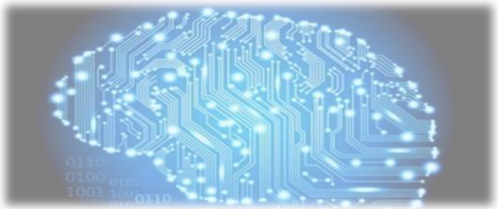
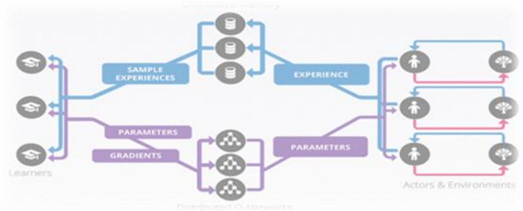




Musical Genre Classification

AutoML 연구실
석사 과정 조성현
22.01.11

To Do List

- Last Week
  - Code analysis
  - Comparison result

- This Week
  - Comparison result
  - Dataset transform
  - Dataset Integrity

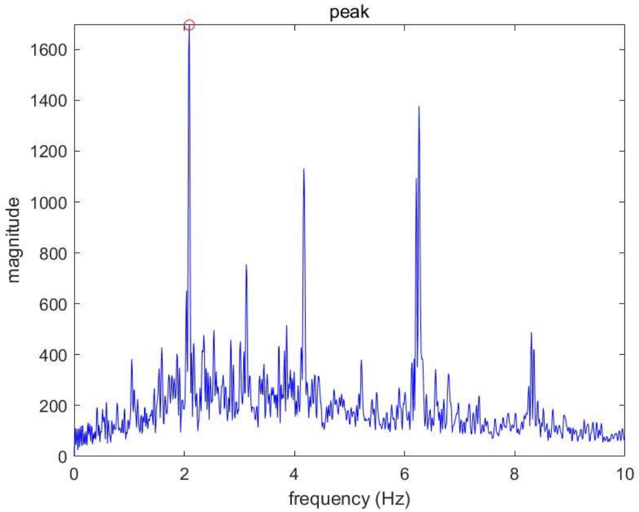
The standard statistics of the datasets

Author	Year	Neural Network	# of employed datasets	Employed Dataset	Input Type	Evaluation Metric	Ref.
Massoudi	2021	CNN	1	Urban Sound Classification Challenge	MFCC	Accuracy	[1]
Ghildiyal	2020	CNN	1	GTZAN	Mel-Spectrogram	Accuracy	[2]
Palanisamy	2020	CNN	3	ESC-50, GTZAN, UrbanSound8K	Log Mel-Spectrogram	Accuracy	[3]
Yu	2019	CNN	2	GTZAN, Extended Ballroom	STFT Spectrogram	Accuracy	[4]
Won	2020	CNN	3	MagnaTagATune, Million Song Dataset, MTG-Jamendo	Mel-Spectrogram	ROC-AUC, PR-AUC	[5]
Bisharad	2019	CNN	1	GTZAN	Mel-Spectrogram	Precision, Recall, F1 score	[6]
Zhang	2019	RNN	1	GTZAN	MFCC, chromagram	Accuracy	[7]
Chillara	2019	CNN	1	FMA-SMALL	Spectrogram	Accuracy	[8]
Zhang	2016	CNN	1	GTZAN	STFT Spectrogram	Accuracy	[9]
Choi	2017	CRNN	1	Million Song Dataset	Log Mel-Spectrogram	ROC-AUC	[10]
Su	2019	CNN	1	UrbanSound8k	Log Mel-Spectrogram, MFCC	Accuracy	[11]
Choi	2016	CNN	2	MagnaTagATune, Million Song Dataset	Mel-Spectrogram, MFCC, STFT	ROC-AUC	[12]
Li	2018	CNN	3	ESC-10, ESC-50, UrbanSound8k	Log Mel-Spectrogram	Accuracy	[13]
Li	2021	CNN	4	FMA, GTZAN5, GTZAN10, EMA	STFT Spectrogram	Accuracy	[14]

