***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 can be translated and slightly adjusted to fit the domain of audio samples of sound producing animals using transfer learning techniques.

Table

Description automatically generatedTable

Description automatically generatedA survey 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:

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. 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).

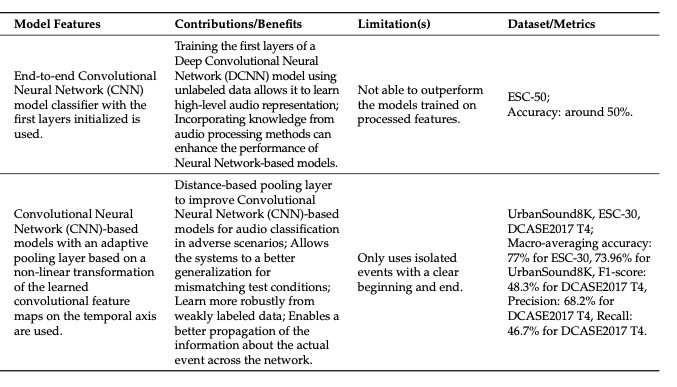
Machine learning algorithms: The survey showed that ensemble approaches are the most popular machine learning algorithms use in bioacoustics classification. Convolutional neural networks (CNN) were the most popular algorithms for general acoustic classifications. The choice of classifiers was motivated by the performance of similar classification tasks from previous studies. Bayesian and hidden Markov models 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

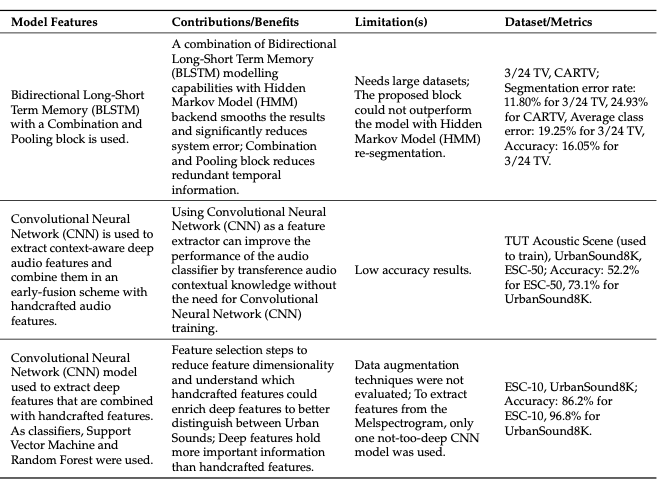
Table

Description automatically generatedSome researchers created models with a hybrid architecture combining transformers with Convolutional Neural network, 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.

Table

Description automatically generated

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 mechanisms.



The study presented by Zhang et al. 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 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.

References

[1]

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, Oct. 2022, doi: [10.3390/s22218361](https://doi.org/10.3390/s22218361).

[2]

T. Andersson, “Audio classification and content description,” p. 63.

[3]

H. Sinha, V. Awasthi, and P. K. Ajmera, “Audio classification using braided convolutional neural networks,” *IET signal process.*, vol. 14, no. 7, pp. 448–454, Sep. 2020, doi: [10.1049/iet-spr.2019.0381](https://doi.org/10.1049/iet-spr.2019.0381).

[4]

F. J. Bravo Sanchez, M. R. Hossain, N. B. English, and S. T. Moore, “Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture,” *Sci Rep*, vol. 11, no. 1, p. 15733, Dec. 2021, doi: [10.1038/s41598-021-95076-6](https://doi.org/10.1038/s41598-021-95076-6).

[5]

W. Chen, Q. Sun, X. Chen, G. Xie, H. Wu, and C. Xu, “Deep Learning Methods for Heart Sounds Classification: A Systematic Review,” *Entropy*, vol. 23, no. 6, p. 667, May 2021, doi: [10.3390/e23060667](https://doi.org/10.3390/e23060667).

[6]

S. Sahoo, P. Dash, B. S. P. Mishra, and S. K. Sabut, “Deep learning-based system to predict cardiac arrhythmia using hybrid features of transform techniques,” *Intelligent Systems with Applications*, vol. 16, p. 200127, Nov. 2022, doi: [10.1016/j.iswa.2022.200127](https://doi.org/10.1016/j.iswa.2022.200127).

[7]

A. Bansal and N. K. Garg, “Environmental Sound Classification: A descriptive review of the literature,” *Intelligent Systems with Applications*, vol. 16, p. 200115, Nov. 2022, doi: [10.1016/j.iswa.2022.200115](https://doi.org/10.1016/j.iswa.2022.200115).

[8]

L. D. Pha, “Robust Deep Learning Frameworks For Acoustic Scene and Respiratory Sound Classification,” Open Science Framework, preprint, Oct. 2021. doi: [10.31219/osf.io/d2tzb](https://doi.org/10.31219/osf.io/d2tzb).

[9]

A. F. R. Nogueira, H. S. Oliveira, J. J. M. Machado, and J. M. R. S. Tavares, “Sound Classification and Processing of Urban Environments: A Systematic Literature Review,” *Sensors*, vol. 22, no. 22, p. 8608, Nov. 2022, doi: [10.3390/s22228608](https://doi.org/10.3390/s22228608).