**Classification of Indoor and Outdoor Scenes with Acoustic Scene Classification**

Ishrat Jahan Ananya, Sarah Suad, Shadab Hafiz Choudhury, Mohammad Ashrafuzzaman Khan

ishrat.jahan16@northsouth.edu, sarah.suad@northsouth.edu, shadab.choudhury@northsouth.edu, mohammad.khan02@northsouth.edu

**Abstract.** Acoustic Scene Classification is the process of recognizing and classifying the source of an audio signal using data from recorders or microphones. As it can be used in many applications, this paper explores the use of convolutional neural networks to classify audio signals from a limited dataset of sounds from common urban scenes. A secondary goal was to utilize the most lightweight algorithm feasible, since most applications of acoustic scene classification would involve relatively low-powered devices such as mobile phones. The audio was converted into spectrograms. Several different models were tested, but due to the limited size of the dataset, overfitting was a major problem with large models. Ultimately, the ShuffleNet model provided the best results in terms of both accuracy and weight.

**Keywords:** Acoustic Scene Classification, Convolutional Neural Network, Deep Learning, ResNet, EfficientNet, ShuffleNet,

**1. Introduction**

Two of the primary senses human beings use to navigate their environment are sight and sound. Audio classification has uses similar to those of computer vision and is one of the most important areas of artificial intelligence.

The human brain is capable of recognizing the source and features of multiple overlapping sounds from various distances simultaneously. The volume, pitch and sequence of these sounds are all acoustic features. From the combination of these features, an individual can easily recognize what environment they are in. The acoustic features of a busy street may include a high level of background noise from footsteps, conversations and vehicles, punctuated by short bursts of loud noises such as from car horns or shouts. By contrast, a shopping mall will have a far lower overall volume of background noise but may have more distinct conversations, music, advertisements and other features.

Acoustic Scene Classification is the process of teaching an artificial intelligence system to recognize acoustic features and classify the scene of the sound based on them. It can be used to detect and understand the context of a scene from multimedia recordings. It can be used in various autonomous systems such as monitoring systems, robots and self-driving vehicles.

It is a relatively newer field of study compared to some others, as large scale datasets on the topic were not available until the annual Detection and Classification of Acoustic Scenes and Events (DCASE) challenges began in 2013.

Convolutional Neural Networks (CNNs) are one of the best contemporary solutions for feature extraction and classification. There are a large number of different models that can be used. Though there are many exceptions, a general idea is that the accuracy of a CNN can be increased by increasing the number of parameters and the depth of the model.

In this paper, we examine three recent and popular models of different sizes that are used in a large number of deep learning applications. The goal was to determine which model would be best suited for the task of acoustic. As these models are used in many other applications, understanding their relative performance in the domain of Acoustic Scene Classification would be useful.

**2. Related Work**

Analysis and classification of auditory signals with artificial intelligence have a long history. Initially, research work was focused on simply detecting and distinguishing acoustic events [4] such as distinct noises like claps and speech, or different individuals speaking [5]. These early examples of the use of neural networks in classification of audio developed from the intersection of signal processing and artificial intelligence.

More mature artificial intelligence techniques such as sophisticated convolutional neural networks have enabled further exploration of Acoustic Scene Classification through different approaches. The DCASE Challenges, initially started in 2013, offer datasets and a platform for exploration of Acoustic Scene Classification [6]. DCASE 2013 highlighted the use of large datasets for acoustic scene classification in various scenes such as a bus, office, market, et cetera.

Early research yielded good outcomes with machine learning models. Good results were achieved in the DCASE 2013 challenge using algorithms such as support vector machines and decision trees [7]. However, as the sophistication and size of datasets increased, neural networks became an effective choice. In this specific dataset, Valenti et al’s approach using a custom CNN model resulted in higher accuracy compared to earlier work - up to 9.7% depending on the technique it is compared to [8].

The majority of recent work on acoustic scene classification has followed up on the CNN approach. Hussein et al. developed a more indepth technique using a deep neural network with only 3 hidden layers that achieved up to 90% accuracy on the DCASE 2016 challenge [9].

In the DCASE 2020 challenge, several attempts were able to reach 96% test accuracy by implementing modern deep convolutional neural networks such as ResNets [10].

Finally, an excellent overview of the development and use of deep learning in acoustic scene classification between 2013 and 2020 is given in a review paper by Abeßer [11].

