CnnSound: Convolutional Neural Networks for the Classification of Environmental Sounds

Özkan İnik, Hüseyin Şeker

***Abstract*—Environmental sounds classification (ESC) has been increasingly studied in recent years. The main reason for this is that these ESC missions are being used widely in our lives. Especially, ESC is used in areas such as managing smart cities, determining location from environmental sounds, surveillance systems, machine hearing, environment monitoring. Classification of these sounds is more difficult than other sounds because there are too many parameters that generate noise in the ESC. In the proposed study, it has been tried to find the most suitable convolutional neural networks (CNN) model for ESC task. For this purpose, 150 different CNN-based models were designed by changing number of layers and values of their tuning parameters used in the layers. In order to test accuracy of the models, the Urbansound8k environmental sound database was used. The sounds in this data set were first converted to an image format of 32x32x3. The proposed CNN-driven model has yielded an accuracy of as much as 82% being higher than its classical counterpart. As there was not that much fine-tuning, the obtained predictive accuracy has been found to be better and satisfactory compared to other studies on the Urbansound8k when both accuracy and computational complexity are considered. The results also suggest further improvement possible in its accuracy due to low complexity of the proposed CNN architecture and its applicability in real-world settings.**

***Index Terms*— Environmental Sound Classification (ESC), Deep Learning, Convolutional Neural Networks (CNN), Urbansound8k.**

I. Introduction

Sound data contains more semantic information than visual data [1]. In particular, sound data becomes more important to obtain information about an environment. In order to realize some applications in daily life, it is necessary to use environmental sounds, unlike speech and music sounds. For this reason, studies on the classification of urban sounds have intensified in recent years. Environmental sounds Classification (ESC), is known as one of the most important issues of the non-speech voice classification task [2]. ESC is of critical importance in many problems such as; noise pollution analysis [3, 4], surveillance systems [5-7], context-aware applications [1, 8-13], machine hearing [14-17], environment monitoring [18], crime alert systems [19], soundscape assessment [20, 21], and smart city [22, 23]. Different data sets have been created for ESC task. ESC-10, ESC-50[24] and Urbansound8k (US8K) [25] datasets are used extensively. Different statistical and machine learning methods have been used for ESC task in the literature [1, 26-33].

The success rates of these methods are relatively low compared to deep learning-based studies in recent years. Deep learning [34] achieved a high success rate in the ImageNet [35] competition in 2012. Due to this success, deep learning models for ESC have been used frequently in recent years, as they have been used in different fields [36-49]. In general, it has been observed that the success rates obtained with deep learning models have better results than other artificial intelligence methods. The main reason for this can be summarized as automatic feature discovery in deep learning models. Recently, it is seen that CNN models [2, 36-38, 41-46, 48-50] are used for ESC task. There are a lot of parameters that need to be adjusted in the design of CNNs. Therefore, the best CNN model can be found in different layer depths and different parameters. In this study, try to find the suitable CNN model for ESC task. The suitable layer number and layer parameters were obtained for CNN. The designed CNN model has been found to perform well in the ESC task compared to most previous work.

This paper is organized as follows. In section 2, information about the features of the Urbansound8k ESC data set is given. In section 3, information about the proposed CNN model is given. Experimental studies have been conducted in section 4. Finally, the conclusion is explained in section 5.

II. Data Set

In this study, Urbansound8k [25] data set is used for ESC task. Urbansound8k data set was obtained from real environment according to 4 seconds recording time. Environmental noise is present in the records obtained. The data set consists of 10 classes. These classes are respectively; air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music. These sounds are transformed into images with the method of scalogram. The scalogram is the absolute value of the continuous wavelet transform (CWT) of a signal plotted as a function of time and frequency. Wavelet Toolbox of Matlab R2020b software was used for the conversion process. The transformed form of each class in the data set into sound signals and images is given in Figure 1. There are a total of 8732 records in the data set. The image resolution for training the CNN model is set to 32x32x3. 80% of the dataset was used for training, 10% for validation and the remaining 10% for testing. The total number of images of each class for training, validation and testing are given in Table 1. For more information on the data set, look at the reference [51].

Manuscript received October 10, 2020

Özkan İnik is with the Department of Computer Engineering, Gaziosmanpasa University, Tokat, Turkey (e-mail: ozkan.inik@gop.edu.tr).

