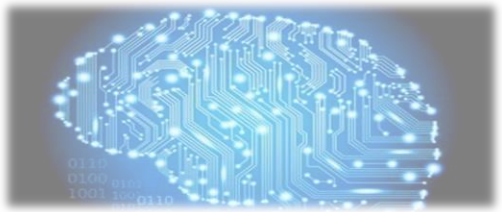
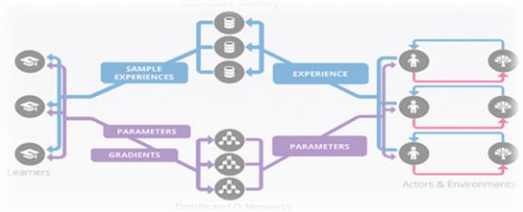



Result of ‘without GA’ and usage in other papers

AutoML 연구실
학부연구생 조성현
21.03.30

Without GA

- **Multilabel accuracy without Genetic Algorithm**

Dataset	Without GA
DEAM	0.007 ± 0.000
GTZAN	0.098 ± 0.005
HOMBURG	0.110 ± 0.002
FMA_SMALL	0.127 ± 0.002
IRMAS	0.195 ± 0.007
MER500	0.177 ± 0.026
emoMusic	0.127 ± 0.004
Soundtracks	0.112 ± 0.003
emotifymusic	0.224 ± 0.018
ballroom	0.122 ± 0.005
SeyerlehnerUnique	0.095 ± 0.004
MIREX-like_mood	0.037 ± 0.002
Medley-solos	0.122 ± 0.004
CAL500_emotion	0.262 ± 0.009
CAL500_ins_voc	0.187 ± 0.010
ISMIR04	0.124 ± 0.019
good-sounds	0.077 ± 0.008
MICM	0.142 ± 0.000
Tropical Genres	0.170 ± 0.027
PCMIR	0.101 ± 0.063
extendedBallroom	0.060 ± 0.004
giantsteps_genre	0.057 ± 0.006
giantsteps_key	0.058 ± 0.004
MagnaTagATune_emotion	0.011 ± 0.000
MagnaTagATune_genre	
MagnaTagATune_instrument	
MagnaTagATune_tempo	0.039 ± 0.001
GMD	0.037 ± 0.017
CAL500_genre	0.153 ± 0.009
FMA_MEDIUM	0.067 ± 0.004

