INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Acoustic Scene Classification using Kervolution-Based SubSpectralNet

Ritika Nandi[1*], *Shashank Shekhar*[2], *Manjunath Mulimani*[1]
[1]Department of Computer Science and Engineering, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, 576 104, India
[2]Department of Electronics and Communication Engineering, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, 576 104, India
{ritika.nandi77, shashankshekhar90210, manjunath.gec}@gmail.com

Abstract

In this paper, a Kervolution-based SubSpectralNet model is proposed for Acoustic Scene Classification (ASC). SubSpectralNet is a competitive model which divides the mel spectrogram into horizontal slices termed as sub-spectrograms that are considered as input to the Convolutional Neural Network (CNN). In this work, the linear convolutional operation of SubSpectralNet is replaced with a non-linear operation using the kernel trick. This is also known as kervolution (kernel convolution)-based SubSpectralNet. The performance of the proposed methodology is evaluated on the DCASE (Detection and Classification of Acoustic Scenes and Events) 2018 development dataset. The proposed method achieves 73.52% and 75.76% accuracy with Polynomial and Gaussian Kernels respectively.

Index Terms: Kervolution-based SubSpectralNet, Acoustic Scene Classification (ASC), Convolutional Neural Network (CNN), Detection and Classification of Acoustic Scenes and Events (DCASE)

1. Introduction

Acoustic Scene Classification (ASC) is the process of assigning a semantic label to an audio recording that represents a specific environment/scene in nature. ASC has many applications such as context-aware systems [1], robotics [2], intelligent wearable systems [3] and management of audio archives [4]. In recent years, Convolutional Neural Network (CNN) or a combination of CNN with other Deep Neural Network (DNN) models have become the state-of-the-art ASC models [5–11]. For instance, a model known as SubSpectralNet [9] divides the mel spectrogram into horizontal slices, termed as sub-spectrograms and trains a CNN for recognition of acoustic scenes.

Kervolutional Neural Network (KNN) has been recently introduced for image classification [12–14] and it outperforms CNN-based systems by generalizing the linear convolutional operation to patch-wise (region-wise) non-linear operation using the kernel trick, also known as kervolution (kernel convolution) operation. A Kervolution-based SubSpectralNet model is being proposed in this paper as an improvement over the the performance of the SubSpectralNet [9] by replacing CNN with KNN. Rather than aiming for state-of-the-art results, our goal is to show how the non-linearity introduced by Kervolution could be used instead of the linear convolution operation in SubSpectralNet to understand acoustic scenes more effectively.

The rest of the paper is organized as follows - Section 2 contains an explanation about the proposed Kervolution-based SubSpectralNet architecture in detail. Experiments carried out in this work are given in section 3. The conclusions are given in section 4.

2. Kervolution-based SubSpectralNet

The proposed ASC system includes two parts: extraction of mel-band energies and Kervolution-based SubSpectralNet. These parts are explained below in brief.

2.1. Extraction of mel-band energies

Two steps are involved in the extraction of mel-band energies. First, Short-Time Fourier Transform (STFT) is computed from each acoustic scene recording using 40ms Hamming windowed frames with 50% overlap. Then, mel-band energies are extracted from each frame through mel filter banks with 40 bands. These mel-band energies are further normalized using zero mean and unit variance scaling. Finally, a mel spectrogram of size $2 \times 40 \times 500$ $(C \times F \times T)$ is obtained from extracted mel band energies from each channel, where *C* is the number of channels, *F* is the number of mel bands and *T* is the number of time frames.

2.2. SubSpectralNet

The SubSpectralNet proposed in [9] includes a combination of SubSpectral layer, convolutional layers and fully connected layers. We replace the convolutional layer of the SubSpectralNet with Kervolutional layer and experiment with different nonlinear functions for modelling the ASC task. Each of the layers of the proposed SubSpectralNet are described below in brief.

2.2.1. SubSpectral layer

Mel spectrograms of size $C \times F \times T$ are considered as input to the SubSpectral layer. This layer divides the spectrogram into *M* frequency-time horizontal slices called sub-spectrograms of $C \times X \times T$ dimension for every recording, where *X* is the sub-spectrogram size, Y is the mel-bin hop-size, and $M = [1 + (F - X)/Y]$.

