Deep Audio Classification for Environmental Sounds

Suryaveer Singh suryavee 50486831
Sai Karthik Vyas Akondi sakondi 50488251
Ayush Rawat arawat2 50478823
Nikhil Naveen Chandra nikhiln2 50469151

Abstract

The report explores the use of deep learning techniques for audio classification using Python, TensorFlow, and PyTorch. We discuss 2 popular feature extraction methods of audio data: MEL spectrogram and MFCC, and how they can be used to represent audio signals as input to deep learning models. We also examine 2 different datasets which are commonly used in classification tasks: ESC50, UrbanSound8k. By training a deep neural network on these datasets, we demonstrate the effectiveness of MEL spectrograms and MFCCs in accurately classifying environmental sounds and other types of environmental sound signals. Our results show that these techniques can achieve high levels of accuracy on the training set in audio classification and have the potential to be applied to a real-life world problem.

1 Introduction

Audio signals are ubiquitous in our daily lives[1], and analyzing them can provide valuable insights into a wide range of real-world problems, from speech recognition to environmental monitoring. With the advent of deep learning techniques, it has become possible to develop highly accurate models for audio classification tasks, which can help automate many audio-related applications. Deep learning techniques have revolutionized the field of audio signal processing by enabling the development of highly accurate models for audio classification tasks. In recent years, deep learning methods such as convolution neural networks and recurrent neural networks have emerged as powerful tools for audio classification, enabling researchers and practitioners to accurately classify a wide range of audio signals.

2 Dataset

For Audio Classification with deep learning, we are using the two widely used Audio datasets
Datasets:
1. ESC-50 Dataset
2. UrbanSound8k Dataset

2.1 ESC-50 Dataset

The ESC-50 dataset[2] consists of 2000 labeled environmental recordings equally balanced between 50 classes (40 clips per class). For convenience, they are grouped into 5 loosely defined major categories (10 classes per category):
1. animal sounds
2. natural soundscapes and water sounds
3. human(non-speech) sounds
4. interior/domestic sounds

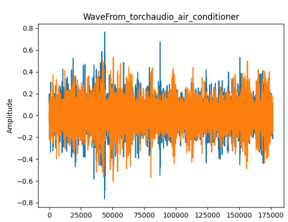
Submitted to Conference on Neural Information Processing Systems (NeurIPS). Do not distribute.

5. exterior/urban noises

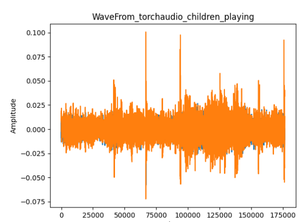
The goal of the extraction process was to keep sound events exposed in the foreground with limited background noise when possible. However, field recordings are far from sterile, thus some clips may still exhibit auditory overlap in the background.

2.2 UrbanSound8k Dataset

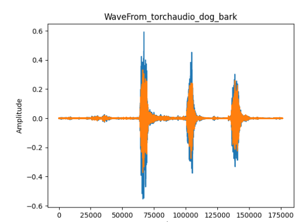
The UrbanSound8K dataset has been widely used in research studies related to audio classification, including studies that use MEL spectrograms and MFCCs. In particular, [3]Salamon et al. (2014) used this dataset to develop a method for environmental sound classification using deep learning. The sound excerpts are drawn from 10 different sound classes that are commonly found in urban environments:
1. air conditioner
2. car horn
3. children playing
4. dog bark
5. drilling
6. engine idling
7. gunshot
8. jackhammer
9. siren
10 street music
The following are the sound wave forms for the UrbanSound8k dataset:



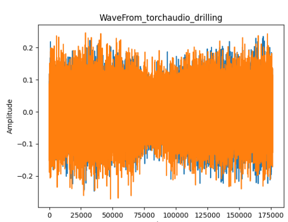
(a) Air Conditioner



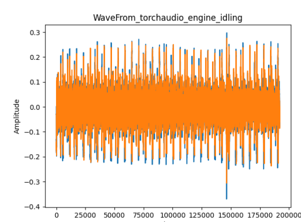
(b) Children Playing



(c) Dog Bark

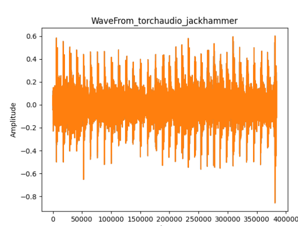


(d) Drilling

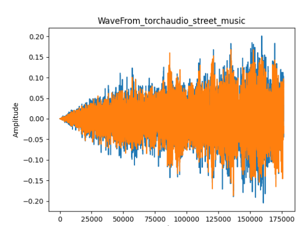


(e) Engine Idling

(f) Gunshot



(g) Jackhammer



(h) Street Music

Figure 1: Audio Waveforms

The UrbanSound8k dataset has been widely used for research in environmental sound classification, including studies of deep learning methods for audio analysis, feature extraction techniques for sound recognition, and applications of machine learning to environmental sound monitoring. The dataset is particularly well-suited for research on sound recognition in urban environments, where many different sounds may be present simultaneously and where classification accuracy is important for applications such as noise pollution monitoring and public safety.

