Scalogram Neural Network Activations with Machine Learning for Domestic Multi-channel Audio Classification

Conference Paper · December 2019

DOI: 10.1109/ISSPIT47144.2019.9001814

4 authors, including:

Stefano Fasciani
University of Oslo
25 PUBLICATIONS **44** CITATIONS

SEE PROFILE

Scalogram Neural Network Activations with Machine Learning for Domestic Multi-channel Audio Classification

Abigail Copiaco, Christian Ritz, Stefano Fasciani†, and Nidhal Abdulaziz*
University of Wollongong, Australia
University of Oslo, Norway†
University of Wollongong in Dubai, UAE*
E-mail/s: {abigailc, critz}@uow.edu.au, stefano.fasciani@imv.uio.no†, nidhalabdulaziz@uowdubai.ac.ae*
Dubai Knowledge Park, P.O. Box 20183, Dubai, UAE, Fax: +971 4 2781801

***Abstract* – Current methodologies explored for audio classification, particularly multi-channel audio, commonly involve the use of individual deep learning approaches. In this paper, we look at domestic multi-channel audio classification through a comparison of various combinations of existing pre-trained Neural Network (NN) models, with Support Vector Machine (SVM) for classification. The NN model is first trained with spectro-temporal features extracted from the audio, characterized by scalogram images that are generated through the Continuous Wavelet Transform (CWT). Activations that are extracted from the selected layer of the concerned neural network model are then sent as features used to train the machine learning approach for classification. Utilization of the network activations learnt from the deep learning component of the classifier strengthens the time-frequency features of the signal that are extracted from the spectrogram, allowing further improvement to the accuracy. For the full SINS development database, best results yielded an F1-score of over 97% for the tenth layer of the Xception network when combined with the multi-class Linear SVM, showing a drastic improvement from the top performing F1-score achieved in the DCASE 2018 Task 5 challenge, which rests at around 89%.**

***Index Terms* – Machine Learning, deep learning, multi-channel audio, classification, support vector machines,**

I. INTRODUCTION

Over the past few years, the popularity of audio applications has continually been rising due to the ethical issues faced by visual-based systems, mostly involving the intrusion of privacy [1,2]. Although majority of previous works are focused on the classification of single-channel sound events [3,4], more recent publications discuss the utilization of multi-channel acoustic scenes [5,6], which are preferred for domestic applications. Multi-channel audio contains data gathered from multiple channels, which can be processed together for better accuracy, as it allows data-driven approaches to thoroughly learn the underlying target sounds in complex and noisy acoustic environments. Furthermore, multi-channel audio detection yielded a 10% increase in accuracy when compared to single channel [7].

This work focuses on multi-channel domestic acoustic scene classification, as per the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Task 5 challenge [8]. Highest performing contributors in the DCASE 2018 Task 5 challenge mostly utilize log-mel energies and Mel-frequency Cepstral Coefficients (MFCC) as features, while pre-trained convolutional neural network models were commonly used for classification [8]. Moreover, other existing methods and applications concerning multi-channel audio classification often utilize machine or deep learning techniques on their own.

In contrast, this paper proposes the combination of deep and machine learning techniques for multi-channel audio classification. Traditionally, classification for CNNs is based on functions, such as the softmax function, applied at the last layer of the network. Although there are a couple of works utilizing features extracted from neural networks [9,19,20], none of these works classified these features using machine learning techniques. Although a few other sources had investigated the use of neural network-learned features for classification, these have been used for different applications. Additionally, different types of features and end classifier methods were utilized to achieve the final results.

In this work, we explore and compare the performance of various combinations of different layers of pre-trained deep learning networks with the Linear Support Vector Machine (SVM) machine learning approach, for training and classification. SVMs are known for their advantages in high accuracy, which is apparent even for unstructured data [10,11]. Scalogram images are then used as features, which are extracted from the average of the Continuous Wavelet Transforms (CWT) for each channel of the audio files [12]. Scalograms are selected as features to the neural networks due to its spectro-temporal nature, taking into account both the time and frequency components of the signal [12], which is advantageous for mapping the properties of the constant movement of continuous signals with minimal loss of information. Other subsets of possible audial features are either cepstral-based, spectral-based, or temporal-based, which either considers time or frequency components individually [12]. For the context of domestic acoustic scenes, an excellent time and frequency localization is necessary in order to take into account several sound events that may happen during its continuous duration.