MIRToolbox

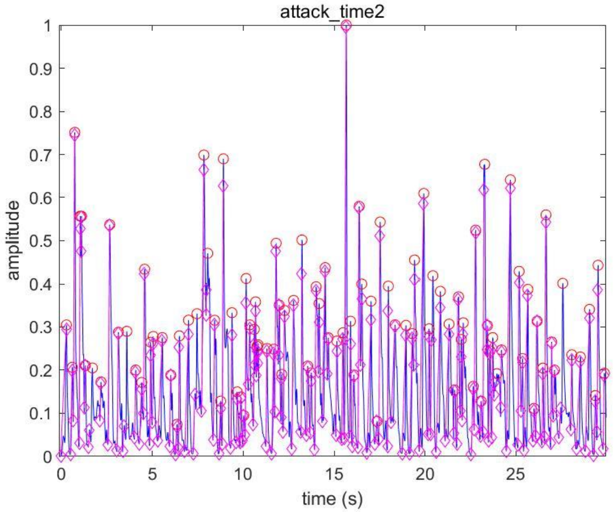
X_axis = frequency

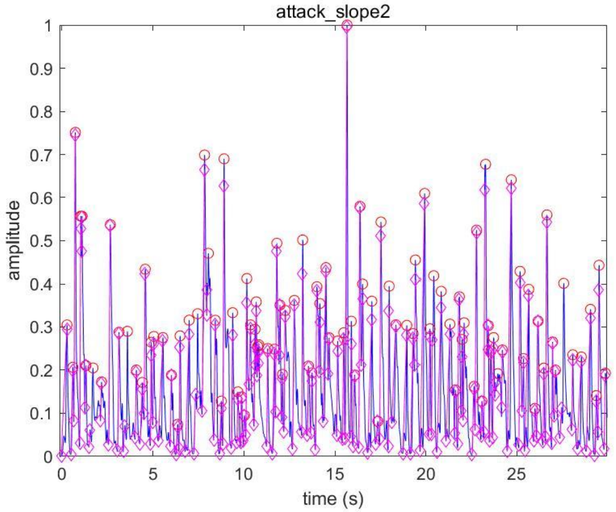
Fluctuation



X_axis = time

Rhythm

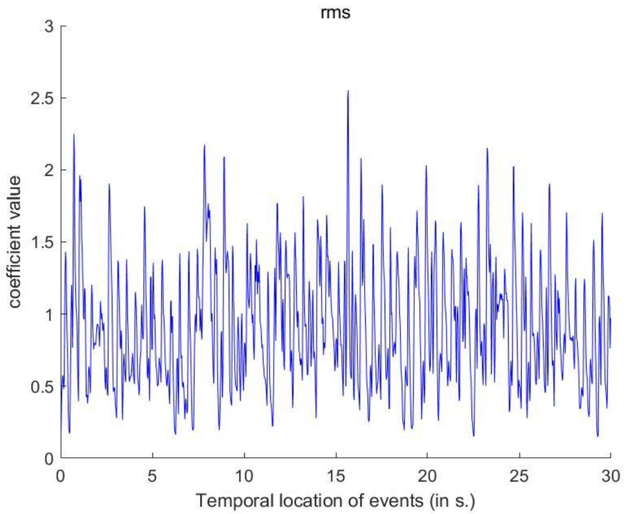




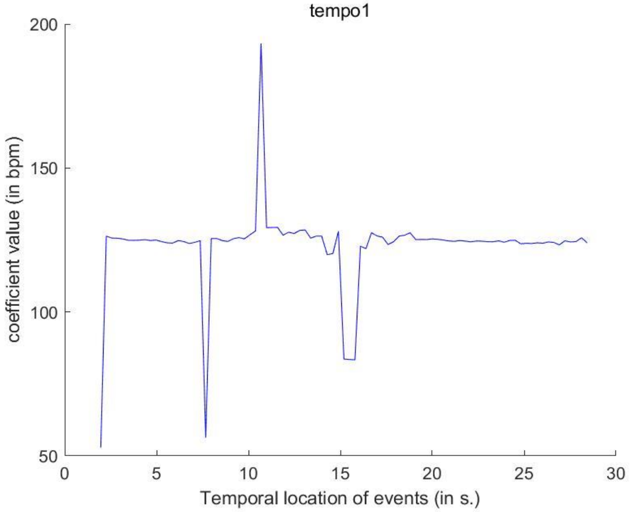
MIRToolbox

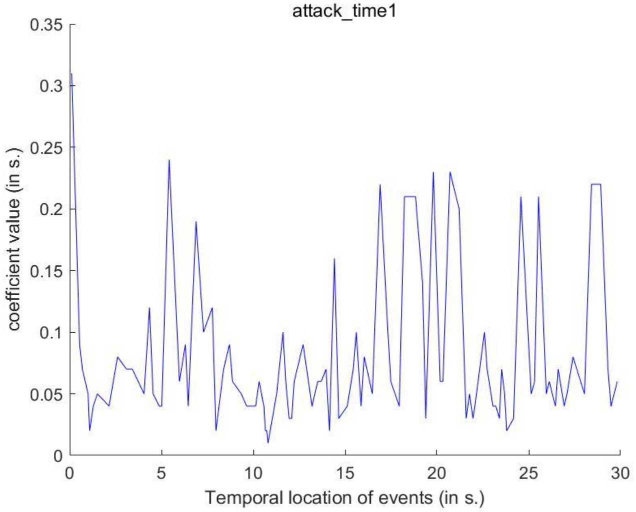
X_axis = Temporal location of events (in s.)

