

Figure 2: The overall architecture of the proposed system. Multiple ConvNet models are individually trained using a various preprocessing methods and combined into an ensemble model. It then calculates the average probabilities for entire audio clip to detect the scene.

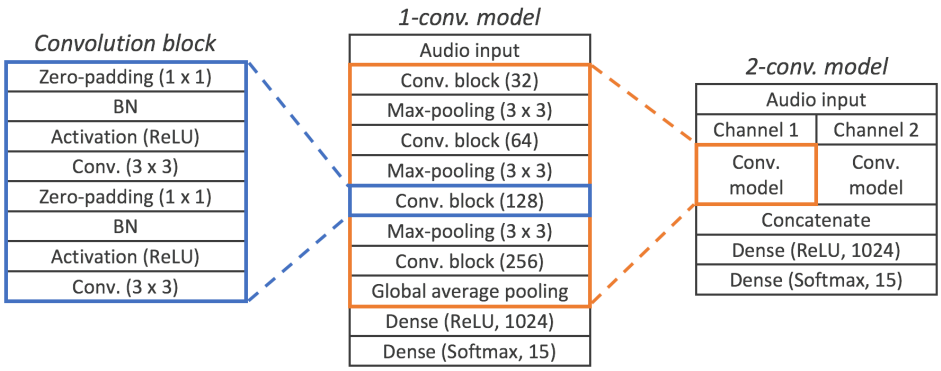

Figure 3: Details of the proposed convolution block, 1-conv. model, and 2-conv. model. The numbers in the brackets are the kernel size for padding/convolution/pooling layers, the number of filters for convolution blocks, and the number of hidden units for dense layer.

connections such as in [22, 23] in the computer vision field. However, we found that it is not highly effective to increase the number of layers or to use a residual connection, at least in our framework, likely due to insufficient amount of data to extract full advantage out of it. The overall architecture of the proposed system is illustrated in Fig.2 which uses two different network architectures: 1-conv. model and 2-conv. model. The former is used for single mel-spectrogram input such as BS, and the latter was used for paired input such as LR, MS, and HPSS. 2-conv. model is similar to 1-conv. model, but processes two channels individually and concatenated before the last fully-connected layer. For both models, we used the same convolution block as illustrated in Fig.3. We employed batch normalization (BN) [24] and rectified linear unit (ReLU) which are de facto standard for modern ConvNets. However, BN and activation function are located before the convolution layer in our proposed network, unlike other networks, because we can get a steady improvement in accuracy. This kind of pre-activation concept can be found in recent residual network papers [22, 23]. We consider that the improvement mainly comes from BN applied for the input data and after max-pooling layers, prior to convolution process.

2.3. Network Ensemble

The results generated by using the same network may be slightly different and model ensemble can generalize this problem [25]. Therefore, we repeated all experiment 3 times using the validation set extracted with different random seeds for each model and took an average probability for each class. In the final decision process, we used two strategies the mean ensemble and ensemble selection method proposed by Caruana et al. [26]. Ensemble selection algorithm aims to find the optimal combination weights by iteratively adding models that maximize the performance of the combination set. We used the mean of test accuracy of all folds as a target value and initialized the ensemble model using LR, MS, and HPSS prior to adding other models. We used 200 iteration and optimal weight we found was 36, 25, 21, 23, 29, 33, 17, 12, and 7 for LR, MS, HPSS, and BS following the order listed in Table1, respectively. Note that this ensemble selection makes use of label from test fold of cross-validation as a hill-climbing set, thus it should not be directly compared to other results of cross-validation.

3. EXPERIMENTS

3.1. DCASE 2017 ASC Dataset

The DCASE 2017 task 1 includes 15 scenes which are bus, café/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office, residential area, train, tram, and urban park. A total 312 segments (52 minutes of audio), recorded at 44.1 kHz with 24-bit resolution in stereo, were provided

Algorithms	Mean Acc.	Algorithms	Mean Acc.
Baseline	0.748	BS (2.0 s, 1)	0.816
Mono	0.844	BS (0.5 s, 11)	0.861
LR	0.871	BS (1.0 s, 11)	0.856
MS	0.879	BS (2.0 s, 11)	0.843
HPSS	0.869	Mean ensemble	0.917
BS (0.5s, 1)	0.801	Ensemble sel.*	0.919
BS (1.0s, 1)	0.805		

Table 1: Mean accuracy for 4-fold cross-validation using proposed ConvNet with various preprocessing and ensemble methods. Baseline and mono are not used for ensemble models, but illustrated for comparison purpose. Note that the result with * used test label, hence it should not be directly compared to other results.

per scene and the length of the audio segments were 10 seconds. The dataset size is increased compare to 2016, but the length of each audio segment was shortened to 10 s from 30 s, so each audio clip contains less information.

3.2. Experiment Settings

The experiment was carried out using 4-fold cross-validation setting provided by the organizer. Network training was performed by optimizing the categorical cross-entropy and stochastic gradient descent (SGD) with Nesterov momentum [27]. The learning rate, decay, and mini-batch size were set to 0.02, 0.0001, and 128, respectively. We trained the network with the NVIDIA GTX 970 and the experiment took about 2 h per model. We used a randomly selected 15% of the training data for validation and the network training was early-stopped if the validation loss did not decrease by more than 20 epochs. The number of examples for training was about 29,800. Baseline system provided by the organizer used mel-spectrogram with 40 mel with a frame size of 40 ms as an input feature for 2 layers x 50 hidden units multilayer perceptron (MLP).