These scalogram features are then used to train the deep learning component of the network for up to a specified layer, from which network activations are to be obtained. Comparison of different combinations tested are then made through performance metrics in terms of the F1-score. Further, the response of the proposed technique to an increase in dataset is also observed.

The rest of the paper is organized as follows. Section II provides some background information and related work involving Deep Convolutional Neural Networks and Support Vector Machines. Section III provides an overview of the proposed methodology, including the combination process of deep and machine learning algorithms for classification. Section IV then describes the experiments conducted, as well as the results used for the evaluation of the performance of the proposed approach. Finally, the concluding section provides observations and propositions to extend the scope for future work.

II. Deep Learning and Machine Learning

A. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) fall under the category of neural networks which uses the combination of signals (convolution) for classification [13]. CNNs are composed of three different layers, inclusive of the convolutional layer, pooling layer, and fully connected layer [14]. It functions such that a compendium of the neurons of the previous layer are connected to the subsequent layer. The concise transmission returns positively in run time, computational complexity, and resource requirements. A combination of several layers of CNNs is called a Deep Convolutional Neural Network (DCNN) [16].

CNN is known for its advantages in terms of computational efficiency [15]. Furthermore, it provides good accuracy for various applications, particularly those that involve data possessing a spatial relationship. For these reasons, CNNs are commonly used as a sole method for classifying large audial data, including multi-channel acoustic scenes, as per the works cited in [17,18].

Nonetheless, as mentioned in the previous section, none of these works had explored the combination of neural networks with machine learning techniques. Although a few other sources had investigated the use of neural network-learned features for classification, these have been used for different applications. Additionally, different types of features and end classifier methods were utilized to achieve the final results [19,20].

B. Support Vector Machines

The Support Vector Machines (SVM) is a type of machine learning technique that aims to maintain the classification error at the lowest point, while keeping the distance between vectors of separate categories at a maximum [10,11]. A variety of sub-types exist for Support Vector Machines. Nonetheless, the focus of this work leans towards Linear SVM.

The process of the Linear SVM starts by creating a linear kernel matrix via pair-wise multiplication of voxels [11]. The locations of the features are defined depending on their intensity [10]. A hyper plane, defined as the optimal linear separation between different classes, is then generated [11].

Although SVMs are highly accurate, it can cause problems in terms of organization and computation, especially when handling large data [10,11]. Such disadvantage accounts for the fact that this technique is rarely used for multi-channel audio classification.

Therefore, in order to overcome the disadvantages of the individual deep and machine learning techniques, this paper tackles their combination by using network activations learned from the deep learning approach, as features to the machine learning technique.

III. Feature Extraction, Network Activations Extraction, and Classification

The overall presented technique begins by obtaining the CWT coefficients from each of the four channels of the audio signal. Scalogram plots, which constitutes time and frequency components of the signal, are then generated accordingly, and are resized into the relevant image size requirement of the pre-trained model through bi-cubic interpolation coupled with antialiasing. After these are fed to specific pre-trained CNN models, network layer activations are extracted. The network activations are then used as features to train the multi-class linear SVM, from which the final prediction results are obtained.

The succeeding sections present details of the proposed methodology, starting from the database used, as well as the extraction of the Scalogram features, the neural network training, and its combination with the SVM. Furthermore, detailed information regarding the system performance evaluation is also provided.