**3. Methodology**

While it is possible to carry out deep learning on audio signals directly, it is both more effective and efficient to use spectrograms [13]. The initial data was converted from audio to images, as CNNs are suitable for image classification.

The CNN models utilized in this experiment were ResNet-18, ResNet-50, EfficientNet and ShuffleNet. Alterations to the structure of the models were kept to a minimum, as a fair comparison could not be made otherwise.

**3.1. Data Organization and Collection**

The dataset was collected from DCASE 2020 challenge [1]. It had three classes of indoor, outdoor and transport. These classes were subdivided into nine more subclasses. The raw dataset contained 10sec audio files in 24-bit .wav format taken from ten different cities in the world.

For this paper, we chose to reduce it to two classes: indoor and outdoor. This left us with six subclasses, three from each. For Indoor Scenes, the subclasses were Metro, Shopping Mall and Airport. For Outdoor Scenes, the subclasses were Park, Pedestrian Street and Public Square.

There were 8328 data samples in tota, adding up to 23 hours of audio. Each subclass had 1388 data points. We ensured that the number of audio samples for each class was equal so that the model was not biased towards any particular category due to unbalanced data. The original audio files were binaural at 44.1 kHz.

**3.2. Data Preprocessing**

The task of data preprocessing requires taking audio samples as input and extracting features from the audio signals. The feature to be extracted from each waveform is known as Mel Frequency Cepstral Coefficients (MFCC) [3]. By performing feature extraction we aim to find components of the audio signals that will help us differentiate it from other signals.

A typical visualization of an audio file most people are familiar with is the waveform plot. This type of graph plots the amplitude of the audio signal against the time. However, there is not enough information in this type of graph to properly utilize a deep learning model. By contrast, extracting the MFCCs from audio files enables the audio to be represented as a spectrogram, which contains a lot more detail. The following figures show the difference between the two.

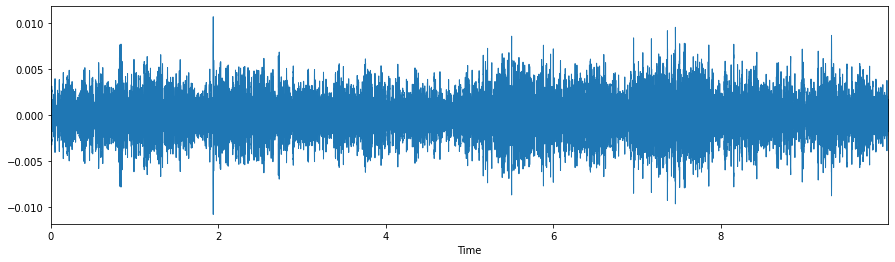


Fig 1. The Waveform Plot for an audio clip recorded in an Airport in Barcelona.

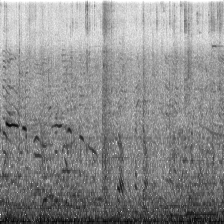


Fig 2. The Spectrogram for the same audio clip.

The process of MFCC calculation requires that the audio signal be sliced into short time frames that allow us to assume that the audio signal has had little to no change. The signal was typically sliced at 20-40ms. The power spectrum for each frame was then calculated. This allowed us to see the distribution of power into frequency components that make up the signal. These steps were accomplished by applying a short-time fourier transform on each frame.

We then applied the mel filter bank to each power spectra and sum the energy in each filter. This is a necessary step to estimate how the human ear perceives sounds at different frequencies and different volumes. The last step in the calculation is to take the discrete cosine transform of the logarithm of each filterbank. However, in practice we are able to perform this whole process with ease through the means of libraries.

**3.3. Development of CNNs**

Since the dataset was somewhat small, we made use of various data augmentation techniques to increase the amount of new data to be trained on in every batch. We implemented a number of image transformation techniques from the torchvision library such as resizing, random horizontal flip and rotating images. Furthermore, the imgaug library [2] was used to add more complex transformations that altered the existing images in the dataset in several ways.

The data augmentation transformations used here are: affine transformation, Gaussian blur and saturation changes.

**3.3.1. ResNet**

The ResNet family of neural network architectures is ideal for image classification tasks. The ResNet architecture uses stacked layers of residual learning blocks using shortcuts between layers to minimize the effect of the vanishing gradient problem [11].