4. RESULTS

4.1. Cross-validation Results

Accuracy is used as a performance metric and the 4-fold mean accuracy of each preprocessing method and ensemble models are presented in Table 1. As a result, the accuracy of the 2.conv-models was 0.87, and BS with various settings (1-conv. models) was generally not as good as 2-conv. models. By combining the results from all the models, it was possible to improve the mean accuracy to 0.917, and ensemble selection slightly pushed it up to 0.919. Because of page limitations, we could not present all class-specific results. However, BS results showed quite different confusion between classes, depending on median filtering size, which is the main reason for the performance improvement of the ensemble. For instance, although the result of BS (0.5 s, 1) are relatively poor compared to other methods, it showed about 16% higher accuracy than the LR for “bus” scene. The confusion matrix of ensemble selection model result is presented in Fig. 4, and it can be observed that the confusion is relatively focused in the home and office, park, and residential area.

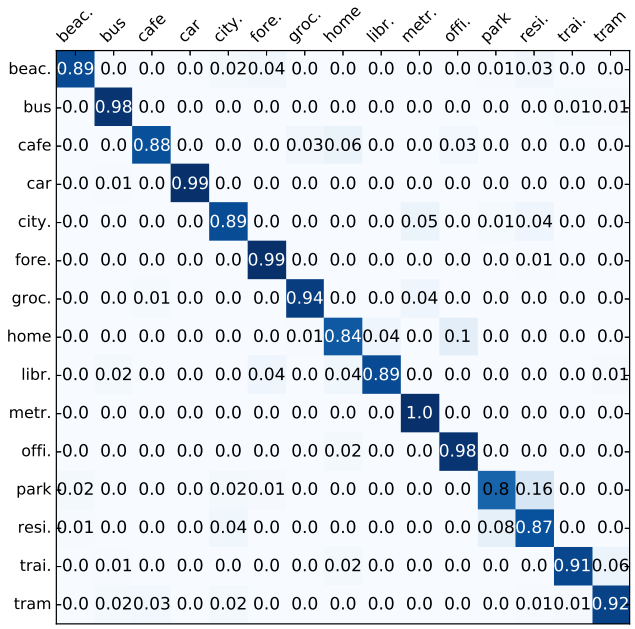

Figure 4: 4-fold mean confusion matrix of the proposed ConvNet with ensemble selection. X-axis indicates the predicted label and Y-axis indicates the true label.

4.2. DCASE 2017 Submission

We used the same experiment settings from development set for the evaluation set. For the final submission, we submitted four slightly different results following the challenge rule. We used a mean probability of 4-fold cross-validation models for submission 1 and 2, and used a newly trained model using a full development set for submission 3 and 4. Regarding ensemble method, we used ensemble selection for submission 1 and 3, and mean ensemble for submission 2 and 4.

4.3. Acknowledgement

This research was supported by Korean government, MSIP provided financial support in the form of Bio&Medical Technology Development Program (2015M3A9D7066980).

5. CONCLUSION

In this paper, we illustrated how we applied ConvNet for identifying the acoustic scene. The main contribution of this paper is presenting a various preprocessing methods that are useful for ConvNet, also having a great synergy when combined together in an ensemble model. As a result, we could obtain an accuracy of 0.917 for 4-fold cross validation on the development set and 0.804 on the evaluation set. In the future, we plan to investigate the optimal kernel size for BS and pre-activation convolution block further, which are currently selected heuristically.

6. REFERENCES

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-

IEEE press, 2006.

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[4] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.

[5] D. Li, J. Tam, and D. Toub, "Auditory scene classification using machine learning techniques," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[6] J. T. Geiger, B. Schuller, and G. Rigoll, "Recognising acoustic scenes with large-scale audio feature extraction and svm," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[7] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 142–153, 2015.

[8] G. Roma, W. Nogueira, P. Herrera, and R. de Boronat, "Recurrence quantification analysis features for auditory scene classification," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, vol. 2, 2013.

[9] J. Nam, Z. Hyung, and K. Lee, "Acoustic scene classification using sparse feature learning and selective max-pooling by event detection," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[10] Y. Han and K. Lee, "Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[11] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 acoustic scene classification using convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[12] J. Kim and K. Lee, "Empirical study on ensemble method of deep neural networks for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[13] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, and B. Schuller, "The up system for the 2016 DCASE challenge using deep recurrent neural network and multiscale kernel subspace learning," DCASE2016 Challenge, Tech. Rep., September 2016.

[14] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," DCASE2016 Challenge, Tech. Rep., September 2016.

[15] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: a hybrid approach using binaural i-vectors and deep convolutional neural networks," DCASE2016 Challenge, Tech. Rep., September 2016.

[16] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[17] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "Dcase 2017 challenge setup: Tasks, datasets and baseline system."

[18] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–4.

[19] J. Park, J. Shin, and K. Lee, "Exploiting continuity/discontinuity of basis vectors in spectrogram decomposition for harmonic-percussive sound separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1061–1074, 2017.

[20] A. W. Moore Jr and J. W. Jorgenson, "Median filtering for removal of low-frequency background drift," *Analytical chemistry*, vol. 65, no. 2, pp. 188–191, 1993.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[23] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[25] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in neural information processing systems*, 1995, pp. 231–238.

[26] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 18.

[27] Y. Nesterov *et al.*, "Gradient methods for minimizing composite objective function," 2007.