***A Review of Audio Classification using Machine Learning: A Systematic Literature Review***

*Project Echo*

***Audio Classification using Machine Learning: A Systematic Literature Review***

Audio classification in the scope of our project will be to unobtrusively classify the different types of species in a rainforest. However, one of the benefits of using AI / ML to classify a specific type of data, is that it can easily be transferred to data in a completely different domain – using some fine-tuning techniques. For example, models developed for the research of cardiovascular diseases using audio samples of the heart [1] can be translated and slightly adjusted to fit the domain of audio samples of sound producing animals using transfer learning techniques.

A recent survey [2] examining the impact of the dataset size and number of classes on the accuracy obtained from acoustic classification shows a correlation between the two values:

Table

Description automatically generated

Figure Bioacoustics dataset and classification accuracy [2]

Table

Description automatically generated

Figure General Acoustics dataset size and classification accuracy [2]

An analysis of the studies that mentioned preprocessing revealed the most popular audio transformation technique as STFT (short-time Fourier transform) among both the bioacoustic and general acoustic studies [3] [4] [5] [6]. STFT breaks a signal into several signals of shorter duration and then transforms them into frequency domains. The other popular technique mentioned was constant-Q transform (CQT) which was used in both Bioacoustic analysis and general acoustic studied. It transforms a data series into a frequency domain. The FFT was also popular mainly in bioacoustic studies.

Feature extraction helps derive the audios short-time energy, zero-crossing rate, and bandwidth, among other useful features when classifying sound. Mel frequency cepstral coefficients (MFCCs) use the MEL scale to divide the frequency band into sub-bands and then extract the Cepstral Coefficients using a discrete cosign transform (DCT).

Machine learning algorithms: The survey showed that ensemble approaches are the most popular machine learning algorithms use in bioacoustics classification. Convolutional neural networks (CNN) were the most popular algorithms for general acoustic classifications. The choice of classifiers was motivated by the performance of similar classification tasks from previous studies. Bayesian [7] and hidden Markov models [8] showed the best accuracy levels for bioacoustic sounds, however only a few studies used them – due to higher computational cost and greater statistical expertise required. CNN algorithms and ensemble approaches were more poplar; however, they had slightly lower accuracy (87-88%).

Chart, scatter chart

Description automatically generated

Figure 3 classification algorithms used for Bioacoustic and general acoustic studies [2]

Table

Description automatically generated

Figure Review of lit [9]

Some researchers created models with a hybrid architecture combining transformers with Convolutional Neural network like Kong et al. [10], proposing a CNN-Transformer and an automatic threshold optimization method. Others focused on models based only on Transformers, presenting Bidirectional Encoder Representations from Transformers (BERT) based models capable of performing sound classification.

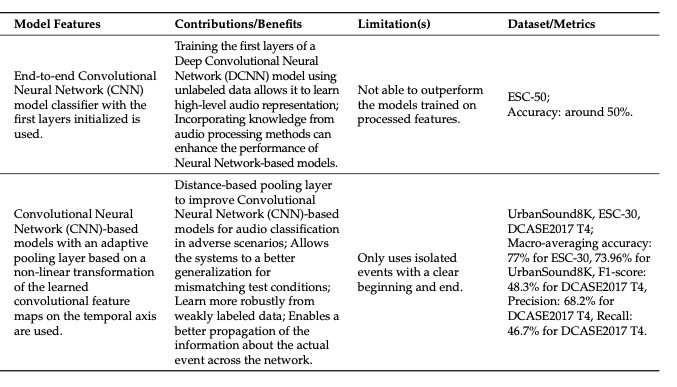


Figure review of lit [9]

Table

Description automatically generated

Figure review of lit [9]

Researchers have shown that seep features include more significant information than handcrafted features, which translates into better results. To further improve the models’ performance, researchers have implemented attention mechanisms that allow focusing on the semantically relevant characteristics. Therefore, the following section is focused on studies that implements different attention mechanism.