A. SINS Database and CWT Scalogram Generation

Implementation of the methodology was carried out through the readily available domestic acoustic dataset provided by the Sound Interfacing through the Swarm (SINS) database [21]. This has been developed in order to aid researchers that were contributing to the DCASE 2018 Task 5 challenge, with the aim of classifying common household tasks [8]. Each sample in the dataset runs 10-seconds in duration, and represents four individual channels, having been recorded through an Acoustic Sensor Network (ASN) which consists of 13 nodes with 4 microphones each [21]. Sampling is achieved at a frequency of 16 kHz and a bit depth of 12 [21]. This database allows classification in nine categories, inclusive of: Absence, Cooking, Eating, Dishwashing, Watching TV, Vacuuming, Movement, Social Activities, and Other Activities. Recordings have been made in a span of a week.

CWT coefficients are used as features to the deep convolutional neural network portion of the proposed methodology. Such coefficients are extricated through MATLAB's Audio System and Data Communications toolboxes. The wavelet computation involves an analytic Morse wavelet with a gamma constant of 3, and a time-bandwidth product of 60. A total number of 144 coefficient scales are calculated for each channel of the audio signals.

The coefficients scales calculation is performed on the average of the four audio channels. Scalograms are then

extracted, and are resized as per the requirement via bi-cubic interpolation. An example of a scalogram image generated under the category of 'Absence' can be seen in Fig.1. Such images are then used as the inputs to the pre-trained deep convolutional neural network models, which will further be detailed in the next sub-section.

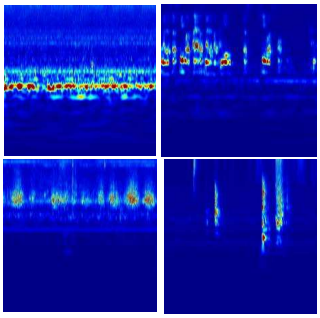

Fig. 1 Scalogram Image for (a) Absence, (b) Social Activities, (c) Vacuum, (d) Cooking

B. *Combining Pre-trained Neural Network Models with Support Vector Machines Classifiers*

In order to combine the advantages of deep learning and machine learning for classification, activations extracted from pre-trained convolutional neural networks are coupled with multi-class linear support vector machines for the scope of this work. The gist of the overall process can be seen in Fig 2. Subsequent to the extraction and concatenation of the scalograms for all channels of the audio files, these are sent into a pre-trained model in order to re-train the model according to the specific application. Several types of pre-trained models are investigated and compared according to their performance, inclusive of the ResNet [22], Xception [23], InceptionResNet [24], GoogleNet [25], and AlexNet [26] models. A summary of these models can be observed below.

Table 1 Pre-trained Convolutional Neural Networks Comparison

Year	Model	Image size	Layers	Top-5 error rate	No. of parameter (million)
2012	AlexNet	227x227	8	15.3%	60
2014	GoogleNet	224x224	22	6.67%	7
2015	ResNet-101	224x224	101	3.6%	44.6
2016	Xception	299x299	71		22.9
2016	Inception-ResNetV2	299x299	164		55.9

After training up to a specified fully-connected layer, the network layer activations are extracted through the MATLAB Parallel computing toolbox and a CUDA enabled NVIDIA GPU.

Having been extracted from neural network training, these activations composes of enhanced time-frequency signal components when compared to their scalogram counterparts. These network layer activations are utilized as features to train the SVM. It is important to note that scalogram images that were used as features to the model cannot be fed directly to the SVM, unless another feature extraction algorithm is used, such as the Bag of Visual Words. This is because wavelets return complex parameters of 144 coefficient scales by 16000 Hz matrix size per label, which cannot be accommodated in arrays to train the SVM with. Extracting network layer activations, however, returns a 1-by-n size of features per label, where "n" is the number of network layer activations extracted from the model.

The Multi-class Linear SVM is then trained through *K(K-1)/2* SVM models, which utilizes a one-versus-one coding design, where K represents the number of categories involved. Results of the multi-class linear SVM then reflect the final prediction of the classification system.