3 Feature Extraction

In our study, we are using two most widely used Audio Signal processing techniques, MEL Spectogram, and MFCC. Both of these techniques convert audio signals to spectrograms. These spectrograms, like images, can be fed into image classification models to get good classification accuracy.

3.1 MEL Spectogram

A diagram that displays how a signal's frequency changes over time is called a mel spectrogram. In our study, we use the short-time Fourier transform to create a logarithmic power spectrogram S from an audio input X. This is accomplished by first dividing the input signal into overlapping frames, applying a window function to each frame, and then performing a Fast Fourier Transform on each frame separately. As can be seen from the Mel Spectrogram, which displays intensities in the 25–30 range for frames 0–40, the air conditioner emits a rather loud and constant level of sound in the frequency range covered by the Mel scale. [4] MEL Spectogram spectrogram is a visual representation that shows how frequencies of a signal change over time. When it is associated with digital signal processing, it can be seen that there are multiple ways to acquire a spectrogram which are produced using filter banks, Fourier transform, etc. In our study, we calculate a logarithmic power spectrogram S from the Sift of an audio signal X (τ,ω) which can be seen in equation 1. S=10Log10|X(τ,ω)|2 —(1) The Short-Time Fourier Transform is a type of transform that is related to the Fourier transform. It is used to analyze a time-domain signal x by determining the magnitude and phase of sinusoidal frequencies ω at different points τ seen in equation 2.

X(τ,ω)= ∞ n=-∞ x[n]w[n-τ]e-jωn —(2)

Taking equation 2 into the case and when computing, it would be needed to split the input signal that overlaps different frames which would be multiplied by the window function w, to which Fast Fourier Transform is going to be applied to each frame one by one.

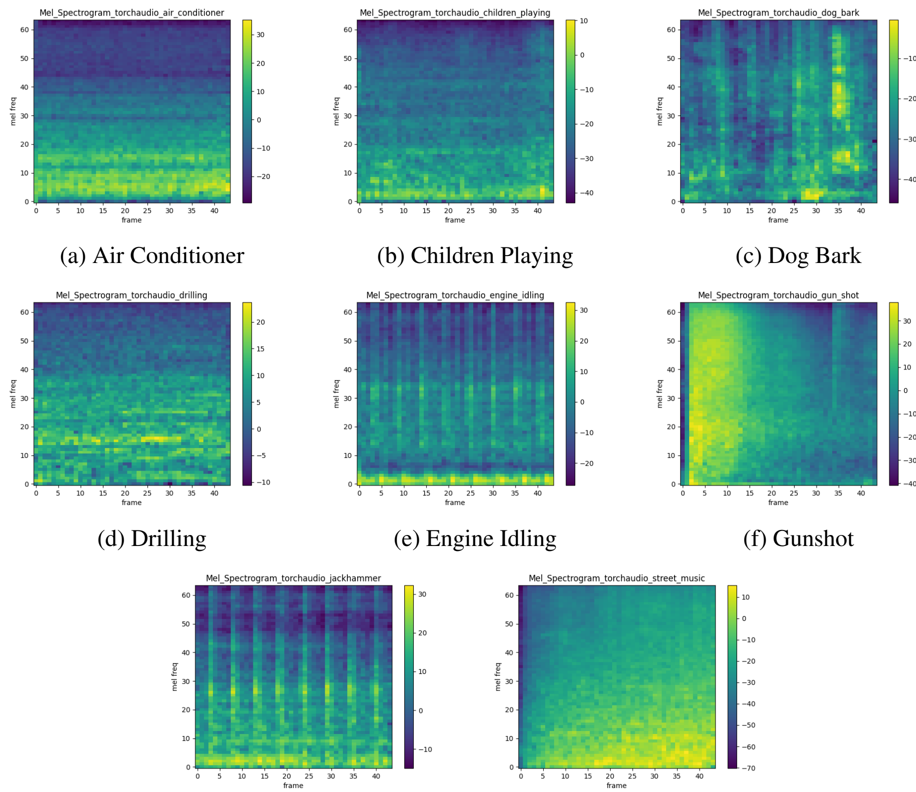
Air conditioner: The Mel Spectrogram of the air conditioner has intensities in the range of 25-30 for frames 0-40. This indicates that the air conditioner produces a relatively high and constant level of sound in the frequency range represented by the Mel scale.

The air conditioner of a Mel Spectrogram has multiple intensities that range between 25-30 that go for frames 0-40. It can be seen that they would produce a high and constant level of sound in the frequency range that would be shown by the Mel scale. The Mel spectrogram exhibits a relatively constant and consistent pattern of color/intensity, which is a hallmark of the sound produced by an air conditioner.

For frames 0-40, the drilling of the Mel Spectrogram shows intensities in the range of 20-50. This shows that the frequency range covered by the Mel scale is covered by the drilling's reasonably loud sound, with some intensity changes. In the provided range of frames, the Mel Spectrogram displays a more varied pattern of color/intensity, which is indicative of the sound made by the drilling machine.