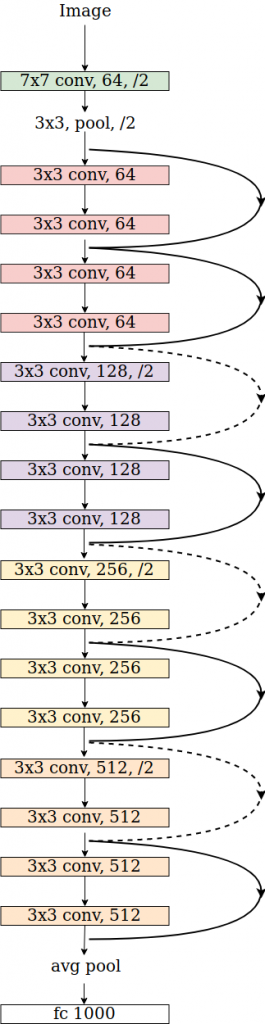


Fig. 3 - ResNet-18

Two ResNet models were implemented: ResNet-50 and ResNet-18. The loss function used was Cross Entropy Loss and the Learning rate was set to 0.02. The results of both models are discussed in the Evaluation section.

3.3.2. **EfficientNet**

EfficientNets are a family of convolutional neural network models that scale dimensional parameters up or down depending on the amount of resources available. This enables EfficientNet models, which range from B0 to B7 in ascending order of size, to be applied to a variety of situations. EfficientNets also implement the same reverse residual block that is utilized in MobileNet for increased efficiency.