C. *Performance Evaluation*

In order to assess the performance of the system proposed in this work, evaluation is done according to the following criteria:

1. Per-category and overall comparison of various combinations of pre-trained neural network models and the multi-class linear SVM
2. Investigation of the effects of extracting network activations at different layers of the same model
3. Evaluation of the effects of the increase in the sample set

Various combinations of pre-trained network models with multi-class linear SVM were compared and evaluated according to their F1-scores. This is defined to be "a measure that takes into consideration both the recall and the precision, which are derived from the ratios of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN)" [9]. The F1-score is defined by (2), where recall and precision are given by (3) and (4) [27]. The prediction parameters used in the calculation of the F1-score are obtained from the confusion matrix.

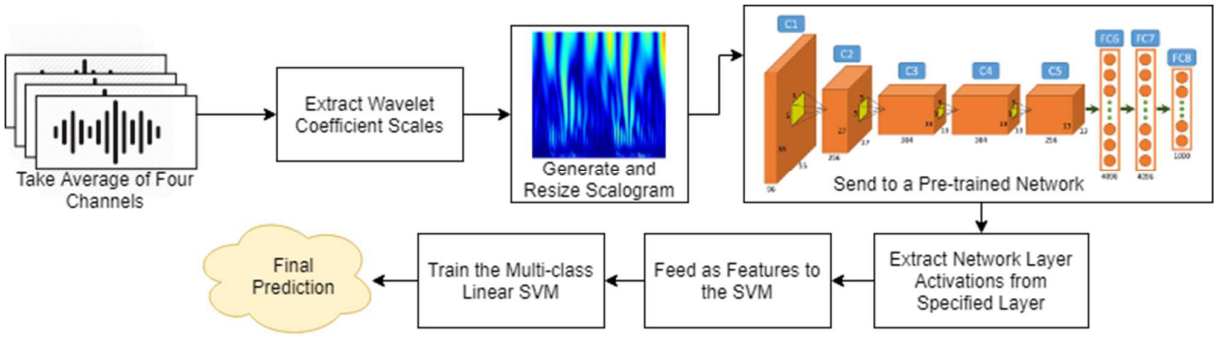

Fig. 2 Combination Layout of Deep and Machine Learning Technique

$$F1score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Performances are evaluated on a sample set of 4293 10-second audio files, with 477 files per level. The SINS database recordings compose of unequal numbers of audio files per category. However, equal number of samples per category were chosen for this work in order to preserve a balanced dataset, and to avoid biasing in favour specific categories with more samples. Overall, 80% of the sample set was used for training, while the remaining 20% were utilized for testing. The following sections detail the results from the performance evaluation.

IV. Experiment Results

A. *Classification through Combinational Methods*

A comprehensive comparison of the performances between the network activations extracted from several pre-trained models in combination with the multi-class linear SVM for multi-channel audio classification can be observed in Tables 2 and 3. CWT-extracted scalograms were used as initial input features for all of the methods compared. Table 2 includes the overall comparison between different method combinations, including the layer from which the network activations were extracted from. Note that network activations can be drawn out from any layer of the CNN model. Table 3, on the other hand, details the per-level comparison of the method performances in terms of their F1-scores for the top five performing classification methods.

Table 2 Pre-trained Neural Network Models with Multi-class Linear SVM Combination Methods Comparison

Method	**Layer**	**F1-Score**
Xception with SVM	10	94.27%
Xception with SVM	6	94.00%
Xception with SVM	8	94.11%
GoogleNet with SVM	15	83.55%
AlexNet with SVM	8	87.71%
ResNet-101 with SVM	170	93.41%
InceptionResNet-V2 with SVM	134	92.98%
InceptionResNet-V2 with SVM	303	93.83%

Table 3 Per-level Methods Comparison

Category	**Xception**			**Inception ResNet-V2 (303)**	**ResNet-101 (170)**
	6	**8**	**10**		
Absence	89	89	90	90	86
Cooking	97	97	97	98	96
Dishwashing	94	95	96	94	94
Eating	94	94	94	92	94
Movement	92	92	91	92	91
Other Act.	81	81	80	78	79
Social Act.	98	98	98	98	99
Television	100	100	100	100	100
Vacuum	100	100	100	100	100

As observed from the results documented in Table 2, Xception, InceptionResNet-V2, and ResNet-101 yields the best network layer activations, which returns the highest F1-scores when combined with multi-class Linear SVM. Such may be due to the fact that these pre-trained models were produced at a later date, and multiple improvements from the previous models have already been made. The Xception model, which returned the highest F1-score for network activations extracted from its tenth layer, functions through a depth-wise separable convolution algorithm.

