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.

A 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:

Dataset	Classes	Instances	Ratio	Average Accuracy
Cat Sound	2	440	220.00	91.13
Birdvox70k—CLO43SD	43	5428	126.20	90.00
Open Source Beehive Project	2	78	39.00	89.33
BIRDZ	50	602,512	12,050.20	89.04
Humboldt-University Animal Sound Archive	2530	120,000	47.40	81.30
MFCC dataset	10	7195	719.50	78.40
Zoological Sound Library	10,000	240,000	24.00	73.04
NIPS4Bplus	87	687	7.90	65.00

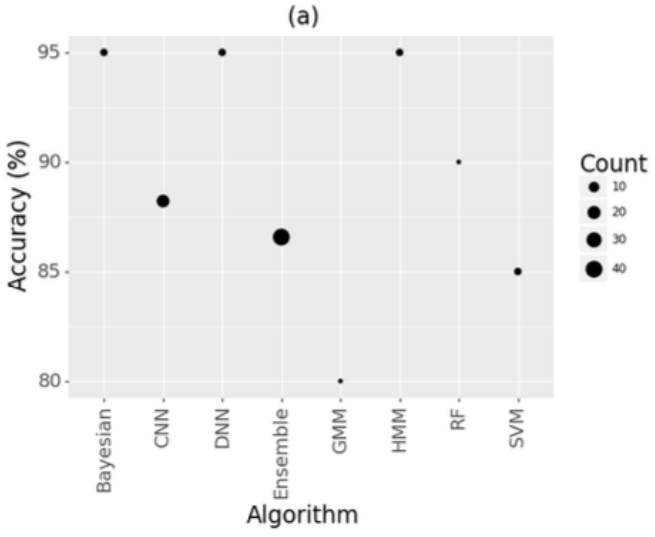
Dataset	Classes	Instances	Ratio	Average Accuracy
ESC-10	10	400	40.00	90.66
DCASE	16	320	20.00	86.70
US8K	10	8732	873.20	83.67
ESC-50	50	2000	40.00	81.54
Ryerson AV DB	8	7356	919.50	71.30
CICESE	20	1367	68.35	68.10

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%).



Some 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.

Model Features	Contributions/Benefits	Limitation(s)	Dataset/Metrics
CNN-Transformer model and an automatic threshold optimization method are used.	Computations are done in parallel; Use weakly labeled datasets to train the model and outputs directly clip-level predictions; Automatic threshold optimization is employed.	Convolutional Neural Network (CNN)-based models need many parameters.	DCASE2017 Task4; F1-score: 64.6% for AT, 57.3% for SED; Precision: 69.1% for AT; Recall: 60.7% for AT; Error rate: 68% for SED.
Bidirectional Encoder Representations from Transformers (BERT)-based Transformer for Environmental Sound Classification (ESC) at the edge is used.	Evaluation of Transformers' performance using several feature extraction techniques and data augmentation; Enables Environmental Sound Classification (ESC) on edge devices.	Models trained in traditional frameworks have little support to be converted to models that run at the edge; Lower competitive results when trained with small datasets.	ESC-50, Office Sounds; Accuracy: 67.71% for ESC-50, 95.31% for Office Sounds.
Bidirectional Encoder Representations from Transformers (BERT)-based Transformer used for Environmental Sound Classification (ESC) on a resource-constrained device applied in noisy environments.	The model trained with noise-augmented data can generalize to audio without noise and prevents having to construct custom acoustic filters to apply the model in real-life environments.	Needs large datasets; Only employed on small edge-end devices.	Office Sounds; Accuracy: 75.4% for non-noisy dataset, 81.2% for noisy dataset, Precision: 76.5% for non-noisy dataset, 79.7% for noisy dataset, Recall: 75.6% for non-noisy dataset, 80.6% for noisy dataset, F1-score: 75% for non-noisy dataset, 80% for noisy dataset.
Audio Spectogram Transformer, a purely attention-based audio classification model, is used.	Even in the lowest layers, it can capture long-range global context; Able to handle different input audio lengths without changing the architecture; Few parameters and fast convergence.	Cannot use rectangular patches due to the inexistence of a pre-trained model that used the same dataset as Vision Transformer (ViT); Unable to use only an AudioSet pre-trained model.	AudioSet, ESC-50, Speech Commands V2; mAP: 48.5% for AudioSet, Accuracy: 95.6% for ESC-50, 98.11% for Speech Commands V2.
Audio Spectrogram Transformer that can handle various output resolutions is used.	Shows that Soft F-loss performs better than binary cross-entropy; Designed to deal with a multiplicity of output resolutions.	Large model size; Evaluates sound event localization and detection using only one dataset.	TAU-NIGENS Spatial Sound Events 2021; Error rate: 50%, F1-score: 65.7%, Recall dominant score: 74.7%.

