Audio Representations for CNN-based Semantic Sound Classification*

Katarzyna Wiltos[1]

[1]Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44100 Gliwice, Poland

Abstract

Audio classification aims to automatically differentiate between sounds that belong to various semantic classes. A crucial step in preparation of efficient classification model is the selection of the most suitable audio signal representation to maximize the model's performance. Currently, most studies treat popular encodings such as mel spectrograms (Mel) or Mel-Frequency-Cepstral-Coefficients (MFCC) as default approaches. This paper provides a systematically conducted evaluation of the influence of eight widely applied time-frequency and alternative visual representations: Mel, Log-Mel, MFCC, Short-Time Fourier Transform (STFT), Chroma, Constant-Q Transform (CQT), Tempogram, and Waveform on the performance of three CNN Models: DenseNet121, MobileNetV2, and Xception, and two distinct datasets: GTZAN (for music genre classification) and UrbanSound8K (for environmental sound classification). It explores if any representation yields noticeably superior results regardless of the CNN model or dataset. The evaluation was carried out with ensured reproducibility through a reliable unified training and assessment pipeline with 5-fold cross validation. The findings reveal that Log-Mel, Mel, CQT, and STFT respectively and consistently contribute to the best classification results.

Keywords

Audio Classification, Audio Representations, Spectrograms, Convolutional Neural Networks, Deep Learning

1. Introduction

Deep Learning techniques have tremendously transformed the area of audio signal processing, with tasks such as sound recognition and classification performed with great preciseness. The cornerstone of effective audio classification, besides the appropriate model selection and its tuning, the data preprocessing or augmentation, lies in the choice of optimal audio representation, which has a tremendous influence on the final results. The question of how each audio representation could impact the results remains underexplored. In the toolbox of models there are multiple options to utilize such as Convolutional Neural Networks (CNNs), Transformers and Attention-based Models, Recurrent Neural Networks (RNNs), hybrid models, such as for e.g. Convolutional Recurrent Neural Networks (CRNNs) or traditional machine learning algorithms [1, 2]. In this paper, the evaluation relies on three CNN models: DenseNet121, MobileNetV2, and Xception as a baseline. Additionally, the results are tested on two unrelated datasets: GTZAN (for music genre classification) and UrbanSound8K (for environmental sound classification). This setup along with the designed exeriments pipeline with 5-fold cross-validation ensures a reliable comparison of the effect of selected audio representations in various circumstances. It can be observed how these representaitons affect the performance with various data availability - low to medium, and in distinct CNN architectures. Performance impact is measured through multiple metrics, including Accuracy, Macro F1 Score, and ROC-AUC (macro OvR). This study enables to explore if any pattern in performance is present among the selected representations and if any of them yields consistently better results in all scenarios. Spectrograms and alternative representaitons are widely adopted, but few studies explore which audio representaiton are most optimal for CNN-based models. This work allows to systematically examine whether certain audio representations yield superior performance across models and datasets, if any pattern emerges across these setups, or how the data quantity will affect the results among the representaitons. The main contributions of this paper are the following:

*IVUS 2025: Information Society and University Studies 2025, May 15, Kaunas, Lithuania

wiltos.katarzyna@gmail.com (K. Wiltos)

- Evaluation of 8 selected most popular audio signal visual encodings in CNN-based model performance.
- Reliable comparison across 2 distinct datasets - GTZAN (for music genre classification) and UrbanSound8K (for environmental sound classification).
- Utillization of 3 CNN Models - DenseNet121, MobileNetV2, and Xception.

This study provides a systematic, reproducible, multi-architectural and multi-dataset evaluation of audio representations in CNN-based semantic classification.