DEAM

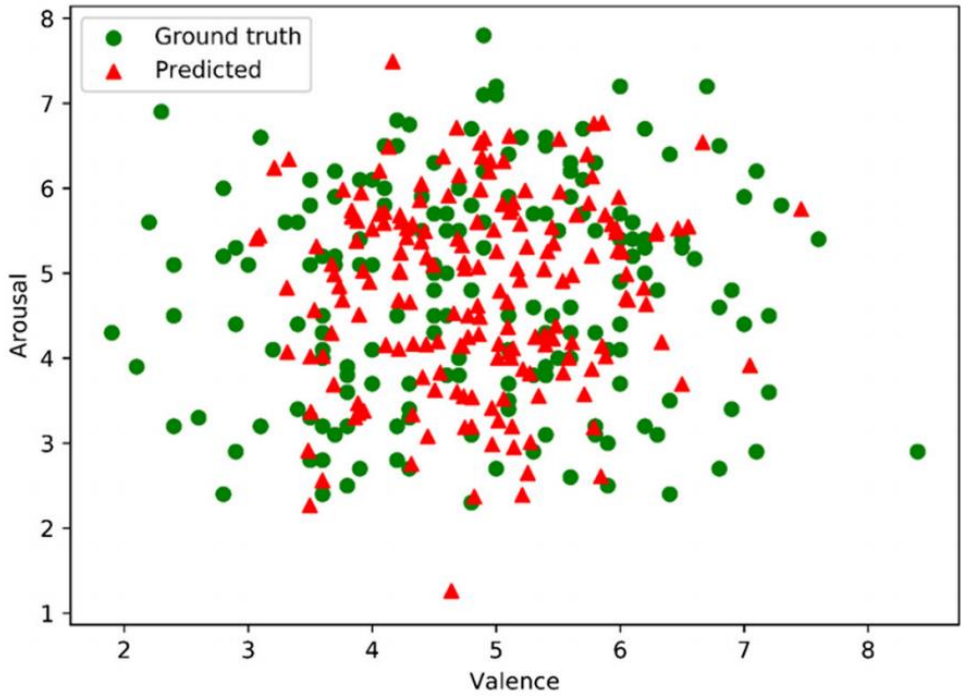


Fig. 7 Predicted versus actual valence and arousal indexes in the DEAM dataset

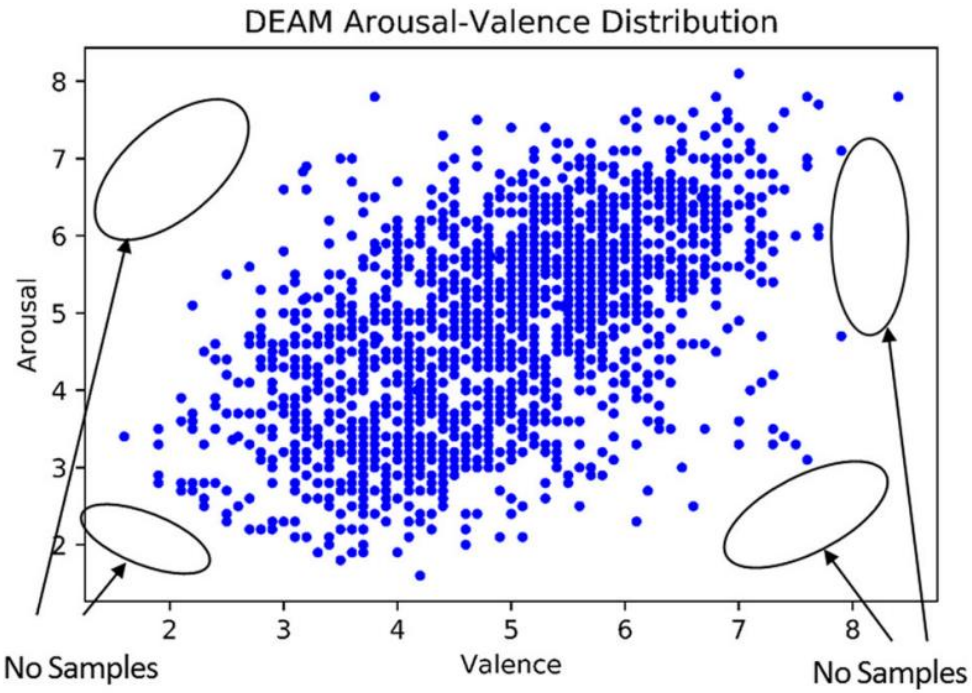


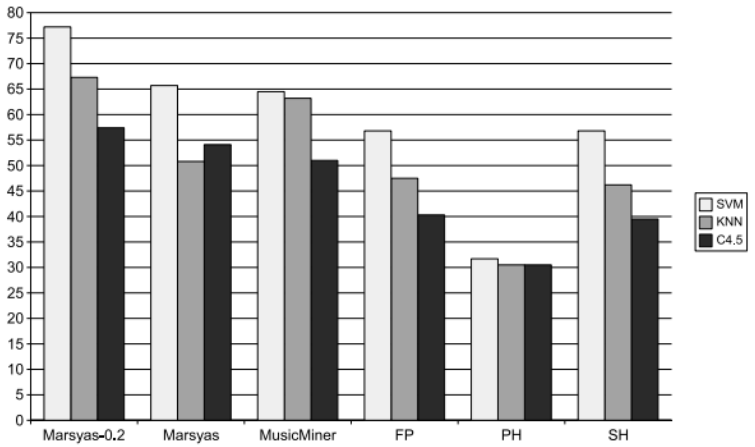
Fig. 16 Depiction of missing samples in DEAM dataset