Hüseyin Şeker is with the School of Computing and Digital Technologies, Staffordshire University, Stoke-on-Trent, The United Kingdom (e-mail: huseyin.seker@staffs.ac.uk).

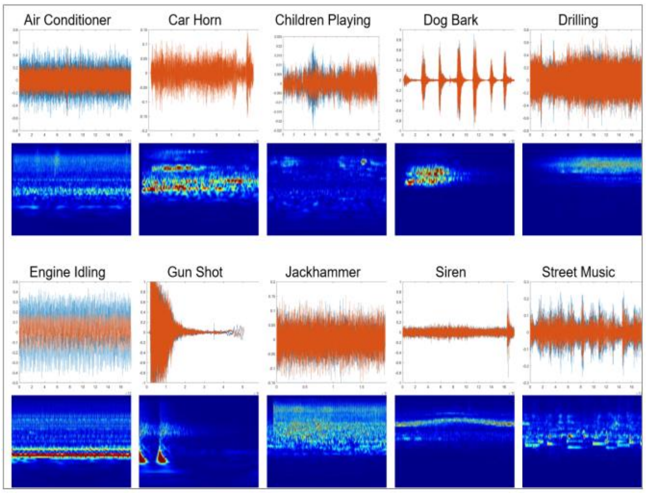


Figure 1. Sound(up) and translated image according to scalogram (down) of classes in Urbansound8k data set.

TABLE 1. NUMBER OF RECORDS USED FOR TRAINING, VERIFICATION AND TESTING IN THE URBANSOUND8K DATA SET

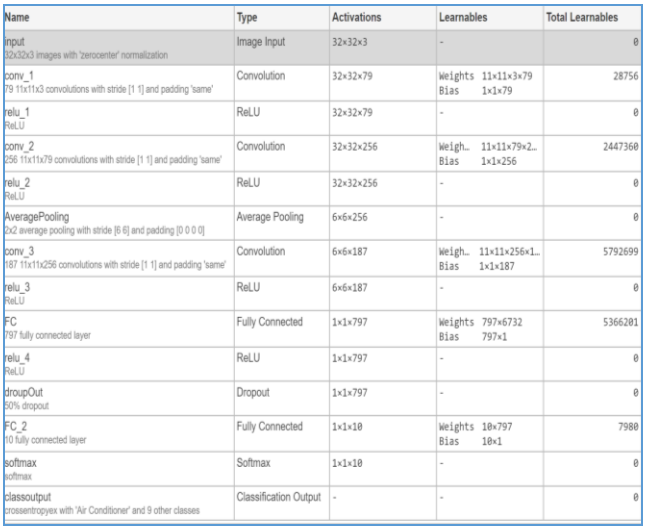
Class	Number of images	Train	Validation	Test
Air Conditioner	1000	600	200	200
Car Horn	429	257	86	86
Children Playing	1000	600	200	200
Dog Bark	1000	600	200	200
Drilling	1000	600	200	200
Engine Idling	1000	600	200	200
Gun Shot	374	224	75	75
Jackhammer	1000	600	200	200
Siren	929	557	186	186
Street Music	1000	600	200	200

III. PROPOSED METHOD

In this study, the most suitable CNN model for the ESC task is tried to be obtained by grid search. 1 CNN models were designed and trained according to the layer depth and the parameter values used in the layers. The layer structure of the model that gives the best result among these models and the parameter values used in the layers are given in Figure 2. Looking at Figure 2, the proposed CNN model consists of 3 convolution layers, 1 pooling layer and 2 fully connected layers. There are 79 filters in the first convolution layer of the proposed model. After the training, the feature maps created by these filters and the effect of the filters on the input image is given in Figure 3.

IV. EXPERIMENTAL STUDIES

Experimental studies have done on a computer with Intel® Core™ i9-7900X 3.30GHz×20 processor, 64 GB Ram and 2 x GeForce RTX2080Ti graphic card. Matlab R2020a 64bit (win64) has used as the software platform. The parameters used for the training of CNN model are given in Table 2.