2.2.2. Kervolutional layer

Convolutional operation is denoted using the notation given in (1).

$$Z = X \bigoplus f, \tag{1}$$

where $X \in \mathbb{R}^{F \times T}$ is a feature matrix, f is a filter and $\bigoplus$ is the convolutional operation. The output of an i^{th} element of a convolutional operation is denoted below in (2).

$$z_i = (x_i, f), \tag{2}$$

where $(*, *)$ denotes the inner product between two vectors. In the similar way, Kervolutional operation is denoted using nota-

[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-656

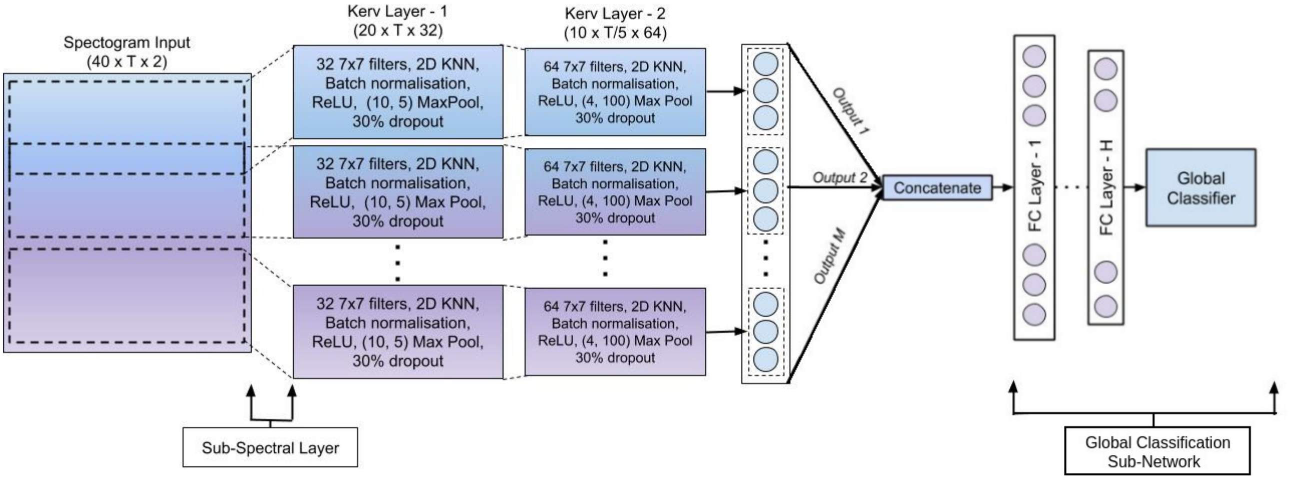

Spectogram Input
(40 x T x 2)
Kerv Layer - 1
(20 x T x 32)
Kerv Layer - 2
(10 x T/5 x 64)
32 7x7 filters, 2D KNN, Batch normalisation, ReLU, (10, 5) MaxPool, 30% dropout
32 7x7 filters, 2D KNN, Batch normalisation, ReLU, (10, 5) Max Pool, 30% dropout
32 7x7 filters, 2D KNN, Batch normalisation, ReLU, (10, 5) Max Pool, 30% dropout
64 7x7 filters, 2D KNN, Batch normalisation, ReLU, (4, 100) Max Pool 30% dropout
64 7x7 filters, 2D KNN, Batch normalisation, ReLU, (4, 100) Max Pool 30% dropout
64 7x7 filters, 2D KNN, Batch normalisation, ReLU, (4, 100) Max Pool 30% dropout
Output 1
Output 2
Output M
Concatenate
FC Layer - 1
FC Layer - H
Global Classifier
Sub-Spectral Layer
Global Classification Sub-Network

Figure 1: *Proposed architecture of the Kervolution-based SubSpectralNet*

tion given in (3).

$$Z = X \bigotimes f, \tag{3}$$

where $\bigotimes$ is the kervolutional operation. The output of an i^{th} element of a kervolutional operation is denoted as given in (4).

$$z_i = (\psi(x_i), \psi(f)), \tag{4}$$

where ψ is a non-linear function and it is computed using kernel trick given in (5) [9],

$$(\psi(x_i), \psi(f)) = \sum_j c_j (x_i^T f)^j = \kappa(x_i, f), \tag{5}$$

where c_j is the coefficient, which balances the order of non-linearity and κ is the kernel function, like the Polynomial or Gaussian kernel. In this work, the linearity of CNN is replaced with the Polynomial and the Gaussian kernel.