Panwar, S., Rad, P., Choo, K.-K. R., & Roopaei, M. (2018). *Are you emotional or depressed? Learning about your emotional state from your music using machine learning. The Journal of Supercomputing.* doi:10.1007/s11227-018-2499-y

GTZAN

TABLE II
GENRE CONFUSION MATRIX

	cl	co	di	hi	ja	ro	bl	re	po	me
cl	**69**	0	0	0	1	0	0	0	0	0
co	0	**53**	2	0	5	8	6	4	2	0
di	0	8	**52**	11	0	13	14	5	9	6
hi	0	3	18	**64**	1	6	3	26	7	6
ja	26	4	0	0	**75**	8	7	1	2	1
ro	5	13	4	1	9	**40**	14	1	7	33
bl	0	7	0	1	3	4	**43**	1	0	0
re	0	9	10	18	2	12	11	**59**	7	1
po	0	2	14	5	3	5	0	3	**66**	0
me	0	1	0	1	0	4	2	0	0	**53**



Genres	Blues	Classical	Country	Disco	Hip Hop	Jazz	Metal	Pop	Reggae	Rock	Accuracies
Blues	863	7	31	10	6	25	38	5	4	45	83,46%
Classical	0	952	15	2	0	33	0	11	1	3	93.60%
Country	18	42	786	49	0	7	18	34	35	139	69.68%
Disco	12	30	35	639	24	2	13	46	97	72	65.87%
Hip Hop	7	1	0	62	805	0	3	30	44	6	84.02%
Jazz	19	9	12	10	0	852	0	9	6	14	91.51%
Metal	0	31	0	4	9	3	851	17	0	81	85.44%
Pop	0	4	7	40	39	2	0	758	31	46	81.76%
Reggae	6	4	23	58	54	4	6	20	710	57	75.37%
Rock	3	59	140	78	18	27	119	58	49	536	49.31%

$$\frac{\text{TP} + \text{TN}}{\text{TN} + \text{FP} + \text{TN}}$$

TABLE I
CONFUSION MATRIX SHOWING THE RESULTS OF 30 REPETITION OF THE GTZAN DATASET DIVIDED IN 66.7% TO TRAIN AND 33.3% TO TEST. THE CELL VALUES CORRESPOND TO THE ACCUMULATION OF CLASSIFICATION RESULTS FOR 30 REPETITION

Tzanetakis, G., & Cook, P. (2002). *Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293–302.* doi:10.1109/tsa.2002.800560

Moerchen, F., Mierswa, I., & Ultsch, A. (2006). *Understandable models Of music collections based on exhaustive feature generation with temporal statistics. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06.* doi:10.1145/1150402.1150523

HOMBURG

prediction \ true	Blues	Electronic	Jazz	Pop	HipHop	Rock	Folk/Country	Alternative	Funk/Soul
Blues	18	4	26	6	16	23	1	6	0
Electronic	2	17	12	6	11	9	0	10	0
Jazz	29	42	171	37	39	64	0	34	0
Pop	4	3	14	15	5	15	0	10	0
HipHop	10	21	25	15	187	21	0	10	0
Rock	55	19	58	31	38	358	1	60	0
Folk/Country	0	0	0	0	0	0	213	0	37
Alternative	2	7	13	6	4	14	1	15	0
Funk/Soul	0	0	0	0	0	0	6	0	10

Table 2: The confusion matrix for k-NN on the genre data.

	Accuracy
Random	26.72
C4.5	45.44
Naive Bayes	43.69
k-NN	53.23

Table 3: The accuracy for the genre classification.

	Accuracy
Random	44.07
C4.5	49.52
Naive Bayes	49.92
k-NN	49.63

Table 4: The averaged accuracy for the user tasks.

Homburg, H., Mierswa, I., Moller, B., Morik, K. and Wurst, M. (2005). *A BENCHMARK DATASET FOR AUDIO CLASSIFICATION AND CLUSTERING. ISMIR`05*

FMA

- Propose an 80/10/10% split into training, validation and test sets to make research on the FMA reproducible.