Name	Type	Activations	Learnables	Total Learnables
input 32x32x3 images with 'zerocenter' normalization	Image Input	32×32×3	-	0
conv_1 79 11x11x3 convolutions with stride [1 1] and padding 'same'	Convolution	32×32×79	Weights 11×11×3×79 Bias 1×1×79	28756
relu_1 ReLU	ReLU	32×32×79	-	0
conv_2 256 11x11x79 convolutions with stride [1 1] and padding 'same'	Convolution	32×32×256	Weigh... 11×11×79×2... Bias 1×1×256	2447360
relu_2 ReLU	ReLU	32×32×256	-	0
AveragePooling 2x2 average pooling with stride [6 6] and padding [0 0 0 0]	Average Pooling	6×6×256	-	0
conv_3 187 11x11x256 convolutions with stride [1 1] and padding 'same'	Convolution	6×6×187	Weigh... 11×11×256×1... Bias 1×1×187	5792699
relu_3 ReLU	ReLU	6×6×187	-	0
FC 797 fully connected layer	Fully Connected	1×1×797	Weights 797×6732 Bias 797×1	5366201
relu_4 ReLU	ReLU	1×1×797	-	0
droupOut 50% dropout	Dropout	1×1×797	-	0
FC_2 10 fully connected layer	Fully Connected	1×1×10	Weights 10×797 Bias 10×1	7980
softmax softmax	Softmax	1×1×10	-	0
classoutput crossentropyex with 'Air Conditioner' and 9 other classes	Classification Output	-	-	0

Figure 2. Architecture of the proposed CNN model and parameter information used in each layer

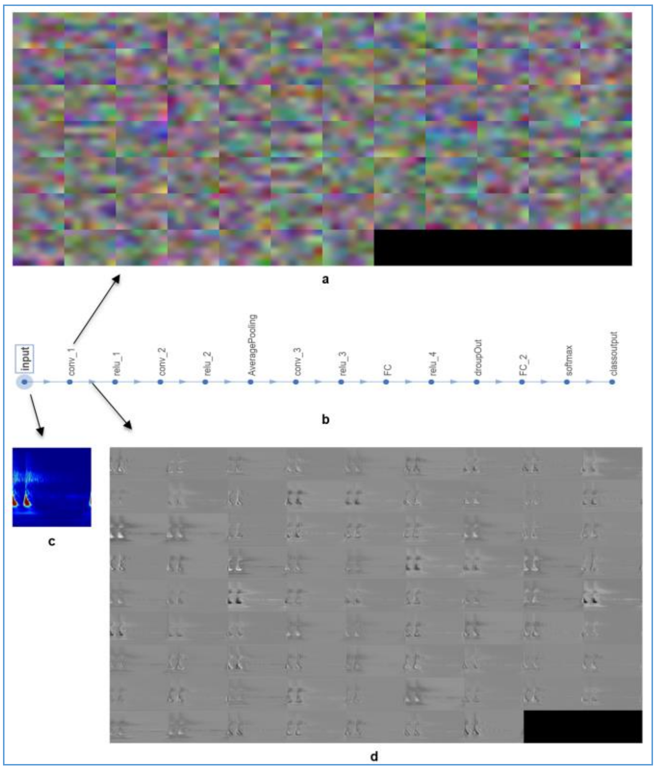


Figure 3. Layer architecture of the proposed CNN model (b), trained filters in the first convolution layer(a), the test image of the Gun shot class (c), the effect of each filter in the first convolution layer on the input image(d).

TABLE 2. CNN TRAINING PARAMETERS

Parameters	Value
Optimizer	SGDM
Epochs	50
Learning rate drop factor	0.1
Learning rate drop period	10
Dropout rate	0.5
Mini Batch Size	256
Initial learning rate	0.001
Validation Frequency	50

In the studies, the accuracy of the most suitable CNN model was obtained as 82.26%. The confusion matrix obtained by this model is given in Figure 4. When confusion matrix is examined, it is seen that the most confused class with each other are Children Playing and Street Music. While the highest classification performance was achieved in the Car Horn class, the lowest classification success was achieved in the Engine Idling class. The graph of accuracy and validation values according to the epoch in the training phase of the CNN model is given in Figure 5 and accuracy and validation loss graph is given in Figure 6. When Figure 5-6 are examined, it is seen that the model at the training stage reach the optimum performance approximately after the 15th epoch.