2. Related Works

Many studies explored various tailored approaches for specific audio classification scenarios achieving diverse results. In [3] Mu et al. proposed a temporal-frequency attention based convolutional neural network model (TFCNN), evaluated on ESC-50 and UrbanSound8K datasets with the use of Log-Mel Spectrograms, which achieved between 84.4% and 93.1% of accuracy. In [4] authors presented a CNN Model with Mel representaitons tested on UrbanSound8K, which with data augmentation applied obtained approximately 79% accuracy. In [5] Wolf-Monheim, F. conducts evaluation of effects of applying different spectograms - Mel, MFCC, STFT Chromagrams, CQT Chromagrams, and Tempograms among others, in CNN-based Model performance, presenting results of 94.06% of training and 57.50% validation accuraacy for Mel, 93.88% and 56% of training and validation accuracy respectively for MFCC, highlighiting their superior performance as compared to the remaining representaitons. In [6] M. Dong achieves approximately 70% accuracy with the use of Mel-spectograms, CNN-based model and GTZAN dataset. A novel hybrid approach was presented in [7] by Jena, K.K. et al., where a multimodal CNN was applied for GTZAN dataset with fused inputs for music genre classification, which obtained 81% of accuracy. In [8] Ahmed, M. et al. proposed a modified CNN Model with applied ensamble learning, which was then trained on music genre classification datasets and obtained 92.7% of accuracy, where the features used for training were based on MFCC. In the [9] a novel method was presented, which combined CNN with capuchin search algorithm and derived features from discrete wavelet transform (DWT), MFCC, and STFT, achieving 96% of accuracy on GTZAN dataset. In [10] Zhao et al. utilized Attention-based model for music genre classification on GTZAN dataset obtaining 97.2% of accuracy with data augmentation and CQT spectrograms. A CNN-based Model combined with traditional machine learning algorithms was applied in [11], which obtained up to approximately 90% of accuracy on 5 music genres of GTZAN dataset.

3. Methodology

To perform the evaluation eight most common audio representaitons were selected, namely Mel, Log-Mel, MFCC, STFT, Chroma, CQT, Tempogram, and Waveform. To ensure that results are reliable, each representation was utilized in training of 3 various CNN architectures, and on 2 diverse datasets, that contained sound samples from unrelated scenarious - music and environmental sounds. For evaluation of how each audio representation impacts CNN-based Models ability to generalize - perform well on unseen input samples, a 5-fold stratified cross-validation was done along with set seed for reproducibility.

3.1. Evaluation Pipeline

Firstly, all audio .wav files were preprocessed and audio representations were saved to separate folders for each dataset - GTZAN[12] and UrbanSound8K[13], and for each representation type. For each .wav file, the sound recording was normalized, resampled to 22.050 Hz, additionally ensuring all audio samples had same duration, 8 representaitons were generated and saved as PNG files, where target image size was 128x128 pixels. If any audio clip was shorter than 4 seconds, a reflection padding was applied,

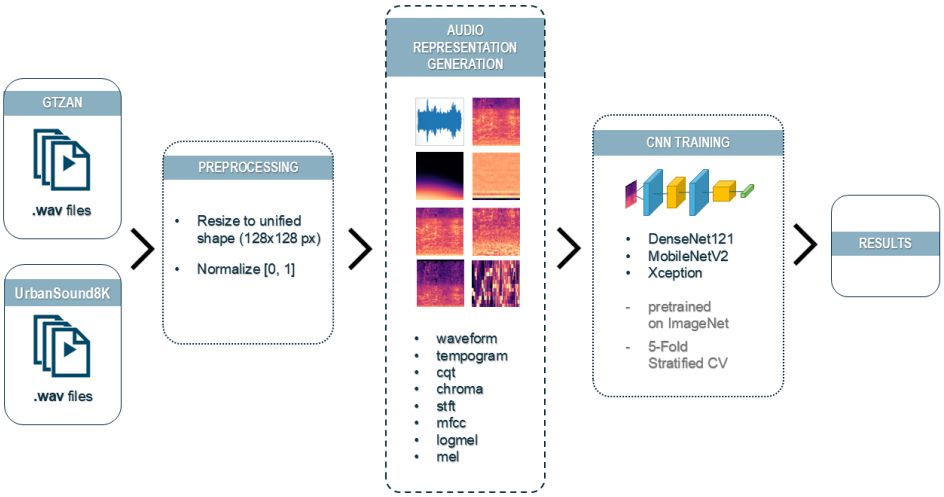

Figure 1: Pipeline diagram for audio representation performance impact evaluation on 2 distinct datasets and 3 CNN architectures with 5-fold stratified cross-validation.