MAP	Artist-label Basic Model	Artist-label Siamese Model	Tag-label Model
GTZAN (fault-filtered)	0.4968	0.5510	0.5508
FMA small	0.2441	0.3203	0.3019
NAVER Korean	0.3152	0.3577	0.3576

Table 1. MAP results on feature similarity-based retrieval.

KNN	Artist-label Basic Model	Artist-label Siamese Model	Tag-label Model
GTZAN (fault-filtered)	0.6655	0.6966	0.6759
FMA small	0.5269	0.5732	0.5332
NAVER Korean	0.6671	0.6393	0.6898

Table 2. KNN similarity-based classification accuracy.

Linear Softmax	Artist-label Basic Model	Artist-label Siamese Model	Tag-label Model
GTZAN (fault-filtered)	0.6721	0.6993	0.7072
FMA small	0.5791	0.5483	0.5641
NAVER Korean	0.6696	0.6623	0.6755

Table 3. Classification accuracy of a linear softmax.

Park, J., Lee, J., Park, J., Ha, J. & Nam, J. (2018). *REPRESENTATION LEARNING OF MUSIC USING ARTIST LABELS. 19th International Society for Music Information Retrieval Conference (ISMIR)*

IRMAS

TABLE III. GENRE CLASSIFICATION WITHOUT SMOTE

Learning Model	Confusion Matrix				
Mid Term Features					
		Classical	Country Folk	Jazz	Pop
Decision Tree	Classical	441	20	173	128
	Country Folk	13	41	14	23
	Jazz	162	32	532	281
	Pop	137	43	271	594
Random Forest	Classical	639	0	87	36
	Country Folk	10	37	15	29
	Jazz	134	0	656	217
	Pop	138	0	168	739
Kernel SVM	Classical	549	0	151	62
	Country Folk	9	0	17	65
	Jazz	144	0	551	312
	Pop	138	0	219	688
Ensemble	Classical	620	1	85	56
	Country Folk	10	40	15	26
	Jazz	117	0	683	207
	Pop	115	1	161	768

Short Term Features					
Decision Tree	Classical	1181	30	424	305
	Country Folk	29	124	53	76
	Jazz	415	78	1291	695
	Pop	315	92	682	1473
Random Forest	Classical	1666	0	202	72
	Country Folk	28	104	49	101
	Jazz	379	0	1621	479
	Pop	342	0	399	1821
Kernel SVM	Classical	1487	0	338	115
	Country Folk	27	2	91	162
	Jazz	425	0	1358	696
	Pop	355	0	524	1683
Ensemble	Classical	1602	2	240	96
	Country Folk	24	90	71	97
	Jazz	334	5	1629	511
	Pop	295	4	382	1881

TABLE II
RECOGNITION PERFORMANCE FOR *IRMAS* DATASET.

Model / #param	Micro			Macro		
	P	R	F1	P	R	F1
Bosch *et al.*	0.504	0.501	0.503	0.41	0.455	0.432
Han *et al.* / 1446k	**0.655**	**0.557**	**0.602**	0.541	0.508	0.503
Single-layer / 62k	0.611	0.516	0.559	0.523	0.480	0.484
Multi-layer / 743k	0.650	0.538	0.589	**0.550**	**0.525**	**0.516**

Goel, S., Pangasa, R., Dawn, S., & Arora, A. (2018). Audio Acoustic Features Based Tagging and Comparative Analysis of its Classifications. 2018 Eleventh International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2018.8530512
Pons, J., Slizovskaia, O., Gong, R., Gomez, E., & Serra, X. (2017). *Timbre analysis of music audio signals with convolutional neural networks. 2017 25th European Signal Processing Conference (EUSIPCO).* doi:10.23919/eusipco.2017.8081710

MER500

ML Classifier	Feature Set (number of features) Accuracy in % FS1 (72)	FS2 (27)	FS3 (17)	FS4 (28)
NB	54.4	51.2	48.4	53.8
MLP	60.2	**70**	64	63.8
SVM	64	62.6	59.2	62
J-48	53.2	53.6	49.2	54.8
RF	61.2	60.8	57.6	60.6