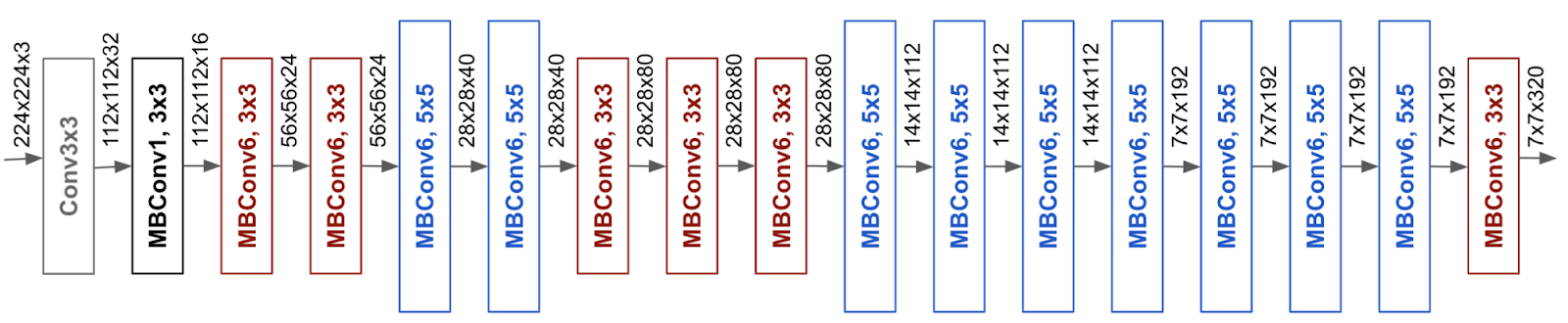


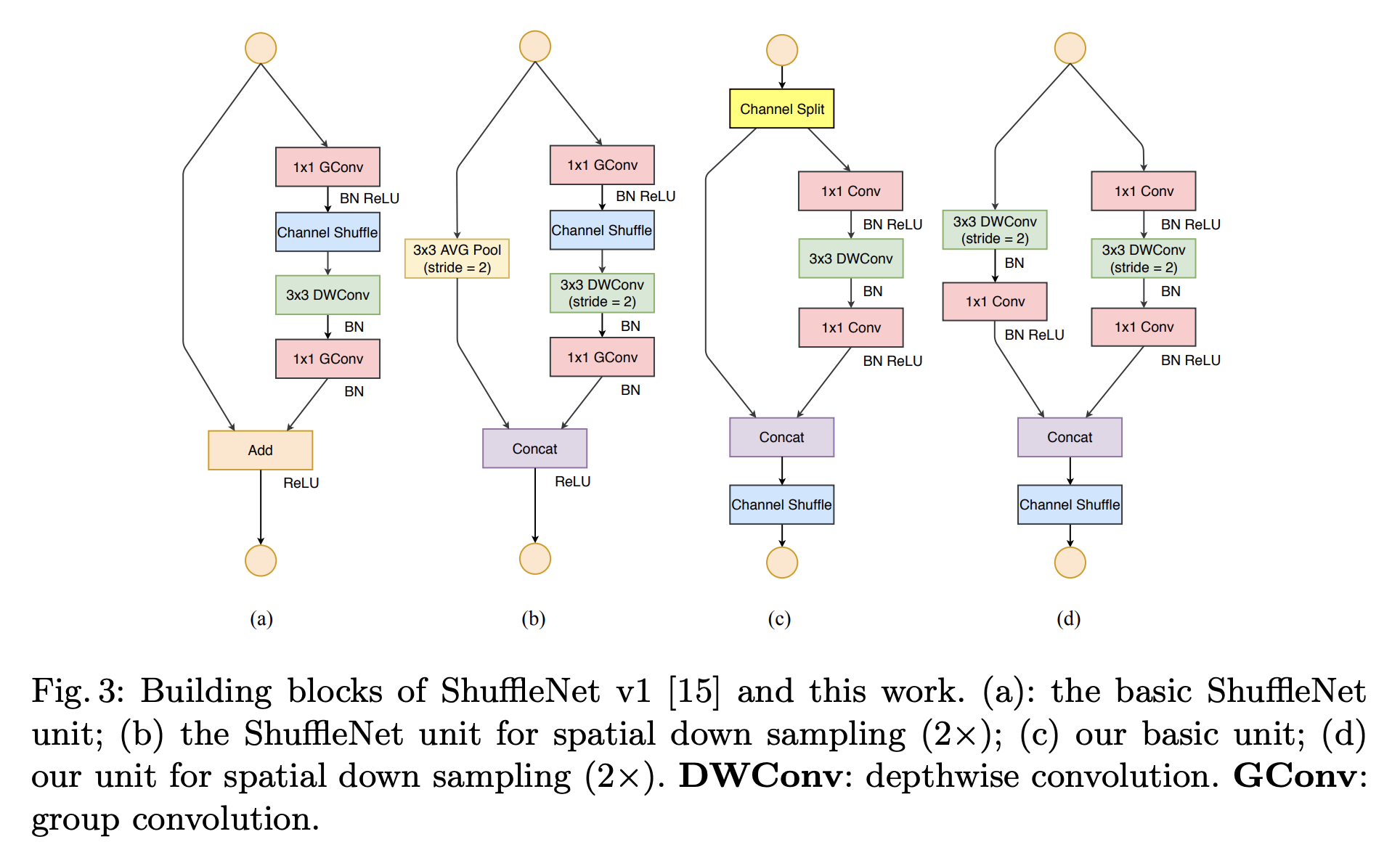
Fig. 3 - EfficientNet

In all image recognition tasks, EfficientNet shows better performance than ResNet while having a smaller size and lower computational requirements. Therefore, we also tested the application of EfficientNet in this task.

**3.3.3. ShuffleNet**

ShuffleNets are computationally efficient CNN architectures. They are particularly designed for mobile devices with limited computing power [14]. In order to maintain accuracy while simultaneously reducing computation cost, Shufflenet makes use of two operations, namely pointwise group convolution and channel shuffle. We use pointwise group convolution and bottleneck like structures to increase the number of channels without increasing FLOPs. Using channel shuffle we improve information communication between different groups of channels. This will improve our accuracy.

We implemented ShuffleNetV2 which has only one difference from ShuffleNetV1, the additional 1x1 convolution layer added right before global average pooling to mix up features.



**4. Evaluation**

We implemented the resnet50 model from scratch instead of using a pretrained model because the existing ones are mostly pretrained on the Imagenet dataset which is very different from ours. The pretrained model , once implemented, produced an incredibly low score of 60% on train data and 21% on test data. Hence we switched to running the resnet50 model from scratch.Considering the size of the architecture, the model has been a little too heavy for the size of the dataset. Thus leading to an accuracy of 100% after running for 40 epochs.This means the model has overfitted on the training data. We then implemented further image augmentation methods discussed above but the problem remained. We quickly learned that the next step was to implement a lighter model that best suited the dataset size.

Resnet18 has less than half the number of layers than resnet50 but it produced the same problem as mentioned above. The dataset being too small affected the test accuracy even after extending the dataset using data augmentation.