For frames 0-40, the Mel Spectrogram for engine idling displays tiny, intermittent intensities in the 0-15 range. This suggests that the engine emits a sound in the frequency range covered by the Mel scale at a relatively modest volume with sporadic intensity swings. In the specified range of frames, the Mel Spectrogram displays a sparse pattern of color/intensity, which is a hallmark of the sound made by an idling engine

. The intensity ranges for the Mel Spectrogram of jackhammer are 0-10 and 20-30 for frames 0-40. This demonstrates that the jackhammer emits a sound at a pretty high volume with considerable intensity changes over the Mel scale's two different frequency bands. The Mel Spectrogram displays a more intricate color/intensity pattern in the range of frames provided, which is an indicative of the sound made by a jackhammer.



(a) Air Conditioner (b) Children Playing (c) Dog Bark

(d) Drilling (e) Engine Idling (f) Gunshot

(g) Jackhammer (h) Street Music

Figure 2: MEL Spectrograms

The intensity range for the Mel Spectrogram of street music is 0-10 for frames 0-40. This suggests that street music has a comparatively low sound output in the range of frequencies covered by the Mel scale. The background street music sound is characterized by a sparse pattern of color/intensity on the Mel Spectrogram in the given range of frames.

3.2 Mel Frequency Cepstral Coefficients- MFCC

The intensity range in MFCC charts represents the size of the MFCC coefficients. The intensity values, which are frequently standardized to a given range (such as 0 to 17.5), can be understood as the energy or strength of a specific frequency band. The intensity value increases with the frequency range's spectral features.

To find specific patterns or features in the audio stream, MFCC plots can be used to classify audio. For example, in the street music plot, the higher intensity values between 0 and 2.5 indicate stronger spectral characteristics in that frequency band, which can be used to identify the genre of the music.

The frequency components of an audio signal can be seen and analyzed using MFCC plots, which can be helpful for tasks like speech recognition, speaker identification, and music genre categorization.

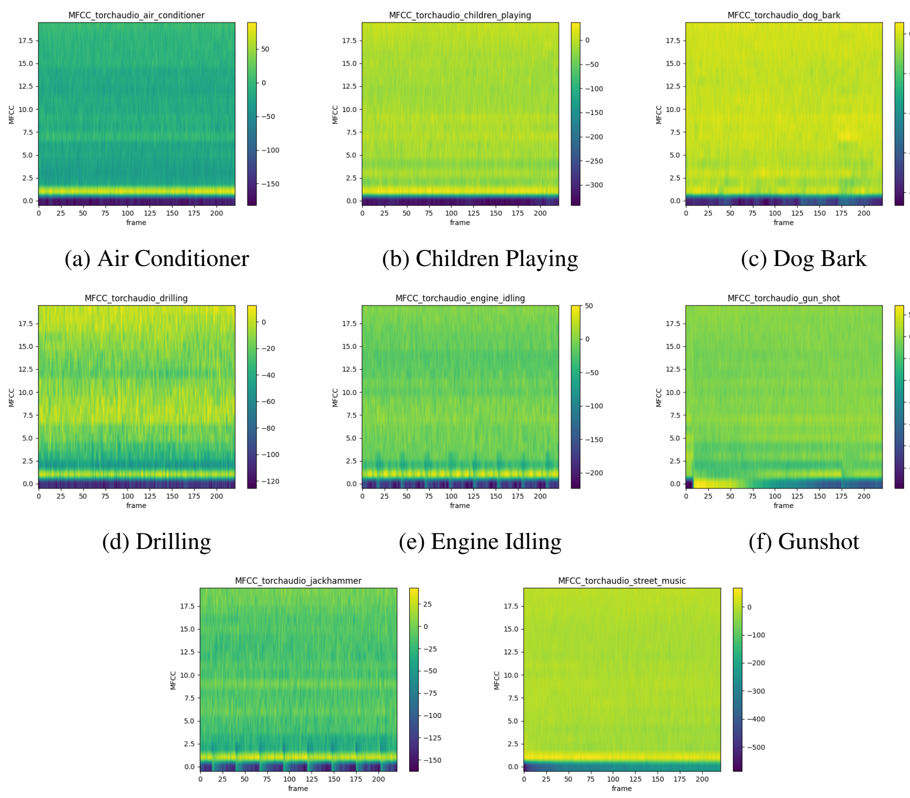
MFCC plot for street music: The plot shows the intensity of values of coefficients (such as speech recognition, speaker recognition and music genre classification) over time. The plot shows intensity values in the range of 0 – 17.5 for the first 200 frames. The intensity is higher in the range of 0 – 2.5 over the frequency range which indicates stronger spectral characteristics.

The plot shows higher intensity/energy in the range of 0 – 5.0 for the frames defined on a scale of 0 – 200. As the jackhammer sound changes over time, the intensity changes as well. This change is shown in the MFCC plot by the change in color. In this case, the jackhammer starts with low a low frequency which increases over time resulting in a change in intensity.

The plot shows the MFCC coefficient change over time (in frames). We can observe a sudden spike for the frames 0 – 25 which represents the high frequency for a short period of time of the gunshot. We can also observe the low-intensity values after the gunshot showing silence in the signal.

The plot shows the intensity range (defined by MFCC coefficient) for frames 0 – 200. The MFCC plot shows continuous high intensity in the range 0 – 1.0. The intensity decreases in the range between 1.0 – 2.5. We can see constant low intensity for the range of 2.5 – 17.5 with occasional variations in intensity.