Table 1. Accuracy of ML classifiers for different feature sets

	Romantic	**Exciting**	**Happy**	**Sad**	**Devotional**
Romantic	**43**	4	6	30	17
Exciting	0	**90**	8	2	0
Happy	3	10	**82**	4	1
Sad	20	4	2	**67**	7
Devotional	19	1	0	12	**68**

Table 2. Confusion matrix for maximum accuracy of 70%with FS2 for MLP classifier

Makarand Velankar and Parag Kulkarni. (20198). *MUSIC EMOTION RECOGNITION FOR INDIAN FILM MUSIC. ISMIR 2019.*

emoMusic

Table 2: To evaluate the estimation models from content features R^2 and Euclidean distances are reported for static estimation and Kendall Tau (ρ) is reported with distance for dynamic estimation. The reported measures on dynamic annotated data are averaged for all the clips. Baseline results are calculated by setting the target to the average score in the training set. The results that are significantly better (Wilcoxon test $p < 0.01$) than the baseline (averaged training targets) are indicated with asterisk (*)(Dist: mean absolute difference and Euclidean distance for the case of two dimensions)

	Static estimation on individual clips					Dynamic estimation on 40 samples				
Features	**Arousal**		**Valence**		**Both**	**Arousal**		**Valence**		**Both**
	Dist	R^2	Dist	R^2	Dist	Dist	ρ	Dist	ρ	Dist
All	$0.10 \pm 0.07^*$	0.54	0.12 ± 0.09	0.07	0.15 ± 0.09	$0.08 \pm 0.05^*$	0.15 ± 0.22	0.09 ± 0.06	0.05 ± 0.20	$0.13 \pm 0.06^*$
MFCC	$0.11 \pm 0.08^*$	0.34	0.12 ± 0.08	0.10	0.14 ± 0.09	$0.09 \pm 0.06^*$	0.12 ± 0.22	0.09 ± 0.06	0.03 ± 0.21	$0.14 \pm 0.07^*$
Shape	$0.13 \pm 0.08^*$	0.24	0.12 ± 0.08	0.12	0.13 ± 0.09	$0.10 \pm 0.06^*$	0.10 ± 0.22	0.09 ± 0.06	0.05 ± 0.25	$0.14 \pm 0.07^*$
Contrast	$0.11 \pm 0.08^*$	0.38	0.12 ± 0.08	0.08	0.15 ± 0.09	$0.09 \pm 0.06^*$	0.10 ± 0.22	0.09 ± 0.06	0.02 ± 0.23	$0.13 \pm 0.06^*$
Chroma	0.15 ± 0.09	0.02	0.12 ± 0.08	0.04	0.12 ± 0.09	0.11 ± 0.07	0.09 ± 0.20	0.09 ± 0.06	0.02 ± 0.19	0.16 ± 0.07
Baseline	0.15 ± 0.09	-	0.13 ± 0.09	-	0.12 ± 0.10	0.12 ± 0.07	0.05 ± 0.43	0.09 ± 0.06	-0.02 ± 0.59	0.16 ± 0.08

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C.-Y., & Yang, Y.-H. (2013). *1000 songs for emotional analysis of music. Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia - CrowdMM '13.* doi:10.1145/2506364.2506365

Soundtracks

- The ratings in the dataset range from 1 to 7.83 and were scaled by a factor of 0.1 before being used for our experiments.
- 10 runs with randomly selected train-test 80:20 splits