and longer clips were truncated, so that all samples had same duration. Fourier transform parameters for applicable spectograms are set to 2048 and 512, for number of points in the Fast Fourier Transform (NFFT), which defines how many data are taken at a time to calculate the frequency content, and the hop length, which determines the distance between windows. In this case, the hop window of 512 indicates that after analyzing the first 2048 data points of the given sound clip, the next FFT calculation starts 512 data points later. Smaller hop length would cause more overlapping windows. After that, each dataset was prepared for 5-fold stratified cross-validation. Three CNN architectures were employed as baseline models for audio representaitons evaluation, namely DenseNet121, MobileNetV2, and Xception. All models were initialized with ImageNet-pretrained weights. The utilization of transfer learning allowed to laverage the pretrained models, while limiting the computational demands and training time. Model performance was measured by Accuracy, Macro ROC-AUC (calculated using the one-vs-rest), Macro F1, and additional visualizations. Random seed was set and applied across the data splits and model initializations to guarantee reproducibility. Each model was trained for 50 epochs with a batch size of 32, using the Adam optimizer and sparse categorical cross-entropy as the loss function, without early stopping to allow for reliable comparison among all settings. In Figure 1 the full experiments pipeline is visualized.

3.2. Audio Representations

The study utilized two well-established benchmark datasets that are used for audio classification tasks evaluation. These selected datasets allow to asses the performance across distinct acoustic domains. The GTZAN dataset, used for music genre classification, consists of 1000 audio samples, including 10 music genre classes - *blues, classical, country, disco, hiphop, jazz, metal, pop, reggae,* and *rock.* Each class containing 100 samples with duration of around 30 seconds. The UrbanSound8K dataset, used for environmental sound classification tasks, consits of 8732 audio samples with duration of around 4 seconds each. The dataset is divided into 10 classes including various incidents - *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren,* and *street music.* For each dataset the audio representaitons were generated using aformentioned pipeline.

In the Figure 2 audio representaitons for both datasets are presented across 10 classes and all 8 representaiton types. All spectrograms were generated with the help of the librosa 0.11.0 Python library [14].