The next step was to implement Efficientnet with the same parameters and running it for 40 epochs. Albeit, performing comparatively better the model also dropped to 0 loss but at a slower rate, leading to overfitting again. The accuracy was 100% even though Efficientnet is a lighter , less memory intensive model than resnet50. It can be safely assumed that increasing the epochs would not have helped since the loss drops to 0 before 40 epochs is reached.

It is clear from both our work and other research that unlike image classification, audio classification does not benefit from deep networks. Rather, higher resolution [15] or additional preprocessing is necessary to achieve high accuracy.

In order to overcome the overfitting issue, we then implemented ShuffleNetV2 from scratch. This turned out to give us the best results in terms of accuracy as ShuffleNet is able to improve accuracy without compromising computational cost.

**5. Conclusion and Future Work**

As the evaluation section shows, the task of Acoustic Scene Classification faces major hurdles when it comes to larger models. Due to the relatively small amount of data and distinguishable features, most large neural networks overfit. The only model that returned a successful result was ShuffleNet, the lightest of the three models tested.

Several models that have been successfully used to carry out Acoustic Scene Classification with a high degree of accuracy. However, they were highly customized, implementing extensive data augmentation and optimizing the model heavily. The need for such extensive work on the model and data means that progress in the field has been relatively slow and it is much harder to implement ASC in real-life or general applications. Our results show that currently, only a ShuffleNet model or another model similar in size can be used easily, despite lacking accuracy compared to customized models.

A solution to this may be to develop a more generalized, extremely lightweight CNN architecture that is designed specifically for classifying audio through spectrograms.

Future work to follow up on this paper would involve two approaches. The first would be to increase the resolution and accuracy of our comparison of models by using a wider selection of models. While the four models we used span a good range, there are large gaps that weaken our comparison. Secondly, there are several datasets available for ASC outside DCASE. Combining multiple datasets may enable the use of larger models and enable us to implement it in wider contexts.

**6. References**

[1] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020). 2020. Submitted. URL: <https://arxiv.org/abs/2005.14623>.

[2] Jung, A., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.M., Weng, C.H., Ayala-Acevedo, A., Meudec, R., Laporte, M., & others. (2020). imgaug. <https://github.com/aleju/imgaug>.

[3] Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing 28:. https://doi.org/10.1109/TASSP.1980.1163420

[4] Temko A, Nadeu C, Macho D, et al (2009) Acoustic Event Detection and Classification. In: Computers in the Human Interaction Loop. Springer London, London

[5] Liu, Z., Wang, Y., & Chen, T. (1998). The Journal of VLSI Signal Processing, 20(1/2), 61–79. doi:10.1023/a:1008066223044

[6] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley. A database and challenge for acoustic scene classification and event detection. In 21st European Signal Processing Conference (EUSIPCO 2013), volume, 1–5. Sep. 2013. [Doi:](https://doi.org/).

[7] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. IEEE Transactions on Multimedia, 17(10):1733–1746, Oct 2015. [doi:10.1109/TMM.2015.2428998](https://doi.org/10.1109/TMM.2015.2428998).

[8] M. Valenti, S. Squartini, A. Diment, G. Parascandolo and T. Virtanen, "A convolutional neural network approach for acoustic scene classification," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1547-1554, doi: 10.1109/IJCNN.2017.7966035.

[9] Hussain, K., Hussain, M., & Khan, M. G. (2017). Improved acoustic scene classification with DNN and CNN. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE).

[10] Heittola, T., Mesaros, A., & Virtanen, T. (2020). Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions. arXiv preprint arXiv:2005.14623.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[12] Abeßer J (2020) A Review of Deep Learning Based Methods for Acoustic Scene Classification. Applied Sciences 10(6). <https://doi.org/10.3390/app10062020>

[13] G. Z. Felipe, Y. Maldonado, G. d. Costa and L. G. Helal, "Acoustic scene classification using spectrograms," 2017 36th International Conference of the Chilean Computer Science Society (SCCC), Arica, 2017, pp. 1-7, doi: 10.1109/SCCC.2017.8405119.

[14] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848-6856).

[15] Zhang, T., Liang, J., & Ding, B. (2020). *Acoustic scene classification using deep CNN with fine-resolution feature. Expert Systems with Applications, 143, 113067.* doi:10.1016/j.eswa.2019.113067