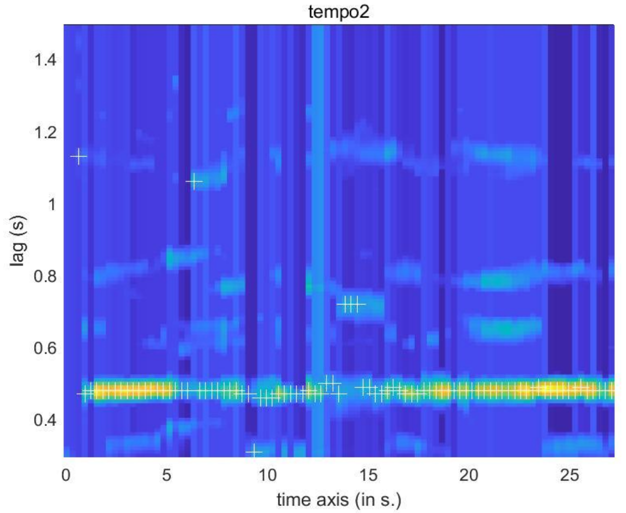
Dynamics

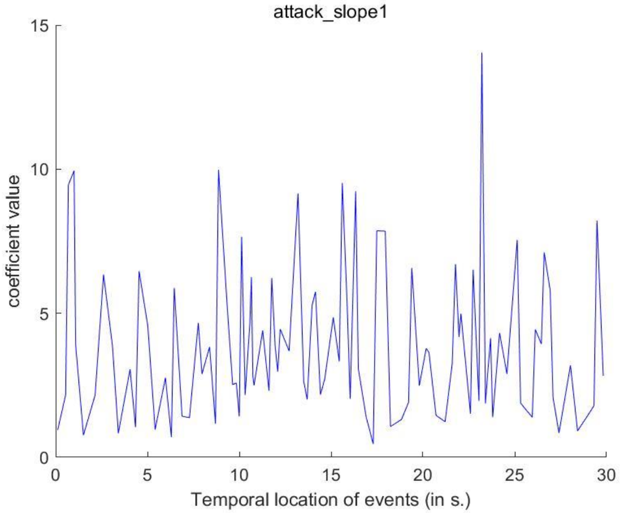


Rhythm





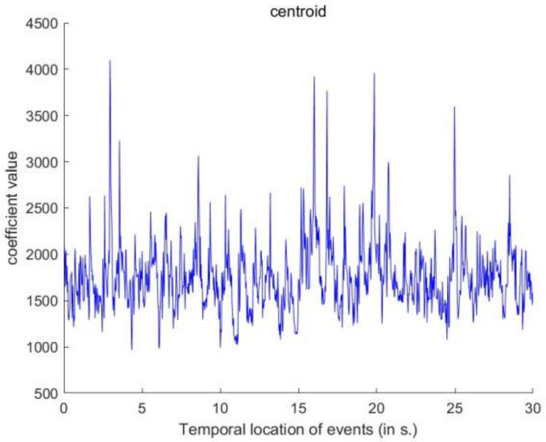


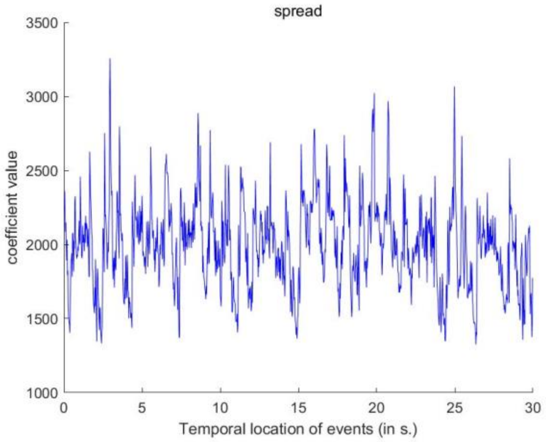


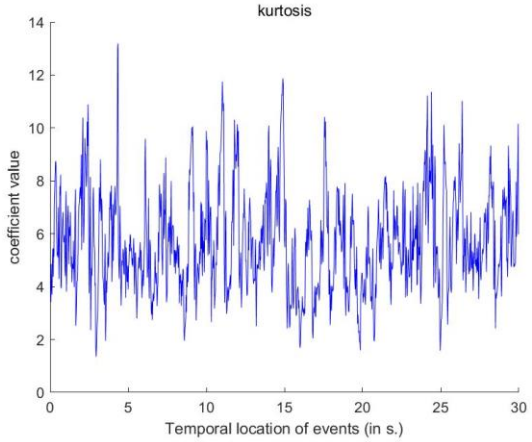
MIRToolbox

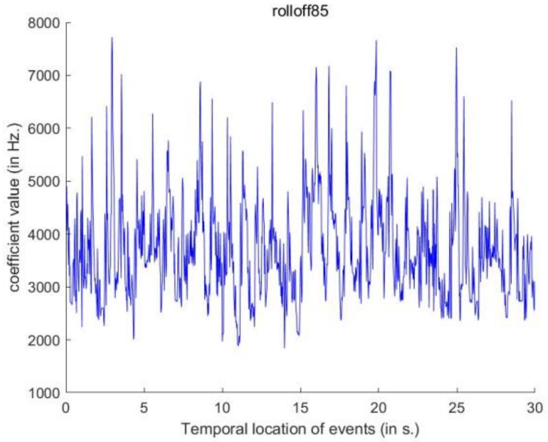
X_axis = Temporal location of events (in s.)