The plot shows intensity or color changes which represent the magnitude of the MFCC coefficient. In the case of a dog bark, the graph shows a sudden increase in intensity over different intervals. The intensity is higher in the range of 0 – 5.0 and decreases afterward. We can visualize the increase in a particular frequency band (5 - 150) which indicates a rise in the pitch of the bark.



(a) Air Conditioner (b) Children Playing (c) Dog Bark

(d) Drilling (e) Engine Idling (f) Gunshot

(g) Jackhammer (h) Street Music

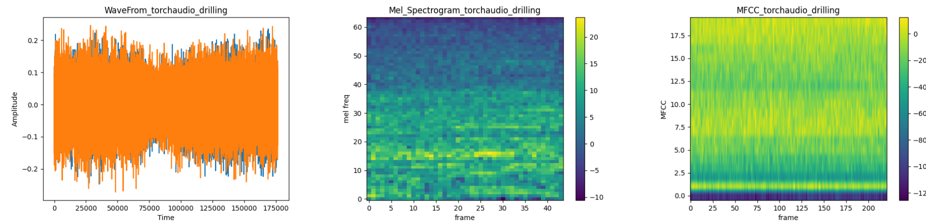
Figure 3: MFCC

3.3 Comparative Analysis of Audio Waveform, MEL Spectrogram, and MFCC

Drilling Audio Waveform:: The waveform of an audio signal shows a sound that is continuous, largely constant, and occasionally changes in volume. The representation of a sound's time-domain properties, such as its duration and amplitude, is called a waveform. indicates that despite brief volume changes, the drilling noise is consistent. However, frequency-domain data is not recorded.

Mel Spectrogram: The spectrogram shows some harmonics at higher frequencies and strong, consistent energy in the lower frequency range. It shows the frequency content of the sound with time. demonstrates that the drilling sound has significant energy in the lower frequency range as well as multiple harmonics at higher frequencies. The range of the spectrogram frames and mel frequency frames can help in interpreting the information that has been presented.

MFCCs: The MFCCs record harmonics at higher frequencies as well as changes in the temporal structure of the sound. It captures both the frequency and temporal structure of the sound and can be used as a compressed representation of the mel spectrogram. displays changes to the sound's temporal structure and higher frequency harmonics. Understanding how to interpret the provided data can be helped by knowing the range of the MFCC frames and coefficients.



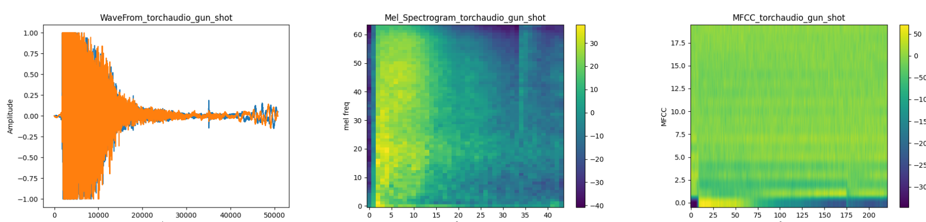
(a) Sound Waveform (b) MEL Spectrogram (c) MFCC

Figure 4: Drilling

Gunshot Audio Waveform:: The waveform of an audio signal shows a sound that is continuous, largely constant, and occasionally changes in volume. The waveform of a gunshot shows a quick, acute amplitude spike that is quickly followed by a decline. The amplitude is considerably higher than the drilling sound.

Mel Spectrogram: The spectrogram shows some harmonics at higher frequencies and strong, consistent energy in the lower frequency range. The spectrogram indicates a distinct and powerful energy at higher frequencies as opposed to the drilling sound. Additionally, there is greater energy in the mid-to high-frequency range.

MFCC: The MFCC records harmonics at higher frequencies as well as changes in the temporal structure of the sound. The MFCC of the gunshot sound exhibit higher energy levels than those of the drilling sound, according to the mel spectrogram. The sound also has a distinct temporal structure.



(a) Sound Waveform (b) MEL Spectrogram (c) MFCC

Figure 5: Gunshot

4 Methodology

In prior audio classification experiments, a variety of neural network topologies for music labeling have been examined. One such technique[5] is the convolutional neural network (CNN), which are used for local feature extraction and then does classification based on the obtained features. When compared to three other architectures used for music tagging, the CNN shows promising results while limiting the number of parameters in relation to each model's performance and training time per sample. Several studies have investigated fully connected deep neural networks (DNNs) [6] for audio categorization, including AlexNet, VGG, Inception, and ResNet. These CNNs demonstrated strong performance in image classification and exhibited promising results in audio classification when applied to a dataset of 70M training videos with 30,871 video-level labels.

Deep learning techniques have produced excellent results in sound recognition tests too. Making the right feeding decisions is crucial for ongoing performance development.

In our research, we have experimented with multiple different Neural Networks that are used for classifying images and we hence concluded that using RESNET-34 would be best for our usecase since it performs pretty well on other audio classification tasks and has 21.5M parameters which is a good number for the tradeoff between accuracy and faster training times.