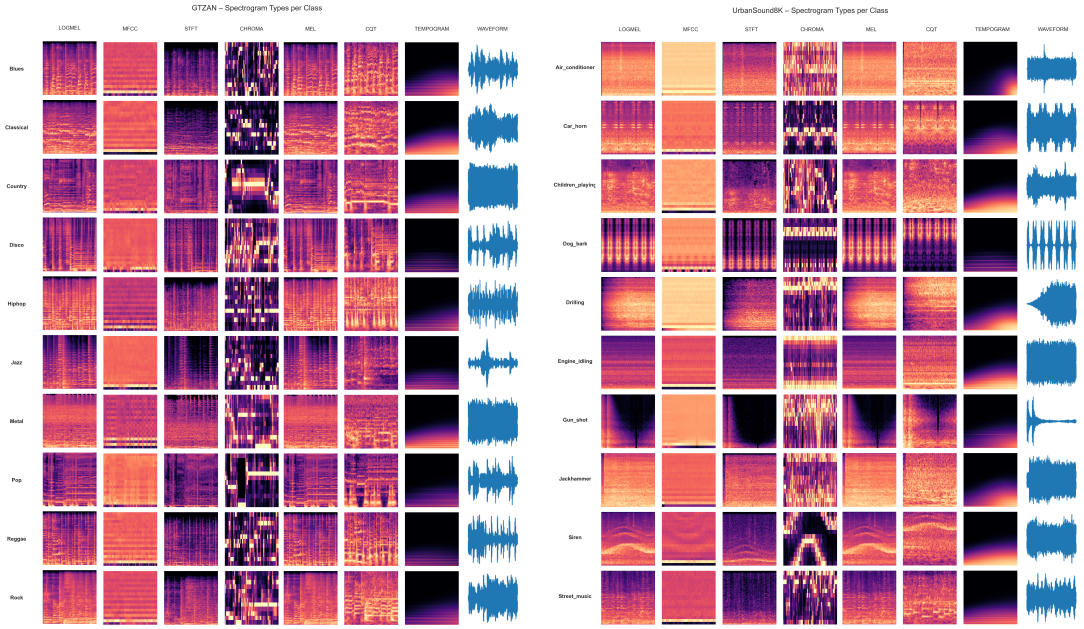

Figure 2: Eight visual audio representation types generated for two distinct datasets - GTZAN (for music genre classification) and UrbanSound8K (for environment sound classification).

3.2.1. Short-Time Fourier Transform (STFT)

The STFT of a signal is based on Discrete Fourier Transform (DFT). STFT is calculated by dividing the original audio clip into short parts of overlapping frames, that are then passed into the Fast-Fourier Transform (FFT):

$$X[n, \lambda) = \sum_{m=-\infty}^{\infty} x[n+m]w[m]e^{-j\lambda m} \tag{1}$$

where $w[n]$ represents the window function (such as Hann or Hamming), the $x[n]$ is the signal, and $x[n+m]$ is the shifted signal, in the end providing a time-frequency representation of the signal [15], and $\lambda = \frac{2\pi f}{F_s}$, where F_s is the sampling rate of the signal.

3.2.2. Mel

The Mel spectrogram is obtained in around three steps. In audio clips, frequencies vary across different timestamps, which is why Fast Fourier Transform would miss a lot of valuable information across the whole clip, thus the use of Short-Time Fourier Transform (STFT), which is FFT applied on small overlapping chunks (windows) of signal [16]. First, the window function is applied to each frame, then the STFT is calculated to obtain frequency-domain representaiton, and Mel filter banks are applied, which are triangular filters, which enhance low-frequency details (where human hearing is sensitive) and reduce the high-frequency nuances. That enhancement is obtained through the following calculation:

$$\text{mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

where $\text{mel}(f)$ represents the pitch, which resembles the human auditory system's perception, and f denotes the frequency in Hertz (Hz) [17].

3.2.3. Log-Mel

The Log-Mel spectrogram applies logarithmic scaling to the Mel spectrogram, which compresses the dynamic range of spectrogram, so the ratio between the loudest and quietest sounds is reduced, making

the loud sounds that may dominate, get balanced [18].

3.2.4. Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are computed by applying the Discrete Cosine Transform (DCT) to the Log-Mel spectrum. In librosa by default, it is the DCT type II:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot \cos\left(k\frac{\pi}{N}\left(n+\frac{1}{2}\right)\right) \tag{3}$$

where N is the number of samples, $x[n]$ is the input, and k is the cepstral coefficient index [19].

3.2.5. Chroma

Chroma features represent the intensity of 12 pitch classes (C, C#, D, …, B) in audio singal, while ignoring octaves through modulation into one octave, and it is also based on the STFT [20].

3.2.6. Constant-Q Transform (CQT)

The CQT represents the signal using geometrically spaced frequency bins:

$$X(n,k) = \sum_{m=0}^{N_k-1} x[m]a_k^*[m-n] \tag{4}$$

where a_k is the kernel for the k-th frequency bin, and N_k is the variable window length depending on the frequency, m is the inner index in the window, and n is shift [21].

3.2.7. Tempogram

Tempogram is a time-tempo representation of an audio signal, where tempos measured in beats per minute (BPM) are presented across various points in time of the audio clip. This representation shows how rhythmic periodicities change over time [22].

3.2.8. Waveform (Raw Audio Signal)

The raw waveform is the time-domain signal represented as:

$$x[n] \tag{5}$$

where n is the discrete-time audio signal, which is defined at a specific time intervals, not continuously [15].

4. Results

The obtained results imply a presence of consistent pattern of superiority of given audio representaitons that is upheld throughout all settings. These audio representations include Log-Mel, Mel, CQT, and STFT. Their utilization ensured stable results of accuracy above 70%, ROC-AUC above 90% and F1 Score above 70%.