This non-linearity is introduced using the kernel trick so that we can establish a hyper-plane in a higher dimensional space into which the non-linear separable dataset can be mapped to [15, 16]. We only need to know κ and not the mapping function itself which means that the amount of computation is reduced by avoiding the math that converts the data from lower to higher dimensions. This trick of calculating high-dimensional relationships without actually transforming the data is what makes the kernel trick better than convolution for classification tasks.

The *Polynomial Kernel* (6) systematically increases dimensions by setting the value of *d*.

$$\kappa_g(x, f) = (x^T f + c_g)^d = \sum_{j=0}^{d} c_g^{d-j} (x^T f)^j, \tag{6}$$

where $d \in \mathbb{Z}^+$, and $c_g \in \mathbb{R}^+$ and is used for balancing of the non-linear orders.

The *Gaussian Kernel* has infinite dimensional features because of the *i*-degree terms given in (7)

$$\kappa_g(x, f) = exp(-\gamma_g ||x - f||^2) = C \sum_{i=0}^{\infty} \frac{(x^T f)^i}{i!}, \tag{7}$$

Here, $\gamma_g(\gamma_g \in \mathbb{R}^+)$ is a hyperparameter. The value of γ_g controls the smoothness of the decision boundary. In equation 7, $C = exp(-\frac{1}{2}(||x||^2 + ||f||^2))$ when $\gamma_g = \frac{1}{2}$.

Introduction of the kernel trick into the SubSpectralNet [9] takes convolution function to a non-linear space, increasing the model capacity without the introduction of extra parameters. Introduction of the kernel trick into the SubSpectralNet [9] takes convolution function to a non-linear space, increasing the model capacity without the introduction of extra parameters.

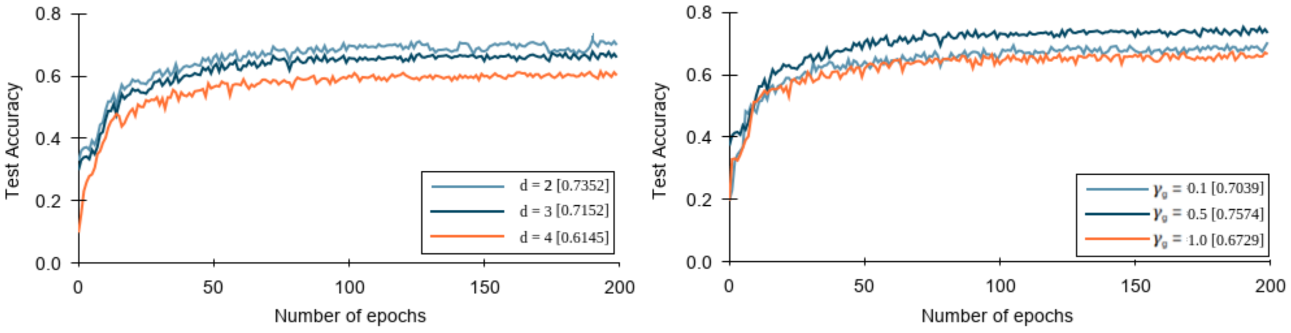
3. Experiments

3.1. Dataset

We use the *TUT Urban Acoustic Scenes 2018 development* dataset [17, 18]. The dataset consists of binaural audio recordings from 10 different acoustic scenes. The entire dataset has 8640 audio segments, out of which, 6122 are used for training and 2518 for testing. 20% of the training data is used for k-fold cross-validation ($k = 5$) in order to eliminate any stochastic factors during optimization and to ensure that the results obtained are not sensitive to initialization. The final accuracy of each configuration is calculated as the average of accuracy obtained at each fold. For feature extraction, we use *dcase_util* toolbox. Each audio segment is converted to signals having 40 mel-bin spectrograms.