This algorithm works such that the height and width dimensions are treated separately. The same logic is applied to the depth of the horizontal, where depth-wise convolution is normally applied, followed by a one-by-one filter. This is done for the purpose of covering the dimension of the depth [23]. The main advantage of such algorithm is the reduction in the number of parameters required in order to produce the same number of channels [23], which therefore avoids over-fitting. Reduction in the number of parameters for the Xception model is also reflected from Table 1, which shows that it requires a significantly lower number of parameters as opposed to most neural network models.

B. *Classification using Individual Neural Network Models*

In order to support the reliability of the proposed methodology, as well as for comparison purposes, we had conducted individual performance evaluations for the different pre-trained models discussed in this work, as well as the multi-class linear SVM. The results yielded are summarized in Table 4. As the methods are not combined, final predictions were extracted directly from the last layer of the various neural network models. In the case of the multi-class SVM, Bag of Visual Words algorithm was used in order to extract features from the scalogram images.

Table 4 Multi-channel Audio Classification using Individual Methods

Method	**Average F1-Score**
Xception	90.12%
GoogleNet	90.17%
AlexNet	93.37%
ResNet-101	89.81%
InceptionResNet-V2	90.23%
Multi-class Linear SVM	87.63%

As observed, with the exception of AlexNet and GoogleNet, the rest of the pre-trained models performed better as network layer activation extractors in combination with the multi-class linear SVM. The higher F1-scores achieved when using network activations as features, can be explained by the enhancement in the time-frequency components of the signal, which is provided by the deep learning component of the classification technique. It is also notable that the F1-score achieved with retraining the AlexNet model is comparable to the top methodology discussed in this work. Nonetheless, this implies that despite being an efficient classifier model, the AlexNet model may not be as proficient for extracting network activations. The architecture of AlexNet involves a series of convolutional networks, max pooling, dropout, and Rectified Linear Unit (ReLU) activations [26,28]. While ReLUs are advantageous in terms of reducing the number of required

parameters and avoiding overfitting for overall predictions [24], they could induce problems when network activations are extracted prior to reaching the last fully-connected layer. ReLu remains active in back-propagation solely when the units are positive. Otherwise, the units remain in an off-state, resulting in dead neurons [28]. Additionally, since negative units are dropped, ReLU features positive biasing in adjacent layers. This causes a positive mean shift in subsequent layers.

C. Cross-fold Validation

For the purpose of further assessing the effectiveness of the proposed methodology, cross-fold validation was performed three times for the top two performing methodologies, namely: the AlexNet Model, and the Xception Model with Multi-class Linear SVM. Cross-fold validation results are detailed in Table 5. For every run, a different set of training and testing data are chosen, allowing the expansion of the performance's effectiveness, while preserving a balanced dataset at the same time. Furthermore, such validation permits the affirmation of the results gathered for this work.

Table 5 Multi-channel Audio Classification using Individual Methods Comparison

Run	AlexNet	Xception with SVM
1	94.59%	94.27%
2	92.70%	94.36%
3	92.82%	94.31%
Average	**93.37%**	**94.31%**

As observed, for the size of the balanced dataset involved, although the results are not too far from the AlexNet, the Xception-extracted network layer activations trained with multi-class linear SVM classification reflects more consistent results despite the differing test samples.

D. Response to the Increase in Dataset

Finally, we investigate the application of the same top performing methodologies on the full SINS development database, which comprises of 72,964 audio samples of 10-seconds duration each. Samples are unevenly distributed throughout the 9 levels, with majority of the samples categorized under 'Absence', 'Movement', and 'TV'. Table 6 summarizes the number of samples categorized under each level, as well as the number of audio files that are used for training and testing purposes. For this application, 80% of the samples are used for training, and 20% are utilized for testing.