The heatmaps in Figure 3 help to visually observe the pattern of how particular audio encodings influence the performance across two metrics in all tested CNN models and both datasets.

The spider plots in Figure 4 offer additional insights into how the pattern in results is present.

Confusion matrices reveal how classification was successful across semantic classes of each dataset. In Figure 5 and in Figure 6 the classification results are visualized for GTZAN and UrbanSound8K respectively, for the best CNN model: MobileNetV2 for the 5th fold from the evaluation. These and

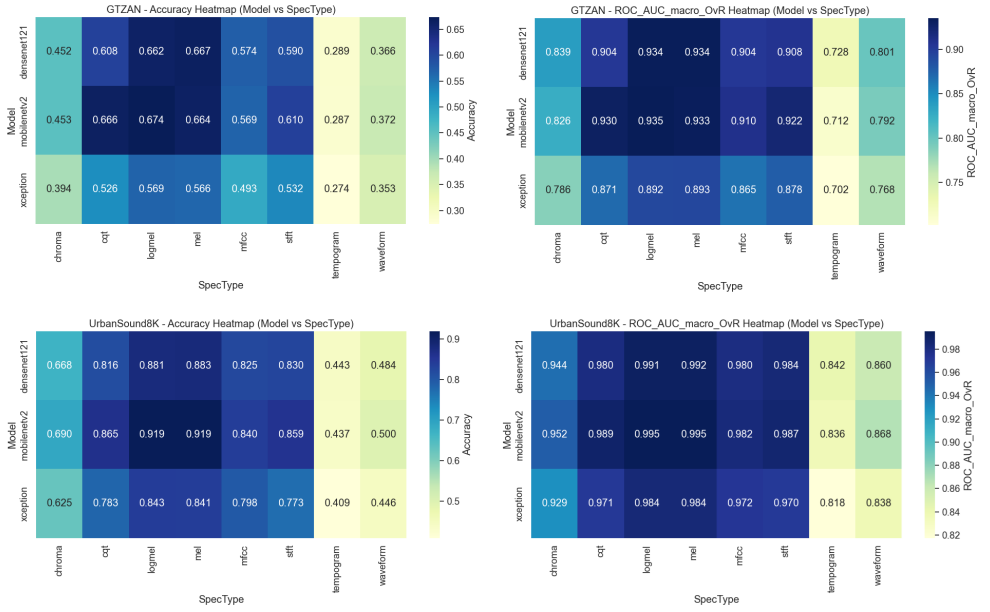

Figure 3: Accuracy and Macro ROC-AUC One-vs-Rest heatmaps for eight audio visual representations, for each dataset and CNN model.