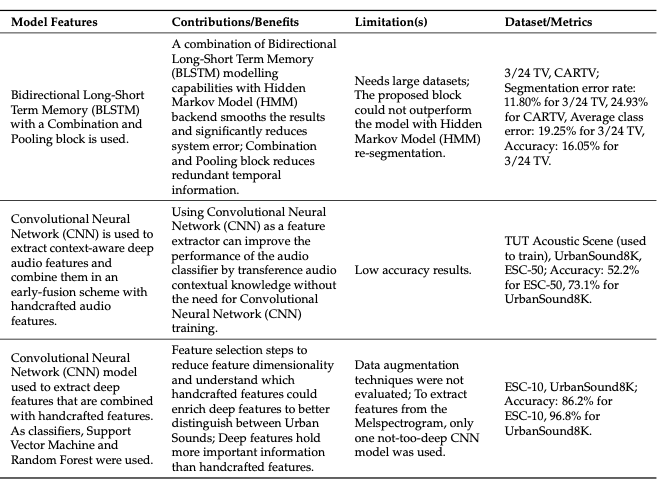


Figure review of lit [9]

The study presented by Zhang et al. (2019) [11] incorporated temporal attention and channel attention mechanisms. His proposal used a Convolutional Recurrent Neural Network (CRNN) model of eight convolution layers to learn high-level representations from the input log-gammatone spectrogram. The channel temporal attention mechanism enhanced the representational power of CNN. Then, two layers of Bidirectional Gated Recurrent Unit (B-GRU) were used to learn the temporal correlation information, to which the CNN learned features were given as input. Finally, SoftMax was used as activation function for the classification task.

Tripathi and Mishra [12] introduced an attention-based Residual Neural Network (ResNet) model that efficiently learns Spatio-temporal relationships in the spectrogram, skipping the irrelevant regions. They also used time shift, adding noise and Spec Augment.

## Potential methods and techniques for Project Echo

##### **Late Fusion**

Some researchers have implemented a late fusion technique to combine several different classifiers resulting in improved accuracy. [6] [13] [14] [15] The approach is to train a deep learning classifier e.g. CNN with some spectrogram representation, along with training some other traditional ML algorithms such as k-means clustering using different features of the same dataset. The predictions given by all the models are then summed at the end to enhance the accuracy of the predictions.

##### **Data Augmentation**

To improve the accuracy of the ML model some researchers [16] [17] [18] have augmented the training data by slightly varying the either the raw audio (e.g. clipping, speeding up, adding noise etc.) or the spectrogram (summing two spectrograms in the same class, shifting the pitch and time axis or warping the image). This approach will probably be necessary to address overfitting as we have a limited data pool.

##### **Audio only training**

Generating spectrograms can be time intensive, as it is processed with CPU and cannot be sped up with GPU, thus creating a bottleneck. Most studies use some form of spectrograms as a way to reduce the dimensionality but there exist methods that train using raw audio data. [7] Using SincNet, a CNN that learns the audio filters needed for dimensionality reduction, an accuracy comparable to image based models trained on spectrograms can be obtained.

Recently, technology has developed a lot, especially in the field of Machine Learning (ML), which is useful for reducing human work. In the field of artificial intelligence, ML integrates statistics and computer science to build algorithms that get more efficient when they are subject to relevant data rather than being given specific instructions Machine learning is commonly used in diverse fields to solve difficult problems that cannot be readily solved based on computer approaches. Recently, these advances in machine learning have helped a lot with sound classification, and sound recognition has shown to be a strong value in automating these tasks to say it another way, birds can make two basic sounds: Call and song. While this approach is time consuming, machine learning approaches may also be useful in establishing differentiating between the different species, even after that since it is done on a species of birds that are still not thought to be discernable. However, machine learning's usage of bird classification has only been examined for a small number of species or mannequin processing on the assumption that it can be applied in the real world only through numerical simulation or hand recordings. The results have proven unpractical for ecologists but can be useful for many people of a wide variety of professions

Capturing and automatically recognizing the acoustic emission resulting from typical behavior, i.e., locomotion, feeding, etc., of the target pests. After acquisition the signals are amplified, filtered, parameterized, and classified by advanced machine learning methods on a portable computer. Specifically, we investigate an advanced signal parameterization scheme that relies on variable size signal segmentation. The feature vector computed for each segment of the signal is composed of the dominant harmonic, which carries information about the periodicity of the signal, and the cepstral coefficients, which carry information about the relative distribution of energy among the different spectral sub-bands. This parameterization offers a reliable representation of both the acoustic emissions of the pests of interest and the interferences from the environment.

The authors have declared that no competing interests exist for North Atlantic right whales.