Table 6 SINS Database Summary

Category	Total Samples	Training	Testing
Absence	18860	15088	3772
Cooking	5124	4099	1025
Dishwashing	1424	1139	285
Eating	2308	1846	462
Movement	18624	14899	3725
Other Act.	2060	1648	412
Social Act.	4944	3955	989
Television	18648	14918	3730
Vacuum	972	778	194
TOTAL	**72964**	**58371**	**14593**

Using the distribution reflected in Table 6, using the AlexNet model for classification results in an average F1-score of 95.58%, yielding a 2% improvement from the results gathered with the lower dataset. This also suggests an improvement of almost 6% from the top performing participant of the DCASE 2018 Task 5 challenge, whose F1-score was 89.95% [29].

Utilizing the network activations extracted from the Xception model with Multi-class Linear SVM returned an F1-score of 97.46% for the SINS development dataset. Although competitive, the application of this classification method for larger datasets came with noticeably higher costs requirements, both in the field of time and in resources. Such observation is supported by the fact that hundreds of thousands of network activations are extracted per audio file from the Xception network. This is then mapped into an array before being fed into the SVM for training. Working with such large array values contributed to the amount of time and memory the machine requires in order to complete the processing.

V. Conclusions and Future Work

In this work, we proposed the idea of combining neural network deep learning techniques with machine learning in the form of the multi-class linear SVM. Performances of several pre-trained models in combination with the multi-class linear SVM for multi-channel acoustic scene classification were evaluated. Throughout the investigation, scalogram images were utilized as features to the neural networks. In order to strengthen our performance study, we had also compared our results with the individual evaluation of the same models. From this, it was found that majority of the pre-trained models work better when used as a network activations extractor for training the SVM. The Xception model yielded the best F1-score for network layer activations extraction, garnering an average F1-score of 94.31% with 4293 samples, and 97.46% for the SINS development dataset. This is due to its depth-wise separable convolution algorithm, which allows the reduction of the required parameters for the same number of output.

On an individual performance level, however, the AlexNet network provided a competitive F1-score of 93.37% for 4293 samples. This was also able to increase to an average F1-score of 95.58% for the SINS development dataset. The AlexNet network utilizes a Rectified Linear Unit (ReLU), enabling rapid training without the cost of the accuracy [22]. Nonetheless, ReLUs are not efficient for extracting network activations due to dead neurons and biasing, which explains why AlexNet does not perform as effectively when combined with multi-class linear SVM. When faced with very large data, however, AlexNet becomes preferable due to the less resource and time requirements. Further, its architecture reacts well with the biasing concerns of an imbalanced dataset.

For future work, improvement of the proposed combination of the Xception and SVM methodologies can be addressed. This can be done through modifying the model to accommodate larger amounts of data without the necessity of large resources. Furthermore, pre-processing the audial inputs can also be considered in the form of data augmentation and noise filtration. Finally, various combination of other deep and machine learning techniques can also be explored towards a common application.

REFERENCES

[1] N. Linda, et al., “Assistive Technologies for people with disabilities – Part II: Current and emerging technologies”, Technical Report, European Parliamentary Research Service, Scientific Foresight Unit (STOA), Jan. 2018.

[2] Bennett and et al., "Assistive Technologies for People with Dementia: Ethical Considerations," Bulletin of the World Health Organization, pp. 1-12, March 2017.

[3] R.M. Alsina-Pags, J. Navarro, F Alas, and M Hervs, "HomeSound: Real-time Audio Event detection based on high performance computing for behaviour and surbeillance remote monitoring," vol. 17, no. 4, 2017.

[4] A. Mitilineos, S.M. Potirakis, N.A. Tatlas, and M. Rangoussi, "A Two-level Sound Classification Platform for Environmental Monitoring," *Hindawi Journal of Sensors*, pp. 2-13, 2018.

[5] K. Xu et al., "Mixup-based Acoustic Scene Classification Using Multi-channel Convolutional Neural Network," Computer Science, Computer Vision and Pattern Recognition, Cornell University, Research 2018.