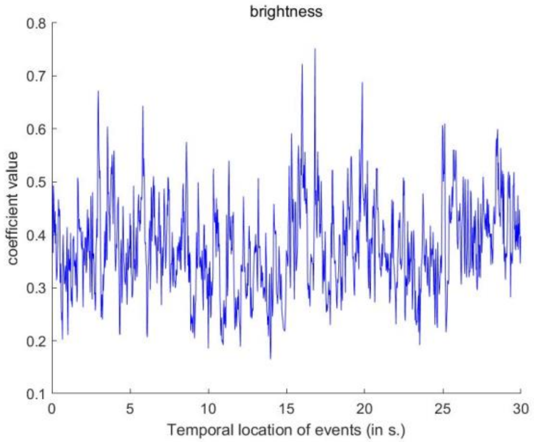
Spectral

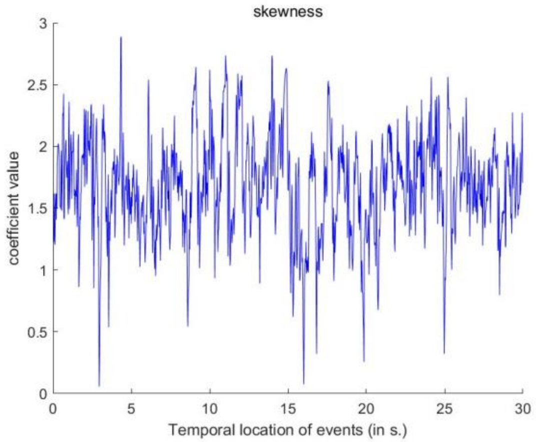


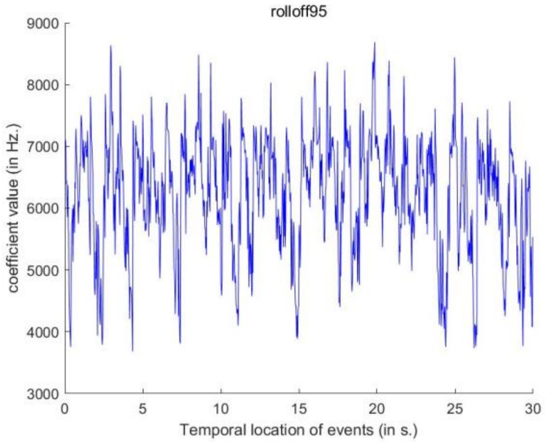








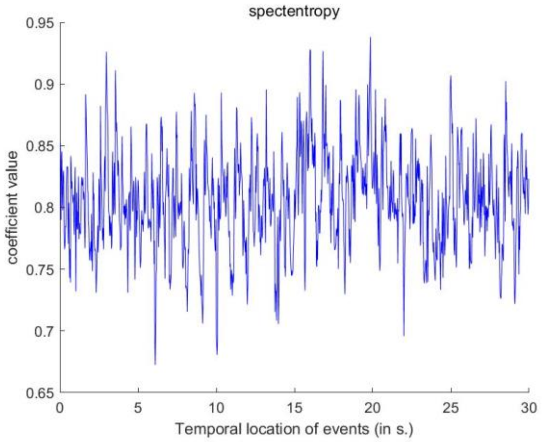


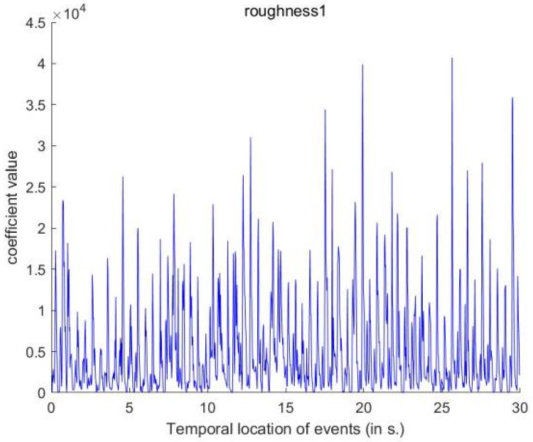


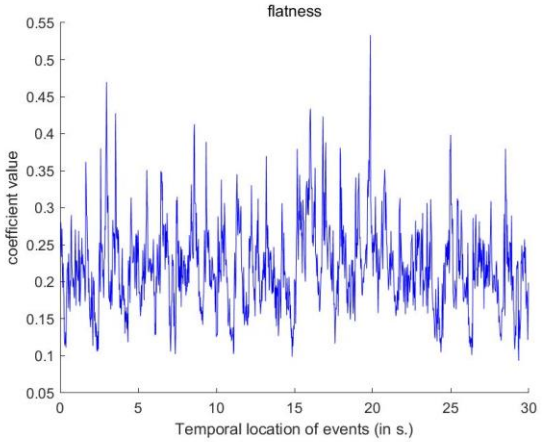
MIRToolbox

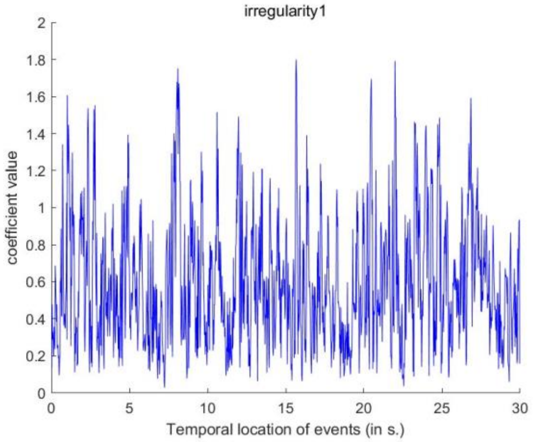
X_axis = Temporal location of events (in s.)