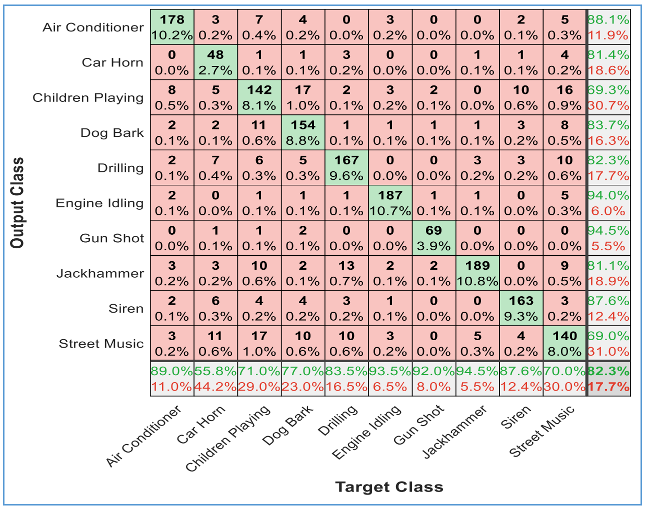


Figure 4. Confusion matrix of proposed CNN model

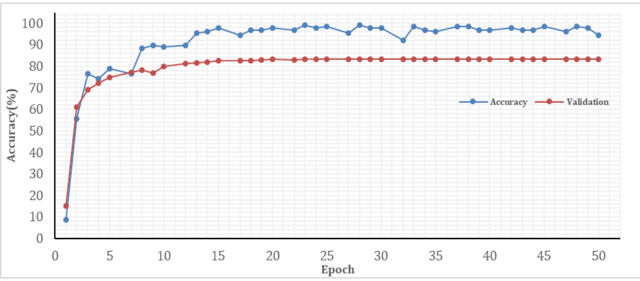


Figure 5. Accuracy and validation of the proposed CNN model for training.

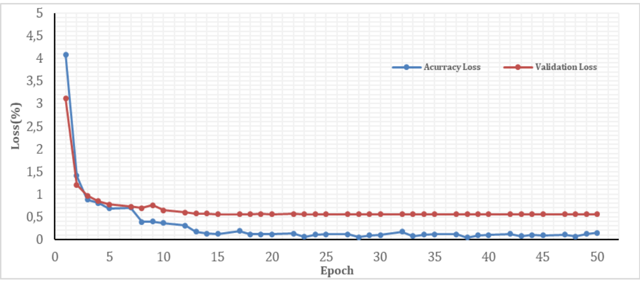


Figure 6. Accuracy loss and validation loss graph of the proposed CNN model during training.

A. Comparison with other studies

Different studies based on deep learning have been conducted on the ESC data set. Accuracy values obtained by some previous studies are given in Table 3. Looking at Table 3, it is seen that the proposed CNN method achieves a very good performance. The method only performed lower than GoogLenet and AlexNet. The reason for this is related to the image size obtained during the transformation of the data set. GoogLenet input image size is 224x224x3 and AlexNet input image size is 227x227x3. In the proposed CNN models, the input image size is 32x32x32. The large input image size causes the model to discover more features. Thus, it enables the model to be more successful.

TABLE 3. COMPARISON OF THE ACCURACY VALUE OBTAINED BY THE PROPOSED METHOD WITH OTHER METHODS

Method	Accuracy(%)
GooLeNet and AlexNet [40]	93
Proposed method(CnnSound)	**82.26**
D-CNN(Activation functions=LeakyReLU) [48]	81.9
CNN [21]	81.5
D-CNN(Activation functions= PReLU) [48]	81.4
D-CNN(Activation functions= ReLU) [48]	81.2
DNN [20]	79.23
SoundNet [52]	79
DCNN + augmentation SB-CNN (DA) [37]	79
D-CNN(Activation functions= ELU) [48]	78.9
EnvNet-v2 + augmentation[39]	78.3
Pyramid-Combined CNN[2]	78.1
EnvNet-v2 (Tokozume et al., 2017)[39]	78
Dilated CNN [45]	78
DCNN [53]	77.36
Unsupervised feature learning SKM (DA)[30]	76
Convolutional layers with max-pooling[36]	74
SKM[30]	74
Deep CNN[37]	74
D-CNN(Activation functions= Softplus) [48]	73.7
CNN (Baseline model) [36]	73.7
Unsupervised feature learning SKM [30]	73.6
M18 CNN (Dai et al., 2017)[54]	72
VGG (Pons & Serra, 2018)[55]	70
SVM [25]	71
Very Deep CNN[54]	69.38
Baseline system[25]	68
SVM[56]	62.4
ANN, KNN + features cascading + optimization[57]	56.4