4.1 RESNET-34

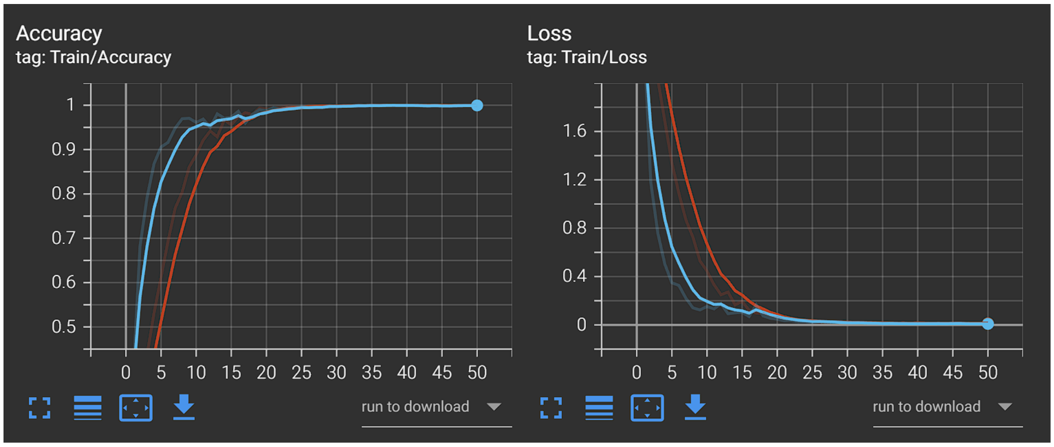
In experiments for picture categorization, a deep convolutional neural network architecture known as ResNet[3] (Residual Network) has displayed outstanding results. Its ability to learn more intricate and detailed properties makes it a preferred choice for many computer vision projects. Its promise in the area of audio has, however, also been looked at more recently. For example, the authors of the Residual Convolutional Neural Network for Music Tagging Using Raw Waveforms[7] research used ResNet to tag music using raw waveform inputs and achieved cutting-edge results using the well-known MagnaTagATune dataset. ResNet was applied in the context of acoustic scene classification in another paper, "Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Discriminative Outputs," and it demonstrated superior performance than other deep learning architectures.

4.2 Training

We used a pretrained resnt34 model from pytorch and applied transfer learning on it for our specific classification task. We modified the input and output layers of the model to accept our data and classify into specified classes for our datasets. We trained the model taking batch size as 128, and learning rate as 0.001 which adapts every 20 epochs and reduces by a factor of 0.1. We used Cross-entropy loss and Adam optimizer for backpropagation. we used L2 regularization for learning rate optimization to reduce overfitting. We trained the model for 50 epochs. The data for the training process was split into training and validation sets with an 80:20 split. We created the dataset by loading the audio files using torchaudio library and transformed it to mel-spectrograms and MFCC using audio libraries.

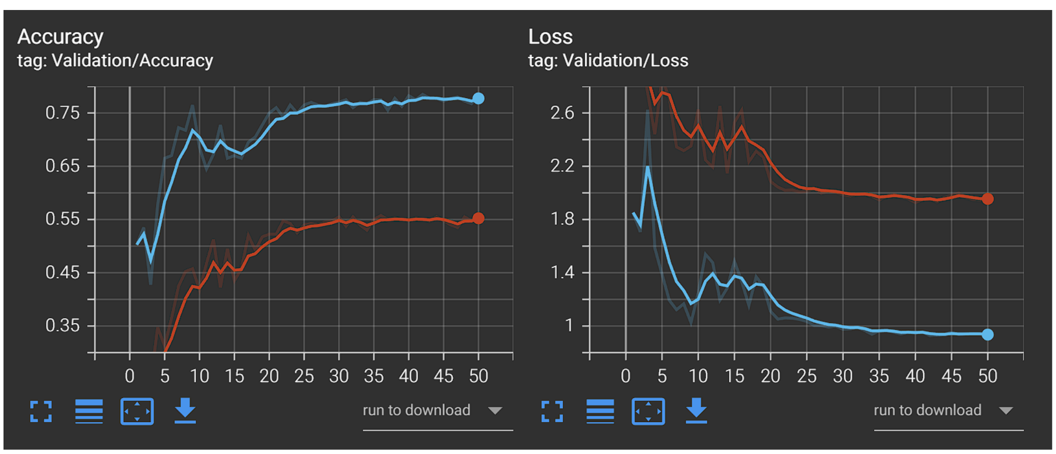
4.2.1 Training on ESC-50

Data for training the model on ESC-50 dataset is loaded using torchaudio library. The obtained waveform was then transformed into Mel-Spectrogram and MFCC. 2 models are trained one for each of these transformations to do a comparative study on the transformations.