Spectral

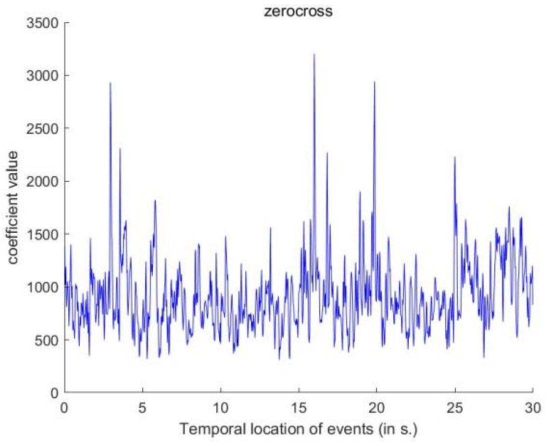


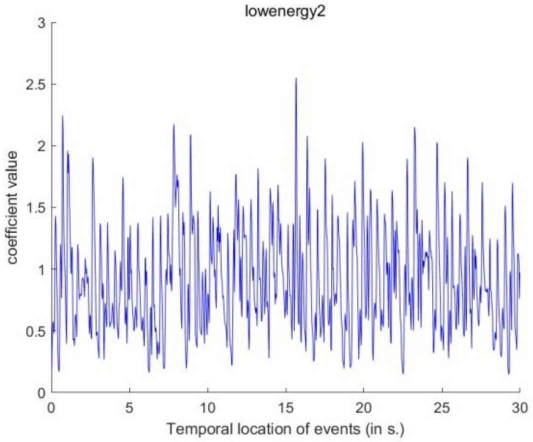


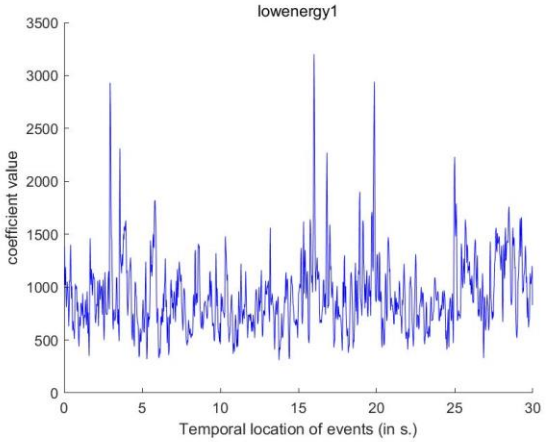


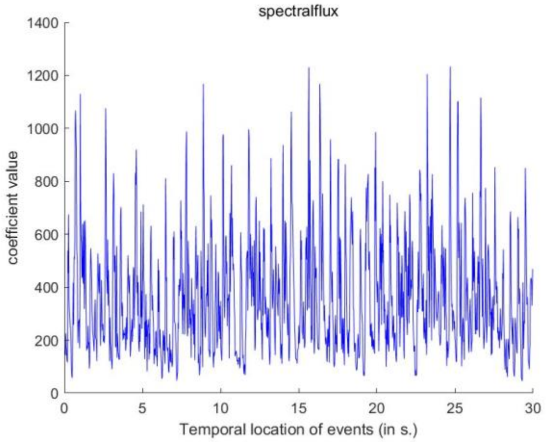


Timbre





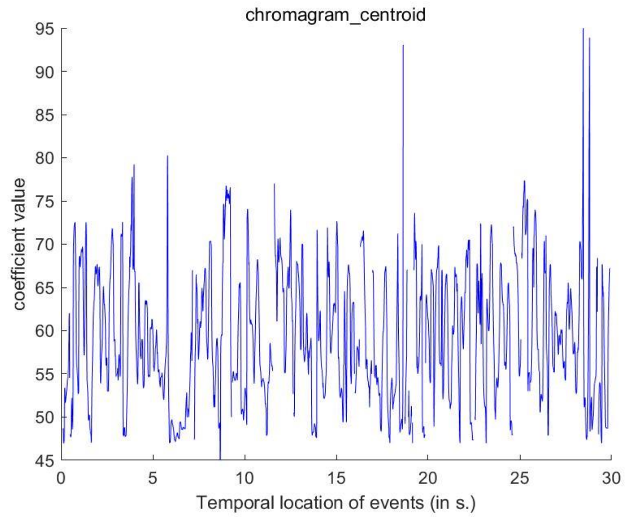


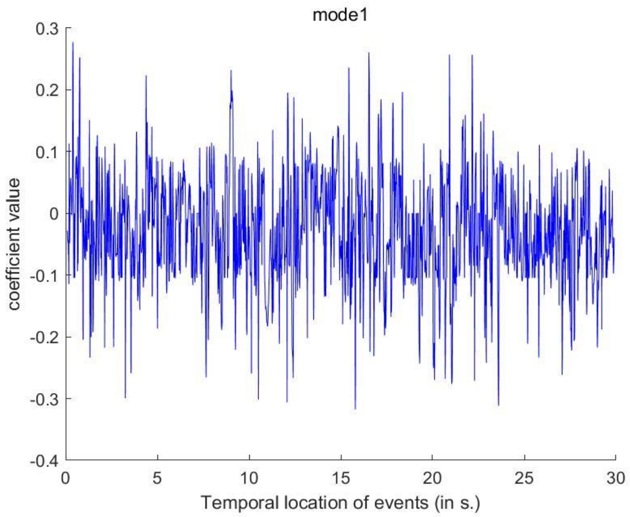


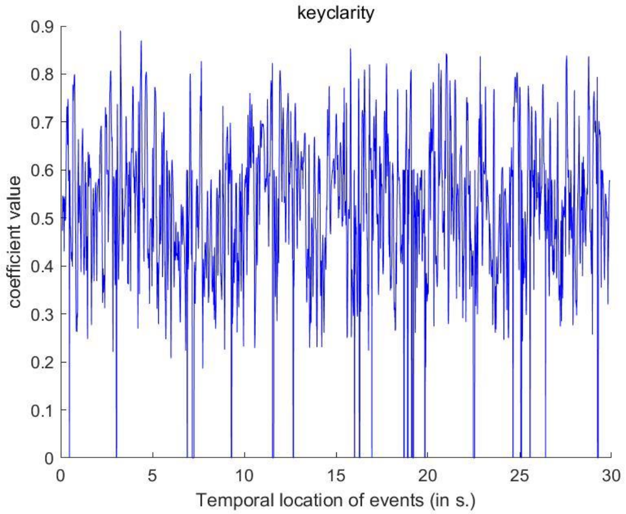
MIRToolbox

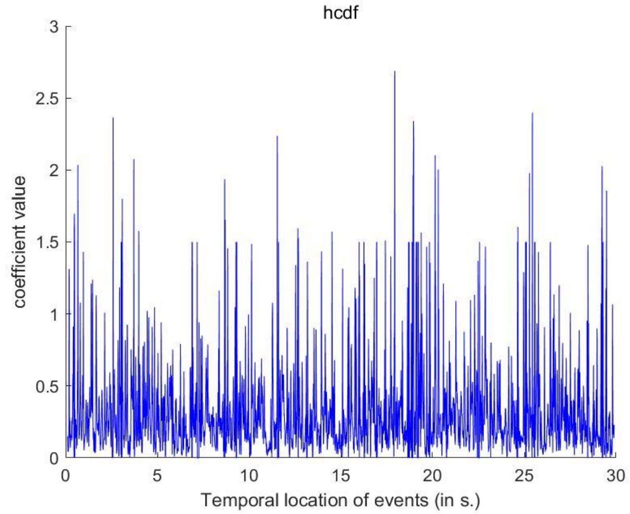
X_axis = Temporal location of events (in s.)