Model Features	Contributions/Benefits	Limitation(s)	Dataset/Metrics
Transformers for multimodal self-supervised learning from raw video, audio and text are used.	Learns effectively semantic video, audio and text representations; DropToken technique reduces the pre-training complexity, which reduces computational costs, the training time and enables the hosting of large models on restricted hardware.	Needs large datasets to be trained due to the large size of the network.	Only 2 out of 10 datasets were from the audio domain: ESC-50, AudioSet. mAP: 39.4% for AudioSet, AUC: 97.1% for AudioSet, d-prime: 2.895 for AudioSet, Accuracy: 84.9% for ESC-50.
Audio Transformer with Patchout which optimizes and regularizes Transformers on audio spectrograms is used.	Patchout improves the generalization and reduces the computation and memory complexity.	Increases the training time	AudioSet, OpenMIC, ESC-50, DCASE20; mAP: 49.6% for AudioSet, 84.3% for OpenMIC, Accuracy: 96.8% for ESC-50, 76.3% for DCASE20.

Model Features	Contributions/Benefits	Limitation(s)	Dataset/Metrics
End-to-end Convolutional Neural Network (CNN) model classifier with the first layers initialized is used.	Training the first layers of a Deep Convolutional Neural Network (DCNN) model using unlabeled data allows it to learn high-level audio representation; Incorporating knowledge from audio processing methods can enhance the performance of Neural Network-based models.	Not able to outperform the models trained on processed features.	ESC-50; Accuracy: around 50%.
Convolutional Neural Network (CNN)-based models with an adaptive pooling layer based on a non-linear transformation of the learned convolutional feature maps on the temporal axis are used.	Distance-based pooling layer to improve Convolutional Neural Network (CNN)-based models for audio classification in adverse scenarios; Allows the systems to a better generalization for mismatching test conditions; Learn more robustly from weakly labeled data; Enables a better propagation of the information about the actual event across the network.	Only uses isolated events with a clear beginning and end.	UrbanSound8K, ESC-30, DCASE2017 T4; Macro-averaging accuracy: 77% for ESC-30, 73.96% for UrbanSound8K, F1-score: 48.3% for DCASE2017 T4, Precision: 68.2% for DCASE2017 T4, Recall: 46.7% for DCASE2017 T4.

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.

Model Features	Contributions/Benefits	Limitation(s)	Dataset/Metrics
Bidirectional Long-Short Term Memory (BLSTM) with a Combination and Pooling block is used.	A combination of Bidirectional Long-Short Term Memory (BLSTM) modelling capabilities with Hidden Markov Model (HMM) backend smooths the results and significantly reduces system error; Combination and Pooling block reduces redundant temporal information.	Needs large datasets; The proposed block could not outperform the model with Hidden Markov Model (HMM) re-segmentation.	3/24 TV, CARTV; Segmentation error rate: 11.80% for 3/24 TV, 24.93% for CARTV, Average class error: 19.25% for 3/24 TV, Accuracy: 16.05% for 3/24 TV.
Convolutional Neural Network (CNN) is used to extract context-aware deep audio features and combine them in an early-fusion scheme with handcrafted audio features.	Using Convolutional Neural Network (CNN) as a feature extractor can improve the performance of the audio classifier by transference audio contextual knowledge without the need for Convolutional Neural Network (CNN) training.	Low accuracy results.	TUT Acoustic Scene (used to train), UrbanSound8K, ESC-50; Accuracy: 52.2% for ESC-50, 73.1% for UrbanSound8K.
Convolutional Neural Network (CNN) model used to extract deep features that are combined with handcrafted features. As classifiers, Support Vector Machine and Random Forest were used.	Feature selection steps to reduce feature dimensionality and understand which handcrafted features could enrich deep features to better distinguish between Urban Sounds; Deep features hold more important information than handcrafted features.	Data augmentation techniques were not evaluated; To extract features from the Melspectrogram, only one not-too-deep CNN model was used.	ESC-10, UrbanSound8K; Accuracy: 86.2% for ESC-10, 96.8% for UrbanSound8K.

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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.