V. CONCLUSIONS

In this study, the most suitable CNN model was obtained with grid search for the classification of environmental sounds. For this purpose, 150 CNN models have been designed and tested over Urbansound8k environmental sounds data set. Among these methods developed, the best performing CNN model (CnnSound) has achieved 82.26% predictive accuracy. When compared with similar studies in the literature, it has been observed that the CnnSound model has a satisfactory performance and there is room for improvement further research will be geared towards further improvement through pre-processing methods, sound representation, optimization methods and further fine-tuning of CNN models. This is further expected to be studied along with other sound libraries to further demonstrate robustness of the deep learning-based frameworks being developed and adapted into sound modelling and classification.

REFERENCES

[1] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, pp. 1142-1158, 2009.

[2] F. Demir, M. Turkoglu, M. Aslan, and A. Sengur, "A new pyramidal concatenated CNN approach for environmental sound classification," Applied Acoustics, vol. 170, p. 107520, 2020.

[3] P. Aumond, C. Lavandier, C. Ribeiro, E. G. Boix, K. Kambona, E. D'Hondt, et al., "A study of the accuracy of mobile technology for measuring urban noise pollution in large scale participatory sensing campaigns," Applied Acoustics, vol. 117, pp. 219-226, 2017.

[4] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P.-P. Vidal, "Urban noise recognition with convolutional neural network," Multimedia Tools and Applications, vol. 78, pp. 29021-29041, 2019.

[5] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005., 2005, pp. 158-161.

[6] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," ACM Computing Surveys (CSUR), vol. 48, pp. 1-46, 2016.

[7] P. Laffitte, Y. Wang, D. Sodoyer, and L. Girin, "Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation," Expert systems with applications, vol. 117, pp. 29-41, 2019.

[8] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in 2010 18th European Signal Processing Conference, 2010, pp. 1272-1276.

[9] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 4, pp. 1-23, 2008.

[10] A. Waibel, H. Steusloff, and R. Stiefelhagen, "CHIL-Computers in the human interaction loop. 5th Intern," in Workshop on Image Analysis for Multimedia Interactive Services, 2004.

[11] D. P. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences, 2004, pp. 39-47.

[12] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, et al., "Audio-based context recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, pp. 321-329, 2005.

[13] 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, pp. 16-34, 2015.

[14] H. Li, S. Ishikawa, Q. Zhao, M. Ebana, H. Yamamoto, and J. Huang, "Robot navigation and sound based position identification," in 2007 IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 2449-2454.

[15] R. F. Lyon, "Machine hearing: An emerging field [exploratory dsp]," IEEE signal processing magazine, vol. 27, pp. 131-139, 2010.

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

[17] J. Huang, "Spatial auditory processing for a hearing robot," in Proceedings. IEEE International Conference on Multimedia and Expo, 2002, pp. 253-256.

[18] M. Green and D. Murphy, "Environmental sound monitoring using machine learning on mobile devices," Applied Acoustics, vol. 159, p. 107041, 2020.

[19] P. Intani and T. Orachon, "Crime warning system using image and sound processing," in 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), 2013, pp. 1751-1753.

[20] A. J. Torija, D. P. Ruiz, and Á. F. Ramos-Ridao, "A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model," Science of the Total Environment, vol. 482, pp. 440-451, 2014.

[21] V. P. Romero, L. Maffei, G. Brambilla, and G. Ciaburro, "Modelling the soundscape quality of urban waterfronts by artificial neural networks," Applied Acoustics, vol. 111, pp. 121-128, 2016.

[22] A. Agha, R. Ranjan, and W.-S. Gan, "Noisy vehicle surveillance camera: A system to deter noisy vehicle in smart city," Applied Acoustics, vol. 117, pp. 236-245, 2017.

[23] S. Ntalampiras, "Universal background modeling for acoustic surveillance of urban traffic," Digital Signal Processing, vol. 31, pp. 69-78, 2014.

[24] K. J. Piczak, "ESC: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015-1018.

[25] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 1041-1044.

[26] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, pp. 1216-1229, 2017.

[27] 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, pp. 1733-1746, 2015.

[28] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using AANN and GMM," Applied soft computing, vol. 11, pp. 716-723, 2011.