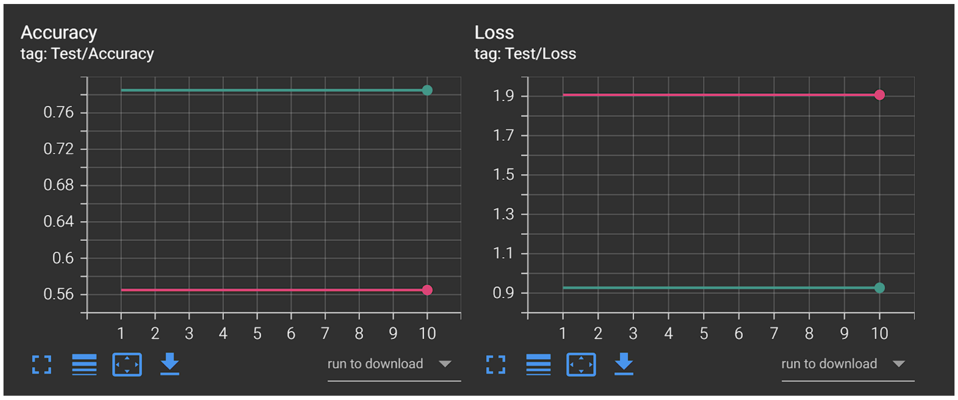
At the end of running 50 epochs, the accuracy obtained on training the ESC-50 dataset is 99 percent for MFCC Spectogram and 99 percent for Mel-Spectogram. Observed a loss of 0.01 for MFCC Spectogram and 0.01 for Mel-Spectogram.

Figure 6: ESCTrain



At the end of running 50 epochs, the accuracy obtained on Validating the ESC-50 dataset is 56 percent for MFCC Spectogram and 78.5 percent for Mel-Spectogram. Observed a loss of 1.9 for MFCC Spectogram and 0.9 for Mel-Spectogram.

Figure 7: ESCValidation

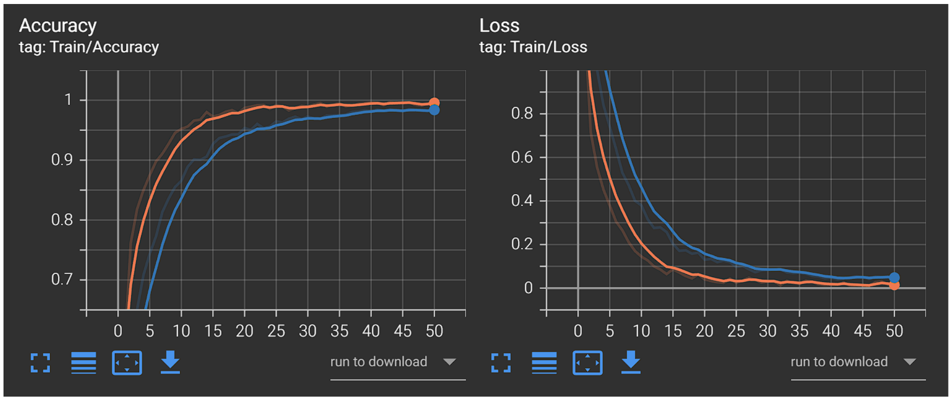


The testing accuracy for the ESC-50 dataset remains constant for 50 epochs at the rate of 56.5 percent for MFCC transformation(Orange) and is 78.5 for Mel-Spectogram Transformation (Grey). The loss observed for MFCC spectrogram is 1.9 and 0.9 for Mel-Spectogram.

Figure 8: ESCTest

4.2.2 Training on UrbanSound8k

Data for training the model on UrbanSound8k dataset is loaded using torchaudio library. The obtained waveform was then transformed into Mel-Spectrogram and MFCC. 2 models are trained one for each of these transformations to do a comparative study on the transformations.

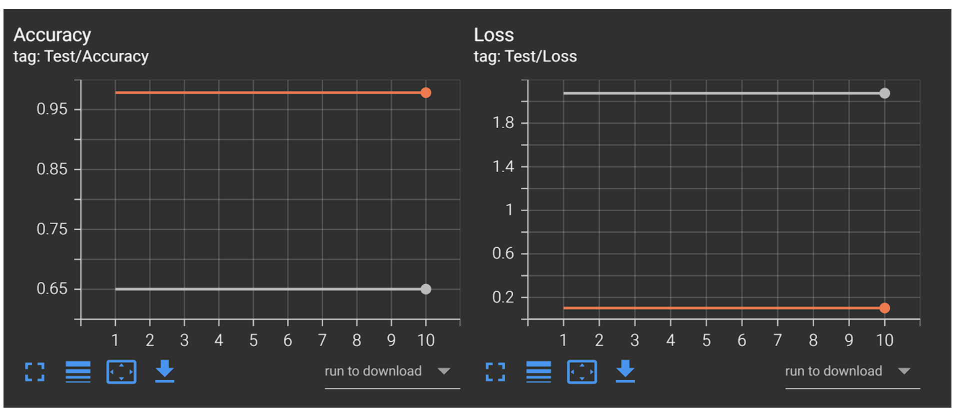


At the end of running 50 epochs, the accuracy obtained on training the UrbanSound8k dataset is 99.74 percent for MFCC Spectogram and 98 percent for Mel-Spectogram. Observed a loss of 0.01 for MFCC Spectogram and 0.04 for Mel-Spectogram.

Figure 9: UrbanTrainpng

At the end of running 50 epochs, the accuracy obtained on Validating the UrbanSound8k dataset is 90.74 percent for MFCC Spectogram and 55 percent for Mel-Spectogram. Observed a loss of 0.552 for MFCC Spectogram and 2.73 for Mel-Spectogram.

Figure 10: UrbanValidation



The testing accuracy for the UrbanSound8k dataset remains constant for 50 epochs at the rate of 97.6 percent for MFCC transformation(Orange) and is 65.3 for Mel-Spectogram Transformation (Grey). The loss observed for MFCC spectrogram is 0.1 and 2.0 for Mel-Spectogram.