The predominant modern ML paradigm is Deep Learning (DL), a representation learning method in which the machine automatically discovers the representations that are required for carrying out a feature detection or classification task using raw input data. In particular, DL enables multilayer computational models to learn representations of data through the hierarchical composition of relatively simple non-linear modules that transform features into progressively higher levels of abstraction. Effectively, by proceeding from low-level to high-level feature abstraction, deep networks—which can be constructed from many layers and many units within layers—are able to learn increasingly complex functions. Importantly, the higher-level representations allow deep networks to extract relevant features from input data, which can be used to accurately perform discrimination, classification, and detection tasks. Artificial Neural Networks (ANNs) are popular realizations of such deep multilayer hierarchies, acting as highly non-linear parametric mappings from the input to the output space, with parameters (“weights”) determined by the optimization of some objective function (“cost”). This optimization is performed by backpropagating the error through the layers of the network in order to produce a sequence of incremental updates of the neural network weights. Popular among a plethora of different ANN architectures are Multi-Layer Perceptron (MLPs) typically used for general classification and regression problems, Convolutional Neural Networks (CNNs) for image classification tasks, and Recurrent Neural Networks (RNNs) for time series sequence prediction problems.

# Bibliography

|  |  |
| --- | --- |
| [1] | S. Sahoo, P. Dash and B. Mishra, "Deep learning-based system to predict cardiac arrhythmia using hybrid features of transform techniques," *Intelligent Systems with Applications,* vol. 16, 2022. |
| [2] | L. Mutanu, J. Gohil, K. Gupta, P. Wagio and G. Kotonya, "A Review of Automated Bioacoustics and General Acoustics Classification Research," *Sensors,* vol. 22, no. 21, p. 8361, 2022. |
| [3] | C. I. Kim, Y. Cho, S. Jung, J. Rew and E. Hwang, "Animal sounds classification scheme based on multi-feature network with mixed datasets," *KSII Transactions on Internet and Information Systems,* vol. 14, no. 8, pp. 3384-3398, 2020. |
| [4] | L. Nanni, S. Brahnam, A. Lumini and G. Maguolo, "Animal Sound Classification Using Dissimilarity Spaces," *Applied Sciences,* vol. 10, no. 23, p. 8578, 2020. |
| [5] | S. Gunasekaran and K. Revathy, "Content-Based Classification and Retrieval of Wild Animal Sounds Using Feature Selection Algorithm," in *Second International Conference on Machine Learning and Computing*, Bangalore, 2010. |
| [6] | J. Noda, C. Travieso and D. Sanchez-Rodriguez, "Methodology for automatic bioacoustic classification of anurans based on feature fusion," *Expert Systems with Applications,* vol. 50, pp. 100-106, 2016. |
| [7] | M. R. Hossain, N. B. English, S. T. Moore and M. R. Hossain, "Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture," *Scientific Reports,* vol. 11, p. 15733, 2021. |
| [8] | C. Bergler, R. X. Cheng, V. Barth, M. Weber, H. Schröter, E. Nöth, H. Hofer and A. Maier, "ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning," *Scientific Reports,* vol. 9, p. 10997, 2019. |
| [9] | A. F. Rodrigues Nogueira, H. S. Oliveira, J. J. Machado and J. M. R. S. Tavares, "Sound Classification and Processing of Urban Environments: A Systematic Literature Review," *Sensors,* vol. 22, p. 8608, 2022. |
| [10] | Q. Kong, Y. Xu, W. Wang and M. D. Plumbley, "Sound Event Detection of Weakly Labelled Data with CNN-Transformer and Automatic Threshold Optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 28, pp. 2450-2460, 2020. |
| [11] | Z. Zhang, S. Xu, S. Zhang, T. Qiao and S. Cao, "Learning Attentive Representations for Environmental Sound Classification," *IEEE Access,* vol. 7, pp. 130327-130339, 2019. |
| [12] | A. M. Tripathi and A. Mishra, "Environment sound classification using an attention-based residual neural network," *Neurocomputing,* vol. 460, pp. 409-423, 2021. |
| [13] | J. Xie and M. Zhu, "Handcrafted features and late fusion with deep learning for bird sound classification," *Ecological Informatics,* vol. 52, pp. 74-81, 2019. |
| [14] | J. Salamon, J. P. Bello, A. Farnsworth and S. Kelling, "Fusing Shallow and Deep Learning for Bioacoustic Bird Species Classification," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 141-145, 2017. |
| [15] | MathWorks, "Acoustic Scene Recognition Using Late Fusion," 2022. [Online]. Available: https://au.mathworks.com/help/wavelet/ug/acoustic-scene-recognition-using-late-fusion.html. |
| [16] | Y. R. Pandeya, D. Kim and J. Lee, "Domestic Cat Sound Classification Using Learned Features from Deep Neural Nets," *Applied Sciences,* vol. 8, no. 10, p. 1949, 2018. |
| [17] | L. Nanni, G. Maguolo and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics,* vol. 57, p. 101084, 2020. |