Tonal





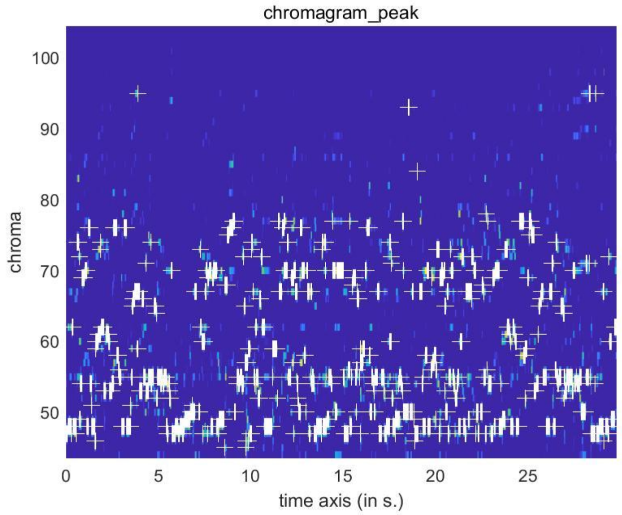


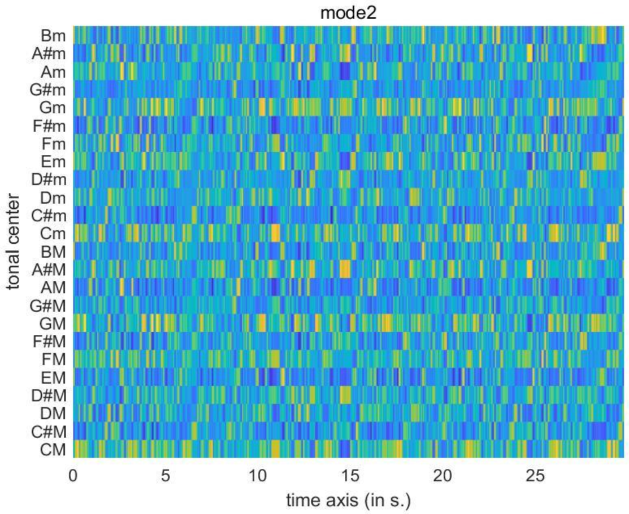


MIRToolbox

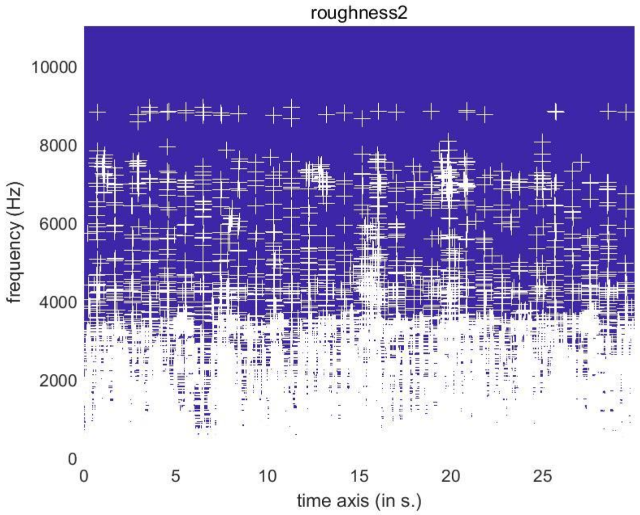
X_axis = time axis (in s.)