Figure 11: UrbanTest

5 Testing and Results

In this study, we applied a ResNet34 deep learning model to classify audio data from two different datasets: Urbansound8k and ESC50. Our results show that the ResNet34 model achieved high accuracy and low loss values on both datasets, indicating its effectiveness in classifying different types of sound. One of the main findings of our study is the importance of the MEL spectrogram and MFCC features in the classification task. Our results show that for Urbansound8k dataset we achive very high accuracy using MFCC data transformation and comparatively low accuracy for Mel-Spectogram. On the other end when using ESC-50 dataset we achieve better accuracy on Mel-Spectogram than MFCC. This shows that for smaller dataset such as Urbansound8k a compressed form of transformation like MFCC can be used, but if we want better accuracies on very large datasets such as ESC-50, we need Mel-Spectogram which has a lot more details.

5.1 Confusion Matrix ESC-50 Dataset

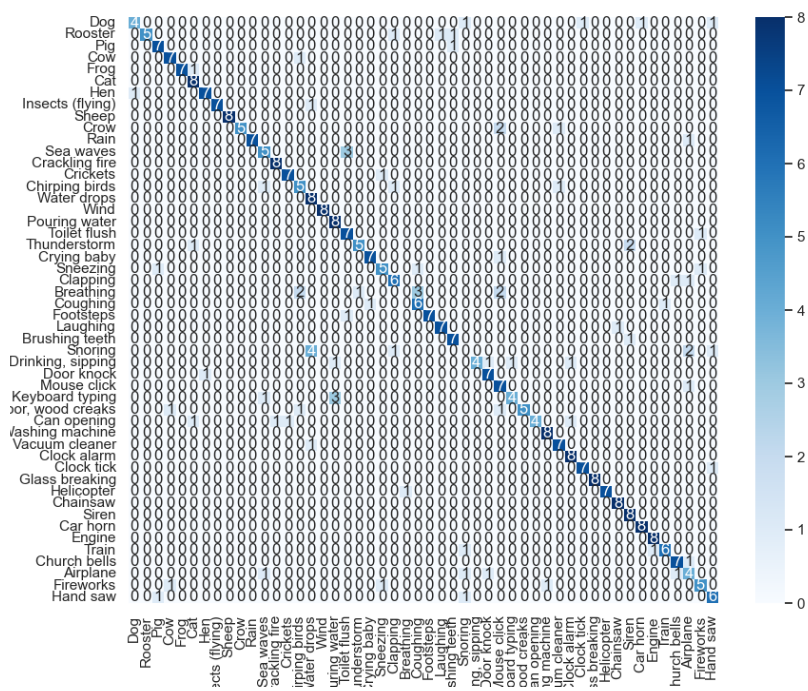


Figure 12: ESC_MEL-Spectrogram

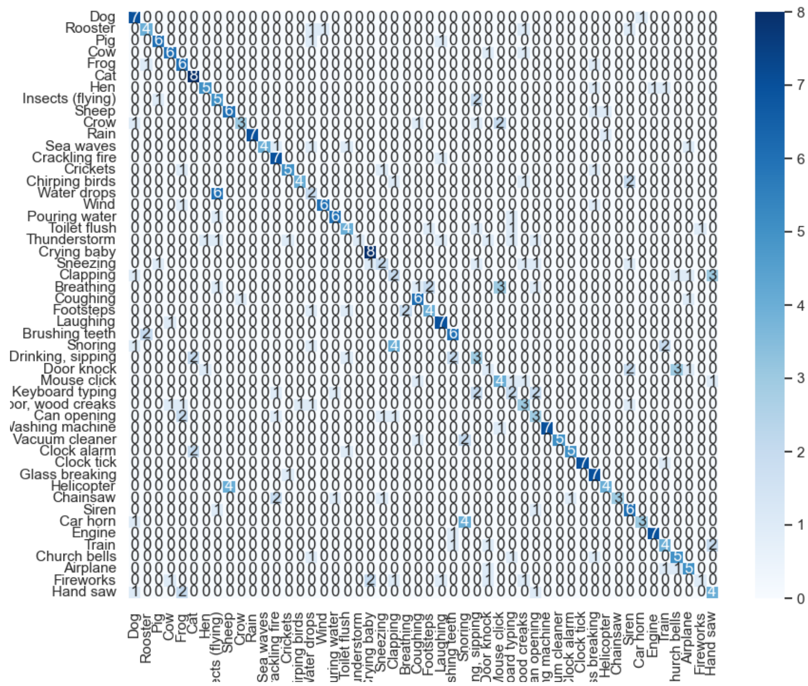


Figure 13: ESC_MFCC

5.2 UrbanSound8k Dataset

Figure 14: MEL_RESNET_50

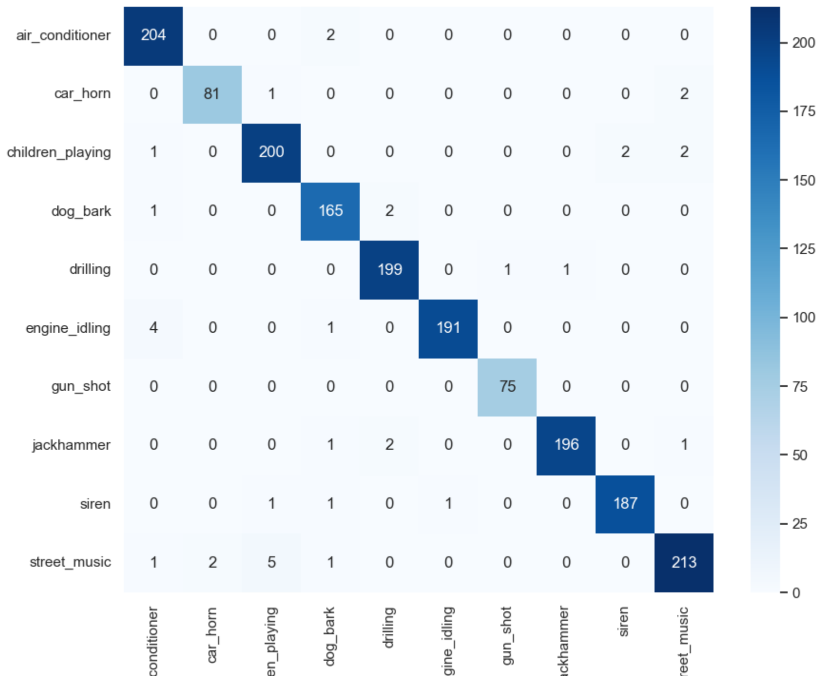


Figure 15: MFCC_RESNET_50

6 Conclusion

In summary, this work shows that the ResNet34 deep learning model can successfully categorize audio data from two independent datasets. Our findings demonstrate the model's capability for usage in practical audio categorization applications, showing good accuracy and low loss values. We also learned about how the size of the dataset and transformation impacts the model and gives a large variation in accuracy and losses.

The significance of the MEL spectrogram and MFCC features in the classification job is further highlighted by our findings, as well as the opportunity for further feature engineering research to enhance the classification performance of deep learning models and explain how these CNN models learn from these transformed spectrograms for feature extraction and classification. We learned that the models had some overfitting due to the size of the dataset being small and the noise being very high due to audio being captured in an outdoor environment. We could get better results by applying more transformations such as pitch shifts and cropping audio files randomly and other data augmentation methods.

7 References

1. J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283, Mar. 2017, doi: https://doi.org/10.1109/lsp.2017.2657381.