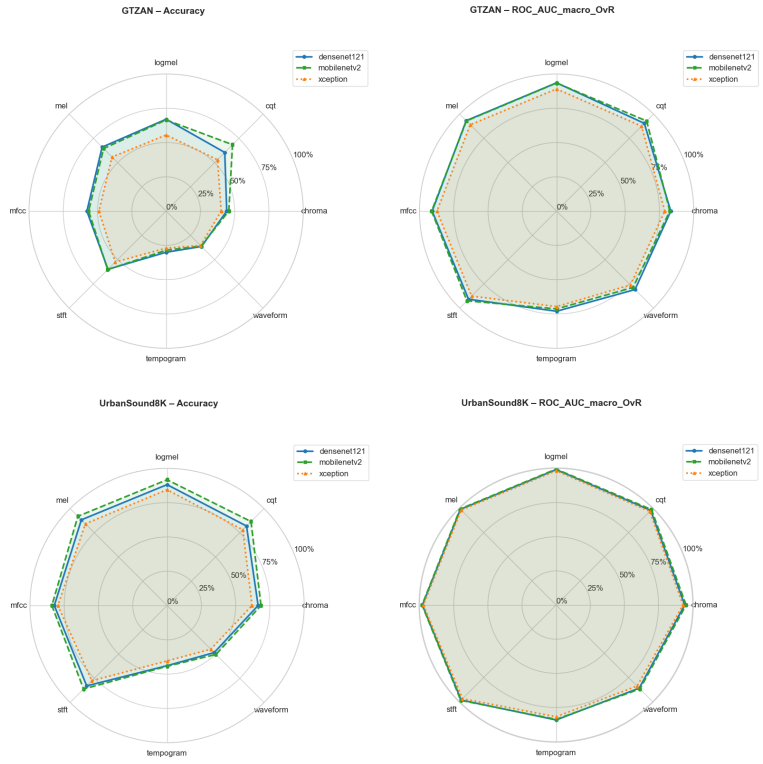

Figure 4: Spider plots visualizing performance among all audio representations, datasets and models.

previous metrics consistently show how Tempogram and Waveform underperform significantly, while the Log-Mel, Mel, CQT, and STFT cause superior results regardless of the setup and settings.

Finally, the rankings in Figure 7 offer concise visualization of how the performance is structured based on the audio representation applied, model, and dataset. It can be noticed that even though the GTZAN dataset contributed to less favourable results across all metrics due to the low data as for CNNs,

the pattern among audio representations and their perfomance ramained unchanged and similar to the one in the case of UrbanSound8K.

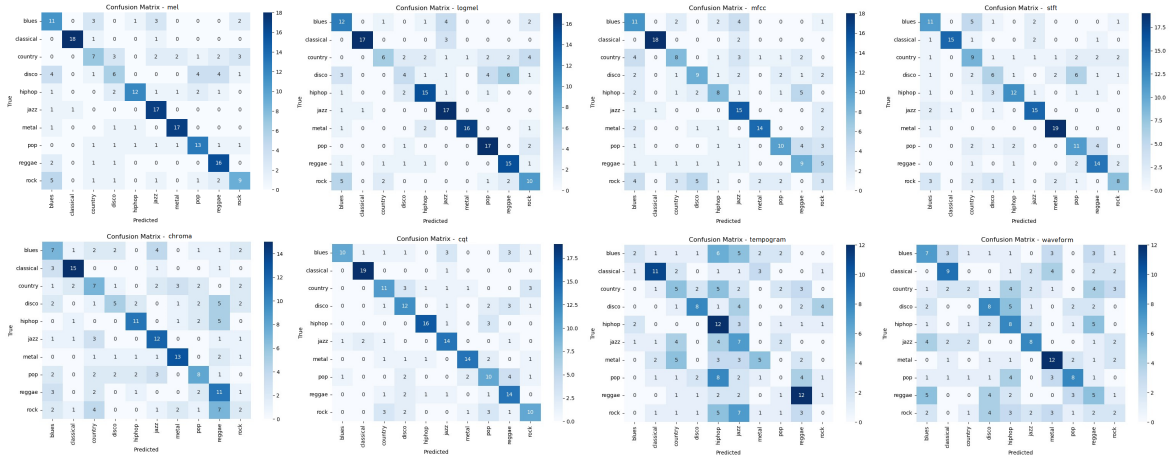

Figure 5: Confusion matrices for the best performing CNN - MobileNetV2 for GTZAN dataset.