[6] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, "Speech and Crosstalk Detection in Multi-channel Audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84-91, 2004.

[7] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features," Detection and Classification of Acoustic Scenes 2016, September 2016.

[8] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," Technical Report, Tokyo, Japan, 2018.

[9] J. Pingel and A. Nehemiah (7 Mar 2017), “Object Recognition: Deep Learning and Machine Learning for Computer Vision”, Webinar 3, Mathworks, Accessible at: https://www.mathworks.com/videos/object-recognition-deep-learning-and-machine-learning-for-computer-vision-121144.html

[10] Z. Pan, H. Lu and Adni L. Huang, "Automated Diagnosis of Alzheimer's Disease with Degenerate SVM-based Adaboost," in Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013 5th International Conference on, Hangzhou, pp. 298-301, 2013.

[11] T.S. Korting. (2014, January) How SVM (Support Vector Machines) Algorithm Works. Video.

[12] S. Mallat and S. Shamma, "Audio Source Separation with Time-Frequency Velocities," ScatBSS - Supported by ERC Invariant Class 320959.

[13] H. Phan, L. Hertel, M. Koch, P. Maass, R. Mazur, and A. Mertins, "Improved Audio Scene Classification based on Label-tree Embeddings and Convolutional Neural Networks," IEEE Transactions on Audio, Speech, and Language Processing, pp. 1278-1290, 2017.

[14] A. Dang, T.H. Vu, and J. Wang, "Acoustic Scene Classification using Convolutional Neural Network and Multi-scale Multi-Feature Extraction," IEEE International Conference on Consumer Electronics, 2018.

[15] Q.V. Le et al., "Tiled convolutional neural networks," Computer Sciene Department, Standford University, 2010.

[16] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, and P. Shaohu, "Acoustic Scene Classification using Deep Convolutional Neural Network and Multiple Spectrograms Fusion," in Detection and Classification of Acoustic Scenes and Events 2017, Munich, Germany, 2017.

[17] D. Chong, Y. Zou, and W. Wang, "Multi-channel Convolutional Neural Network with Multi-Level Feature Fusion for Environmental Sound Classification," in 25th International Conference, MMM 2019, Thessaloniki, Greece, 2019.

[18] Y. Han and K. Lee, "Convolutional Neural network with multiple-width frequency-delta data augmentation for acoustic scene classification," Detection and Classification of Acoustic Scenes and Events 2016, Technical 2016.

[19] Z. Ren, et al., “Deep Scalogram Representations for Acoustic Scene Classification”, IEEE/CAA Journal of Automatica Sinica, vol. 5, no. 3, pp. 662-669, May 2018.

[20] Y.R. Pandeya, D. Kim, and J. Lee, “Domestic Cat Sound Classification using Learned Features from Deep Neural Nets”, MDPI Applied Sciences, 2018.

[21] G. Dekkers et al., "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), 2017.

[22] K. He, et al., “Deep Residual Learning for Image Recognition”, Technical Report, submitted 10 Dec 2015.

[23] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolution”, Technical Report, submitted 7 Oct 2016, revised 4 Apr 2017.

[24] C. Szegedy, et al., “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”, Technical Report, submitted 23 Feb 2016, revised 23 Aug 2016.

[25] C. Szegedy, et al. “Going deeper with convolutions”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.

[26] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks” NIPS Proceedings, 2012.

[27] S.L. Phung, A. Bouzerdoum, and G.H. Nguyen, “Learning pattern classification tasks with imbalanced data sets”, in P.Yin (Eds.), Pattern recognition, Vukovar, Croatia: In-Tech, pp. 193-208, 2009.

[28] D. Pedamonti, “Comparison of non-linear activation functions for deep neural networks on MNIST classification task”, arXiv, 8 Apr 2018.

[29] T. Inou et al., "Domestic Activities Classification based on CNN using Shuffling and Mixing Data Augmentation," DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics, Tokyo, Japan, Technical 2018.