3.2. SubSpectralNet configuration

The proposed SubSpectralNet framework is given in Figure 1. Since we are using binaural input, two-channel sub-spectrograms are connected separately to two kervolutional layers with kernel-size of (7, 7) and same padding, having 32 and 64 kernels respectively. Each kervolutional layer is followed by a batch normalization layer, an activation layer with ReLU function, max-pooling layers of dimensions $(X/10, 5)$ and $(4, 100)$ respectively, and a 30% dropout. The output from the second max-pooling function is flattened and a fully connected layer with 32 neurons and ReLU activation is added. This is followed



(a) (b)

Figure 2: *Comparison of performance of - (a) Polynomial Kervolution SubSpectralNet with degree of Polynomial 2, 3, and 4; (b) Gaussian Kervolution SubSpectralNet with value of γ_g equal to 0.1, 0.5, and 1.*

by a 30% dropout, and a softmax layer. These are called the sub-classifiers.

Finally, the fully connected (ReLU) layers of the sub-networks are concatenated and a Deep Neural Network is trained with H hidden layers with R_i neurons, where:

$$H = max([log_2(M)] - 1, 0); R_i = 2^{\text{6+H-i}}, 1 \leq i \leq H \quad (8)$$

This is done to capture the global correlation (or de-correlation) between frequency bands and is hence called the global classification sub-network.

3.3. Ablation Study

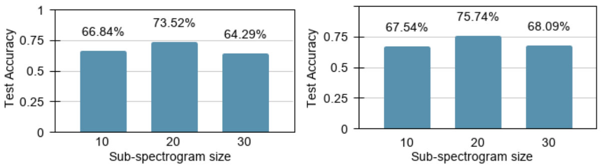
In this section, we investigate the effect of using different kernels and hyperparameters for kervolution; as well as different number of subspectrograms for the proposed architecture. As part of the ablation study, simple kervolution has been added to the baseline to see how it impacts the model and the performance has been tabulated in Table 1 along with detailed explanation in Section 3.5.

3.3.1. *Kernels and Hyperparameters*

As explained in 3.2, we replace the convolutional layers of the baseline model released along with the dataset [18] as well as SubSpectralNet, with kervolutional layers using two kernel function - Polynomial kernel and Gaussian kernel.

Prior experimentation shows that the Polynomial kernel works best with $c_g = 1$ [12]. Keeping all the other hyperparameters same as the baseline model, we repeat the experiments for three degrees of Polynomials, *i.e.* $d = 2$, $d = 3$ and $d = 4$. We plot a curve of the number of epochs versus test accuracy, comparing the performance of the Polynomial kervolution with different values of d in Figure 2a. The graph highlights that the model performs slightly better than CNN-based SubSpectralNet when the degree of Polynomial is two *i.e.* 73.52%, and the accuracy decreases as the value of d increases.

Similarly, in case of Gaussian kernel, we explore how classification performance varies in terms of γ_g, defined as the inverse of the radius of influence of samples. We consider three values of the hyperparameter, *i.e.* $\gamma_g = 0.1$, $\gamma_g = 0.5$ and $\gamma_g = 1$. Figure 2b shows the comparison in test accuracy for the three values of γ_g. When the value of $\gamma_g = 0.5$, we achieve the highest test accuracy of 75.76%, and the model performs poorly for $\gamma_g = 0.1$ and $\gamma_g = 1$, achieving an accuracy of 70.39% and 67.29% respectively. The reason for this variation is that when gamma is too large, the radius of the area of influence would only include a small set of nearby samples, resulting in a Gaussian function with a small variance. Conversely, when gamma is too small, the model is unable to capture the complexity or shape of the data as the region of influence would include the data points that are placed further away and should have less influence.



(a) (b)

Figure 3: *Results obtained by SubSpectralNet on a 40 mel-bin spectrogram and 10 mel-bin hop-size using - (a) Polynomial Kernel ($d = 2$) and (b) Gaussian Kernel ($\gamma_g = 0.5$)*

3.3.2. *Different number of sub-spectrograms*

In Section 3.3.1, we concluded that the Polynomial kernel works best with $d = 2$ and the Gaussian kernel works best with $\gamma_g = 0.5$. Keeping all other parameters same, we investigate the effect of size of sub-spectrogram on the model performance. As shown in Figure 3, considering sub-spectrograms of size 20 achieves a higher test accuracy of 73.52% and 75.76%, than sub-spectrograms of size 10 and 30 in Polynomial and Gaussian kernel respectively.