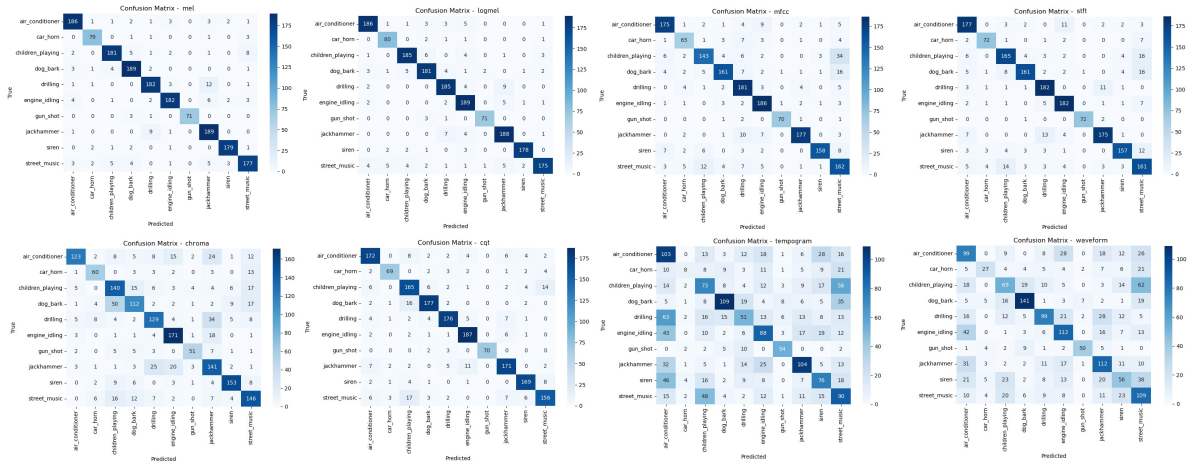

Figure 6: Confusion matrices for the best performing CNN - MobileNetV2 for UrbanSound8K dataset.

5. Conclusions and Future Work

This study provided a systematic evaluation of various audio representaitons in the context of CNN-based semantic classification tasks. In the pursuit of finding out if any audio representaiton yields consistently better performance across various CNN models and distinct datasets, this study assessed 8 commonly used time-frequency, alternative and signal-domain visual encodings, namely Mel, Log-Mel, MFCC, STFT, Chroma, CQT, Tempogram, and Waveform. The experiments were carried out across two datasets - GTZAN for music genre classification and UrbanSound8K for environmental sound classification. The obtained results highlighted the superior performance of the Log-Mel, Mel, CQT, and STFT representations across all tested scenarios. These representations demonstrated higher average classification scores in comparison to alternative encodings. On the other hand, representations like Tempogram and Waveform, showed significantly weaker results in this setup. The observed pattern in performance was similar across all models and regardless of dataset, so the rank of audio representaitons

Figure 7: Rankings of spectrograms and alternative audio representations among three metrics - Accuracy, Macro ROC-AUC OvR, and Macro F1.

remained stable. This observation suggests that the choice of the spectrogram type may have a greater impact on performance outcomes than the specific CNN architecture employed.

While this study provides evaluation of various audio representations in diverse scenarios, several limitations should be acknowledged. Other potentially relevant architectures, such as Convolutional Recurrent Neural Networks (CRNNs), Vision Transformers (ViTs), or attention-based models, which

were not included in this analysis could offer additional insights. Their incorporation could reveal different patterns in terms of representation effectiveness. Future work could extend this evaluation toward more recent model architectures, such as Transformer-based approaches, hybrid architectures, or additional audio feature engineering. In audio classification new classes may occur, making the approach of unsupervised learning and continual novel class discovery (CNCD) a promising approach in zero-shot like models[23]. For improved audio visual representations a high-resolution spectograms with uncertanity modeling could be applied [24], similarly as in the case of high dynamic range images. Exploring data augmentation strategies and their interplay with different representations may also further improve performance, particularly in low-resource scenarios. For the purpose of improving image quality, which may significantly influence the CNN-based model performance in the task of sound classification, another approach could provide an innovative strategy by enhancing visual features[25]. In case of noise or overlapping regions of audio recordings other strategies may be applied to mitigate their influence on classification performance [26], for e.g. through adaptive region selection. The careful selection of audio representation proves to be a vital choice, and Log-Mel, Mel, CQT, and STFT encodings remain a highly reliable option for CNN-based sound classification.

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] K. Zaman, M. Sah, C. Direkoglu, M. Unoki, A survey of audio classification using deep learning, IEEE Access 11 (2023) 106620–106649.

[2] Y. Zheng, R. Zhang, S. Atito, S. Yang, W. Wang, Y. Mei, Asit-crnn: A method for sound event detection with fine-tuning of self-supervised pre-trained asit-based model, Digital Signal Processing 160 (2025) 105055. URL: https://www.sciencedirect.com/science/article/pii/S1051200425000776. doi:https://doi.org/10.1016/j.dsp.2025.105055.