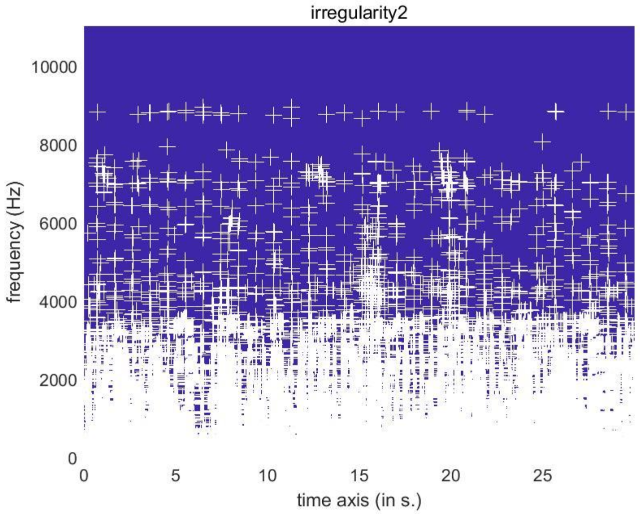
Tonal





Spectral

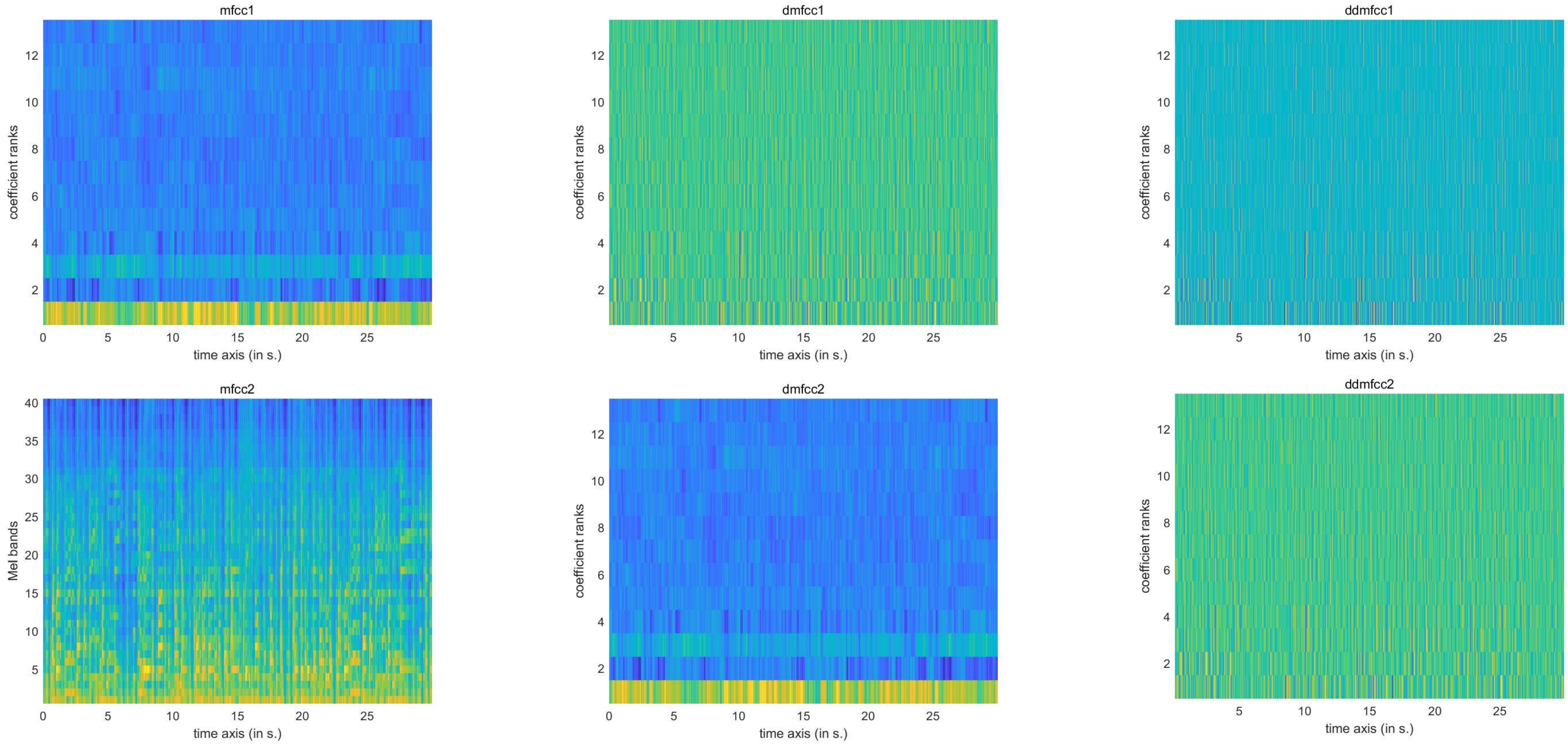




MIRToolbox

X_axis = time axis (in s.)

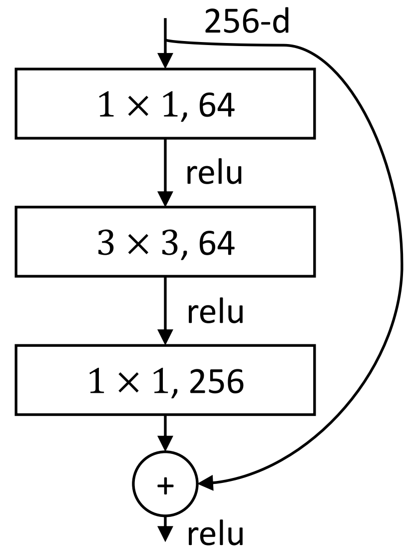
Spectral



Code Analysis

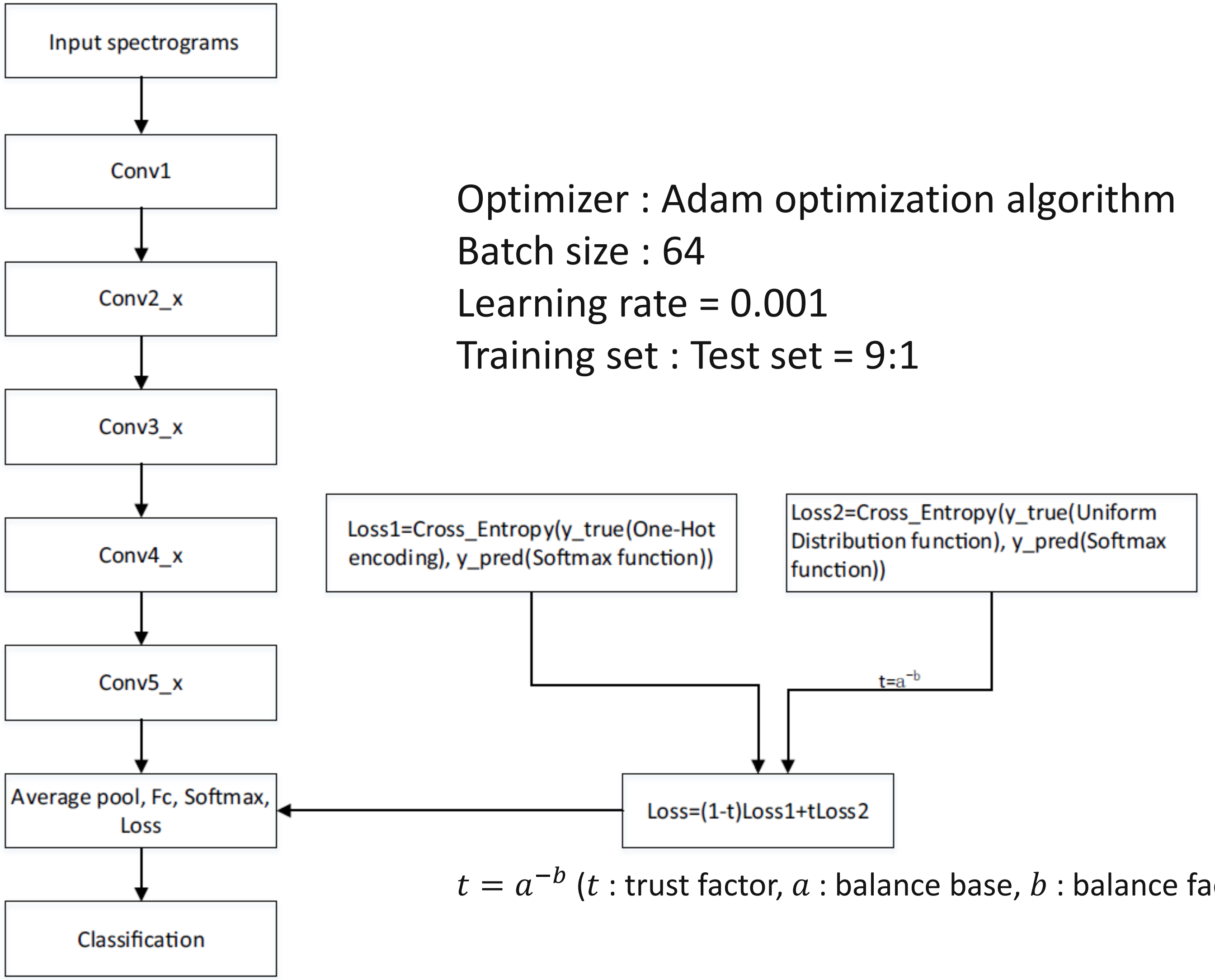
[14] Li, J., Han, L., Li, X., Zhu, J., Yuan, B., & Gou, Z. (2021). An evaluation of deep neural network models for music classification using spectrograms. *Multimedia Tools and Applications*, 1-27.