3.4. Performance comparison

The performance of the proposed Kervolution-based SubSpectralNet is compared with the following state-of-the-art models.

3.4.1. *DCASE 2018 baseline*

The baseline considers 40 monaural log mel band energies as input features. The architecture includes two convolutional layers with 32 and 64 filters respectively, followed by batch normalization, max-pooling and 30% dropout. The output feature map of the convolutional layer was considered as input to a fully connected layer with 100 neurons. The number of neurons in the output layer is equal to the number of acoustic scenes in the input dataset.

Table 1: *Performance comparison of the proposed Kervolution-based SubSpectralNet with DCASE 2018 Baseline [18] and CNN-based SubSpectralNet [9]*

Model	**Overall Accuracy**	**Class-wise Accuracy**									
		Airport	Bus	Metro	Metro Station	Park	Public Square	Shopping Mall	Street Pedestrian	Street Traffic	Tram
Using Convolutional Neural Network											
Baseline	58.3	58.5	66.5	51.3	52.1	74.4	50.9	59.5	46.6	80.1	42.9
SubSpectral-Net	72.18	63.77	77.69	69.35	69.88	88.02	57.41	78.85	71.66	91.46	56.32
Using Polynomial Kervolution ($d = 2$)											
Baseline	61.20	44.15	65.70	51.34	58.69	79.75	47.22	78.85	53.44	84.55	47.51
SubSpectral-Net	73.52	65.75	68.22	71.67	76.04	88.88	63.8	76.68	66.94	94.43	62.58
Using Gaussian Kervolution ($\gamma_g = 0.5$)											
Baseline	65.85	60.75	64.88	49.81	72.20	82.64	50.46	66.67	61.94	82.11	66.28
SubSpectral-Net	75.76	68.02	70.48	73.94	78.91	91.15	65.74	78.95	68.41	96.7	64.85

3.4.2. *CNN-based SubSpectralNet*

It considers 40 binaural mel band energies as input features to the independent SubSpectral layers. Subspectrograms from SubSpectral layers are considered as input to the two convolutional layers with 32 and 64 filters respectively, followed by batch normalization, max-pooling and 30% dropout. The output feature map of a convolutional layer was considered as input to a fully connected layer with 32 neurons. The number of neurons in the output layer is equal to the number of acoustic scenes in the input dataset. The above architecture is termed as sub-classifiers of the SubSpectralNet. Finally, the output of the fully connected layers of each sub-network on subspectrogram is concatenated to form a global classifier.

3.5. Results

Performance of the Kervolution-based baseline model and the proposed Kervolution-based SubSpectralNet has been compared with the DCASE 2018 baseline [18] and CNN-based SubSpectralNet [9] (see Table 1). It highlights that even without the introduction of subspectrograms, *i.e.*, using only kervolution, better results on the baseline model are achieved than convolution. Kervolution using both the polynomial and the Gaussian kernel has been done on the baseline and the results are better than the baseline in seven of the ten classes providing proof to the fact that kervolution is increasing the performance of the model without even using SubSpectralNet. This is because, convolution operation only captures linear features whereas kervolution operation captures non-linear features from input features maps and consequently the higher-order terms generate more discriminative features than linear convolution.

CNN-based SubSpectralNet largely outperforms the baseline system. However, we can observe that the proposed Kervolution-based SubSpectralNet outperforms all other methods. KNN using a Polynomial or Gaussian kernel captures higher-order interactions of the input sub-spectrograms from the SubSpectral layer. Hence, Kervolution-based SubSpectralNet is more suitable for acoustic scene classification as compared to CNN-based SubSpectralNet. Gaussian kernel performs better than Polynomial kernel in all classes as well as in terms of overall accuracy. This is because the function space of Gaussian Kernel is a lot larger than that of the Polynomial kernel, as it extends the kervolution function to infinite dimensions.