2. K. J. Piczak, "ESC," Proceedings of the 23rd ACM international conference on Multimedia, Oct. 2015, doi: https://doi.org/10.1145/2733373.2806390.
3. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv.org, Dec. 10, 2015. https://arxiv.org/abs/1512.03385
4. A. Guzhov, F. Raue, J. Hees, and A. Dengel, "ESResNet: Environmental Sound Classification Based on Visual Domain Models," arXiv:2004.07301 [cs, eess], Apr. 2020, Accessed: May 05, 2023. [Online]. Available: https://arxiv.org/abs/2004.07301

5. S. O. Arik et al., “Convolutional Recurrent Neural Networks for Small-Footprint Keyword Spotting,” arXiv:1703.05390 [cs], Jul. 2017, Accessed: May 05, 2023. [Online]. Available: https://arxiv.org/abs/1703.05390
6. H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. Sainath, “Deep Learning for Audio Signal Processing,” IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206–219, May 2019, doi: https://doi.org/10.1109/jstsp.2019.2908700.
7. T. Kim, J. Lee, and J. Nam, “Sample-level CNN Architectures for Music Auto-tagging Using Raw Waveforms,” arXiv:1710.10451 [cs, eess], Feb. 2018, Accessed: May 06, 2023. [Online]. Available: https://arxiv.org/abs/1710.10451

H. Lu, H. Zhang, and A. Nayak, “A Deep Neural Network for Audio Classification with a Classifier Attention Mechanism,” arXiv:2006.09815 [cs, eess, stat], Jun. 2020, Accessed: May 06, 2023. [Online]. Available: https://arxiv.org/abs/2006.09815

8. T. Ye, S. Si, J. Wang, N. Cheng, and J. Xiao, “Uncertainty Calibration for Deep Audio Classifiers,” arXiv:2206.13071 [cs, eess], Jun. 2022, Accessed: May 06, 2023. [Online]. Available: https://arxiv.org/abs/2206.13071
9. M. Ravanelli, B. Elizalde, K. Ni, and G. Friedland, “Audio Concept Classification with Hierarchical Deep Neural Networks,” arXiv:1710.04288 [cs, eess], Oct. 2017, Accessed: May 06, 2023. [Online]. Available: https://arxiv.org/abs/1710.04288
10. K. Palanisamy, D. Singhania, and A. Yao, “Rethinking CNN Models for Audio Classification,” arXiv:2007.11154 [cs, eess], Nov. 2020, Available: https://arxiv.org/abs/2007.11154
11. Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” arXiv:2104.01778 [cs], Jul. 2021, Available: https://arxiv.org/abs/2104.01778
12. Y. Gong, S. Khurana, A. Rouditchenko, and J. Glass, “CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification,” arXiv:2203.06760 [cs, eess], Mar. 2022, Accessed: May 06, 2023. [Online]. Available: https://arxiv.org/abs/2203.06760