[29] J. Ludena-Choez and A. Gallardo-Antolin, "Acoustic Event Classification using spectral band selection and Non-Negative Matrix Factorization-based features," Expert Systems with Applications, vol. 46, pp. 77-86, 2016.

[30] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 171-175.

[31] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in The Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), 2015, pp. 714-718.

[32] M. Mulimani and S. G. Koolagudi, "Segmentation and characterization of acoustic event spectrograms using singular value decomposition," Expert Systems with Applications, vol. 120, pp. 413-425, 2019.

[33] J. Xie and M. Zhu, "Investigation of acoustic and visual features for acoustic scene classification," Expert Systems with Applications, vol. 126, pp. 20-29, 2019.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[35] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "Imagenet large scale visual recognition competition 2012 (ILSVRC2012)," See net. org/challenges/LSVRC, p. 41, 2012.

[36] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1-6.

[37] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Processing Letters, vol. 24, pp. 279-283, 2017.

[38] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," arXiv preprint arXiv:1604.07160, 2016.

[39] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," arXiv preprint arXiv:1711.10282, 2017.

[40] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," Procedia computer science, vol. 112, pp. 2048-2056, 2017.

[41] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," Applied Sciences, vol. 8, p. 1152, 2018.

[42] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," Sensors, vol. 19, p. 1733, 2019.

[43] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," Applied Acoustics, vol. 167, p. 107389, 2020.

[44] Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," Applied Acoustics, vol. 172, p. 107581, 2021.

[45] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, "Environmental sound classification with dilated convolutions," Applied Acoustics, vol. 148, pp. 123-132, 2019.

[46] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," Expert Systems with Applications, vol. 136, pp. 252-263, 2019.

[47] F. Medhat, D. Chesmore, and J. Robinson, "Masked Conditional Neural Networks for sound classification," Applied Soft Computing, vol. 90, p. 106073, 2020.

[48] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in 2017 22nd International Conference on Digital Signal Processing (DSP), 2017, pp. 1-5.

[49] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J.-S. Park, et al., "Convolutional Neural Network based Audio Event Classification," KSII Transactions on Internet and Information Systems, Vol. 12, No.6, 2018.

[50] Z. Zhang, S. Xu, T. Qiao, S. Zhang, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," in Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2019, pp. 261-271.

[51] K. Chng, "Classify Urban Sound using Machine Learning & Deep Learning

(https://github.com/KevinChngJY/classifyurbansound_matlab), GitHub. Retrieved September 28, 2020.," 2020.
[52] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in Advances in neural information processing systems, 2016, pp. 892-900.
[53] J. Ye, T. Kobayashi, and M. Murakawa, "Urban sound event classification based on local and global features aggregation," Applied Acoustics, vol. 117, pp. 246-256, 2017.
[54] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 421-425.
[55] J. Pons and X. Serra, "Randomly weighted CNNs for (music) audio classification," in ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2019, pp. 336-340.
[56] A. Kumar and B. Raj, "Features and kernels for audio event recognition," arXiv preprint arXiv:1607.05765, 2016. (https://arxiv.org/abs/1607.05765)
[57] B. da Silva, A. W Happi, A. Braeken, and A. Touhafi, "Evaluation of Classical Machine Learning Techniques towards Urban Sound Recognition on Embedded Systems," Applied Sciences, vol. 9, p. 3885, 2019.

Özkan İnik received B.Eng degree in computer engineering from Selçuk University, Konya, Turkey, in 2010, an the Ph.D in Computer Engineering from Konya Technical University, Konya, Turkey in 2019. During his doctorate, he worked on classification/detection/segmentation of medical images with deep learning methods. He has more than one articles during his doctoral education. He is currently working at Gaziosmanpaşa University, Department of Computer Engineering as associate professor, Tokat, Turkey. His current research interests include machine learning, deep learning, optimization techniques and computer vision.

Hüseyin Şeker is a Professor of Computing Sciences and Associate Dean responsible for Research and Enterprise activities for the School of Computing and Digital Technologies of Staffordshire University, Stoke-on-Trent, UK. He has academic and industry experiences in artificial intelligence, machine learning, data science and emerging & disruptive technologies/systems. His current research interests include artificial intelligence, data science and machine learning and their interdisciplinary applications (http://smartdatacrew.com/).