The class-wise performance of the Baseline Model [18], CNN-based SubSpectralNet [9], and our proposed Kervolution-based SubSpectralNet is also reported in Table 1. The proposed Kervolution-based SubSpectralNet recognizes the different acoustic scenes with an accuracy of more than 70% for most classes, out-performing the CNN-based SubSpectralNet in eight out of ten classes. We obtain a high accuracy of 73.52% using the Polynomial kernel, which is an almost +15% improvement over the baseline. The best accuracy achieved was 75.76%, using the Gaussian kernel ($\gamma_g = 0.5$) with sub-spectrograms of size 20 and mel-bin hop-size of 10, which is an overall increase of around +18% over the DCASE 2018 baseline and 4% over the SubSpectralNet architecture.

4. Conclusions

This paper proposes Kervolution-based SubSpectralNet for Acoustic Scene Classification, which replaces the convolutional layer of SubSpectralNet with a kervolutional layer. We also conduct an ablation study where we investigate the effect of using different kernels and hyperparameters for kervolution; as well as different number of subspectrograms for the proposed model.

Results show that the Kervolution-based SubSpectralNet outperforms the reported works in the literature. It also indicates that the proposed model has a significant contribution towards the Acoustic Scene Classification. The effectiveness of Kervolution-based SubSpectralNet can be inferred by noticing a relative improvement of around +18% accuracy over the DCASE 2018 baseline model.

In the future, we plan to improve the Kervolution-based SubSpectralNet by considering 200 or higher mel-bin spectrograms as [9] reports a higher accuracy using 200 mel-bin spectrograms. We further surmise that the performance of this model can be improved by exploring the effect of incorporating complex architectures [19–22] on the kervolution model.

5. References

[1] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Workshop on Mobile Computing Systems and Applications*. IEEE, 1994, pp. 85–90.

[2] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *IEEE International conference on multimedia and expo*. IEEE, 2006, pp. 885–888.

[3] Y. Xu, W. J. Li, and K. K. C. Lee, *Intelligent wearable interfaces*. Wiley Online Library, 2008.

[4] I. Damnjanovic, J. Reiss, and D. Barry, "Enabling access to sound archives through integration, enrichment and retrieval," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE Computer Society, 2008.

[5] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017, pp. 1–5.

[6] Z. Ren, Q. Kong, J. Han, M. D. Plumbley, and B. W. Schuller, "Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 56–60.

[7] A. M. Basbug and M. Sert, "Acoustic scene classification using spatial pyramid pooling with convolutional neural networks," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, 2019, pp. 128–131.

[8] K. Koutini, H. Eghbal-zadeh, G. Widmer, and J. Kepler, "CP-JKU submissions to DCASE'19: Acoustic scene classification and audio tagging with receptive-field-regularized cnns," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA*, 2019, pp. 25–26.

[9] S. S. R. Phaye, E. Benetos, and Y. Wang, "Subspectralnet–using sub-spectrogram based convolutional neural networks for acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 825–829.

[10] M. Mulimani, A. B. Kademani, and S. G. Koolagudi, "A deep neural network-driven feature learning method for polyphonic acoustic event detection from real-life recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 291–295.

[11] T. Heittola, E. Çakır, and T. Virtanen, "The machine learning approach for analysis of sound scenes and events," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 13–40.

[12] C. Wang, J. Yang, L. Xie, and J. Yuan, "Kervolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 31–40.

[13] M. A. Mahmoudi, A. Chetouani, F. Boufera, and H. Tabia, "Kernelized dense layers for facial expression recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2226–2230.

[14] J. Malik, S. Kiranyaz, and M. Gabbouj, "Operational vs convolutional neural networks for image denoising," *arXiv preprint arXiv:2009.00612*, 2020.

[15] P. Vincent and Y. Bengio, "K-local hyperplane and convex distance nearest neighbor algorithms," in *NIPS*, vol. 14, 2001, pp. 985–992.

[16] M. Hofmann, "Support vector machines-kernels and the kernel trick," *Notes*, vol. 26, no. 3, pp. 1–16, 2006.

[17] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 147–179.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop*, 2018, pp. 9–13.

[19] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," 05 2017.

[20] S. S. R. Phaye, A. Sikka, A. Dhall, and D. R. Bathula, "Multi-level dense capsule networks," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 577–592.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.