layer name	ouput size	50-layer
conv1	112×112	7×7, 64, stride 2
conv2 _x	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
conv5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax
FLOPs		3.8×10^9



A “bottleneck” building block for ResNet-50

Code Analysis



Optimizer : Adam optimization algorithm
Batch size : 64
Learning rate = 0.001
Training set : Test set = 9:1

$t = a^{-b}$ (t : trust factor, a : balance base, b : balance factor)

Code Analysis

Dataset	Types(the number of music audios)	Duration of each audio	Types(the number of music spectrograms)	Total number of music spectrograms
FMA	Classical(300) Electronic(300) Jazz(306) Rock(300)	30s	Classical(1800) Electronic(1800) Jazz(1836) Rock(1800)	7236
GTZAN_4	Classical(100) Disco(100) Jazz(100) Rock(100)	30s	Classical(600) Disco(600) Jazz(600) Rock(600)	2400
GTZAN_10	Blues(100) Classical(100) Country(100) Disco(100) Hiphop(100) Jazz(100) Metal(100) Pop(100) Reggae(100) Rock(100)	30s	Blues(600) Classical(600) Country(600) Disco(600) Hiphop(600) Jazz(600) Metal(600) Pop(600) Reggae(600) Rock(600)	6000
EMA	Angry(160) Happy(188) Quiet(160) Sad(133)	60s	Angry(1574) Happy(1833) Quiet(1536) Sad(1307)	6250

Comparison Results

ResNet50_trust					
Dataset (Genre)	$t = 10^{-1}$	$t = 10^{-2}$	$t = 10^{-3}$	$t = 10^{-4}$	$t = 10^{-5}$
Ballroom	-	44.29	-	-	-
Ballroom-Extended	-	70.81	-	-	-
emoMusic -G	-	-	-	-	-
Emotify-G	-	-	-	-	-
FMA-MEDIUM	-	-	-	-	-
FMA-SMALL	-	-	-	-	-
GiantStepsKey-G	-	-	-	-	-
GMD-G	-	-	-	-	-
GTZAN	65.00	71.00	67.00	67.00	68.00
HOMBURG	-	-	-	-	-
ISMIR04	-	-	-	-	-
MICM	-	37.72	-	-	-
Seyerlehner:Unique	-	-	-	-	-
Tropical Genres	-	-	-	-	-
Avg . Rank	-	-	-	-	-

Loss = (1-t)Loss1 + tLoss2

$t = a^{-b}$ (t : trust factor, a : balance base, b : balance factor)

Comparison Results

Dataset (Genre)	Proposed	ResNet50_trust	-	-
Ballroom	-	-	-	-
Ballroom-Extended	-	-	-	-
emoMusic -G	-	-	-	-
Emotify-G	-	-	-	-
FMA-MEDIUM	-	-	-	-
FMA-SMALL	-	-	-	-
GiantStepsKey-G	-	-	-	-
GMD-G	-	-	-	-
GTZAN	-	71.00 (t =2)	-	-
HOMBURG	-	-	-	-
ISMIR04	-	-	-	-
MICM	-	-	-	-
Seyerlehner:Unique	-	-	-	-
Tropical Genres	-	-	-	-
Avg . Rank	-	-	-	-

Dataset Transform

Dataset (Genre)	STFT	Mel Spectrogram	MFCC
Ballroom	O	O	O
Ballroom-Extended	O	O	O
emoMusic -G	-	-	-
Emotify-G	-	-	-
FMA-MEDIUM	-	-	-
FMA-SMALL	O	O	O
GiantStepsKey-G	-	-	-
GMD-G	-	-	-
GTZAN	O	O	O
HOMBURG	O	-	-
ISMIR04	O	O	O
MICM	O	O	O
Seyerlehner:Unique	O	-	-
Tropical Genres	O	O	O

Reference

[1] Massoudi, M., Verma, S., & Jain, R. (2021, January). Urban Sound Classification using CNN. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (pp. 583-589). IEEE.
[2]Ghildiyal, A., Singh, K., & Sharma, S. (2020, November). Music genre classification using machine learning. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1368-1372). IEEE.
[3] Palanisamy, K., Singhania, D., & Yao, A. (2020). Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*.
[4] Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., & Feng, L. (2020). Deep attention based music genre classification. *Neurocomputing*, *372*, 84-91.
[5] Won, M., Ferraro, A., Bogdanov, D., & Serra, X. (2020). Evaluation of cnn-based automatic music tagging models. *arXiv preprint arXiv:2006.00751*.
[6] Bisharad, D., & Laskar, R. H. (2019, October). Music Genre Recognition Using Residual Neural Networks. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (pp. 2063-2068). IEEE.
[7] Zhang, S., Gu, H., & Li, R. (2019). MUSIC GENRE CLASSIFICATION: NEAR-REALTIME VS SEQUENTIAL APPROACH.
[8] Chillara, S., Kavitha, A. S., Neginhal, S. A., Haldia, S., & Vidyullatha, K. S. (2019). Music genre classification using machine learning algorithms: a comparison. *Int Res J Eng Technol*, *6*(5), 851-858.
[9] Zhang, W., Lei, W., Xu, X., & Xing, X. (2016, September). Improved music genre classification with convolutional neural networks. In *Interspeech* (pp. 3304-3308).
[10] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017, March). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2392-2396). IEEE.
[11] Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream CNN based on decision-level fusion. *Sensors*, *19*(7), 1733.
[12] Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*.
[13] Li, S., Yao, Y., Hu, J., Liu, G., Yao, X., & Hu, J. (2018). An ensemble stacked convolutional neural network model for environmental event sound recognition. *Applied Sciences*, *8*(7), 1152.
[14] Li, J., Han, L., Li, X., Zhu, J., Yuan, B., & Gou, Z. (2021). An evaluation of deep neural network models for music classification using spectrograms. *Multimedia Tools and Applications*, 1-27.