[3] W. Mu, B. Yin, X. Huang, J. Xu, Z. Du, Environmental sound classification using temporal-frequency attention based convolutional neural network, Scientific Reports 11 (2021) 21552.

[4] J. Salamon, J. P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification, IEEE Signal processing letters 24 (2017) 279–283.

[5] F. Wolf-Monheim, Spectral and rhythm features for audio classification with deep convolutional neural networks, arXiv preprint arXiv:2410.06927 (2024).

[6] M. Dong, Convolutional neural network achieves human-level accuracy in music genre classification, arXiv preprint arXiv:1802.09697 (2018).

[7] K. K. Jena, S. K. Bhoi, S. Mohapatra, S. Bakshi, A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis, Neural Computing and Applications 35 (2023) 11223–11248.

[8] M. Ahmed, U. Rozario, M. M. Kabir, Z. Aung, J. Shin, M. Mridha, Musical genre classification using advanced audio analysis and deep learning techniques, IEEE Open Journal of the Computer Society (2024).

[9] Y. Zhang, T. Li, Music genre classification with parallel convolutional neural networks and capuchin search algorithm, Scientific Reports 15 (2025) 9580.

[10] H. Zhao, C. Zhang, B. Zhu, Z. Ma, K. Zhang, S3t: Self-supervised pre-training with swin transformer for music classification, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 606–610.

[11] H. Li, S. Xue, J. Zhang, Combining cnn and classical algorithms for music genre classification, 2018.

[12] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on speech and audio processing 10 (2002) 293–302.

[13] J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound research, in: 22nd ACM International Conference on Multimedia (ACM-MM'14), Orlando, FL, USA, 2014, pp. 1041–1044.
[14] B. McFee, et al., librosa/librosa: 0.11.0, 2025. doi:10.5281/zenodo.15006942.
[15] R. W. Schafer, A. V. Oppenheim, Discrete-time signal processing, Prentice Hall, 2010.
[16] Y. Astuti, R. Hidayat, A. Bejo, A mel-weighted spectrogram feature extraction for improved speaker recognition system., International Journal of Intelligent Engineering & Systems 15 (2022).
[17] T. Tran, J. Lundgren, Drill fault diagnosis based on the scalogram and mel spectrogram of sound signals using artificial intelligence, Ieee Access 8 (2020) 203655–203666.
[18] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3d log-mel spectrograms with deep learning network, IEEE access 7 (2019) 125868–125881.
[19] N. Ahmed, T. Natarajan, K. R. Rao, Discrete cosine transform, IEEE transactions on Computers 100 (2006) 90–93.
[20] D. Ellis, Chroma feature analysis and synthesis, Resources of laboratory for the recognition and organization of speech and Audio-LabROSA 5 (2007).
[21] Y. D. Mistry, G. K. Birajdar, A. M. Khodke, Time-frequency visual representation and texture features for audio applications: a comprehensive review, recent trends, and challenges, Multimedia Tools and Applications 82 (2023) 36143–36177.
[22] P. Grosche, M. Müller, F. Kurth, Cyclic tempogram—a mid-level tempo representation for musicsignals, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 5522–5525.
[23] Q. Yan, Y. Yang, Y. Dai, X. Zhang, K. Wiltos, M. Woźniak, W. Dong, Y. Zhang, Clip-guided continual novel class discovery, Knowledge-Based Systems 310 (2025) 112920.
[24] Q. Yan, H. Wang, Y. Ma, Y. Liu, W. Dong, M. Woźniak, Y. Zhang, Uncertainty estimation in hdr imaging with bayesian neural networks, Pattern Recognition 156 (2024) 110802.
[25] X. Yang, Y. Yang, S. Ma, Z. Li, W. Dong, M. Woźniak, Samt-generator: A second-attention for image captioning based on multi-stage transformer network, Neurocomputing 593 (2024) 127823.
[26] Y. Hu, A. Niu, J. Sun, Y. Zhu, Q. Yan, W. Dong, M. Woźniak, Y. Zhang, Dynamic center point learning for multiple object tracking under severe occlusions, Knowledge-Based Systems 300 (2024) 112130.