Layer	Classifier	Training–Testing Data Splitting	Accuracy Before Augmentation (%)	Accuracy After Augmentation (%)
Conv5	SVM	50%–50%	25.0	45.0
Conv5	SVM	60%–40%	25.6	46.3
Conv5	SVM	70%–30%	31.5	49.7
Conv5	Softmax	50%–50%	25.0	40.4
Conv5	Softmax	60%–40%	27.8	42.6
Conv5	Softmax	70%–30%	30.6	43.5
Fc6	SVM	50%–50%	32.2	71.5
Fc6	SVM	60%–40%	40.3	73.8
Fc6	SVM	70%–30%	40.7	74.4
Fc6	Softmax	50%–50%	36.7	69.6
Fc6	Softmax	60%–40%	38.8	71.1
Fc6	Softmax	70%–30%	46.3	72.5
Fc7	SVM	50%–50%	30.0	65.7
Fc7	SVM	60%–40%	36.1	69.2
Fc7	SVM	70%–30%	38.9	72.2
Fc7	Softmax	50%–50%	27.8	63.5
Fc7	Softmax	60%–40%	37.5	67.1
Fc7	Softmax	70%–30%	40.7	70.4
Fc8	SVM	50%–50%	31.1	61.3
Fc8	SVM	60%–40%	31.9	64.1
Fc8	SVM	70%–30%	33.3	67.9
Fc8	Softmax	50%–50%	30.0	59.8
Fc8	Softmax	60%–40%	36.1	60.6
Fc8	Softmax	70%–30%	38.9	66.0

Table 8 The classification results for each layers of the AlexNet on soundtracks dataset.

Approach	Accuracy
Ren *et al.* [18]	43.56
Panagakis and Kotropoulos [36]	39.44
Mo and Niu [37]	44.68
Proposed System (Before Data Augmentation)	46.3
Proposed System (After Data Augmentation)	75.5

Table 11 Performance comparison with other approaches on the soundtracks dataset.

Chowdhury, S., Vall, A., Haunschmid, V. & Widmer, G. (2019). *TOWARDS EXPLAINABLE MUSIC EMOTION RECOGNITION: THE ROUTE VIA MID-LEVEL FEATURES. International Society for Music Information Retrieval Conference 2019*
Bilal, M. and Aydilek, I. (2019) *Music Emotion Recognition by Using Chroma Spectrogram and Deep Visual Features. International Journal of Computational Intelligence Systems 2019*

emotifymusic

- Emotify, is different from MoodSwings in several respects: it uses a categorical emotional model, it collects data on induced (not perceived) emotion, and the measurements are discrete rather than continuous.
- In the Emotify dataset, lyrics are not comprised in the dataset and this is certainly a limiting factor in consideration of scenarios where lyrics are available for the music

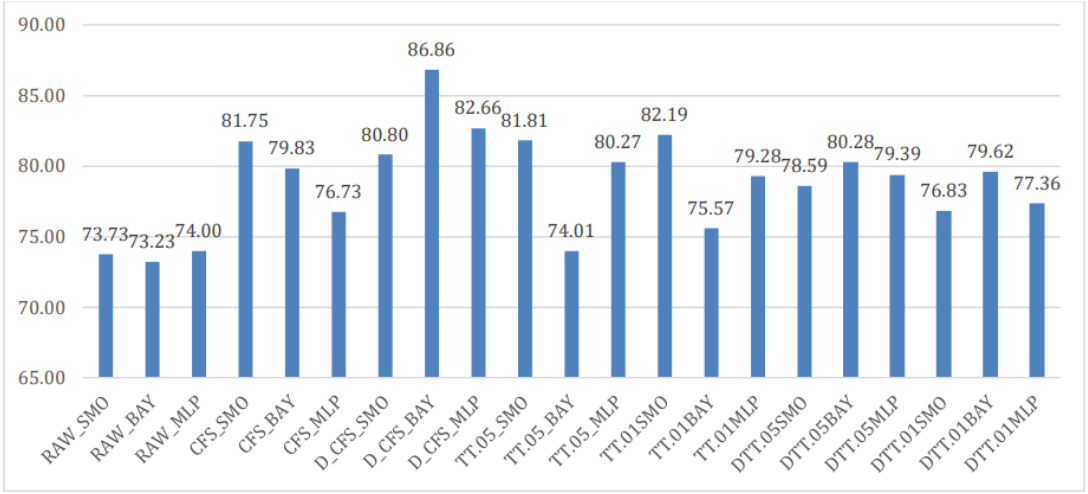


Figure 3. Classifiers mean efficiency using consensus threshold at 25%.

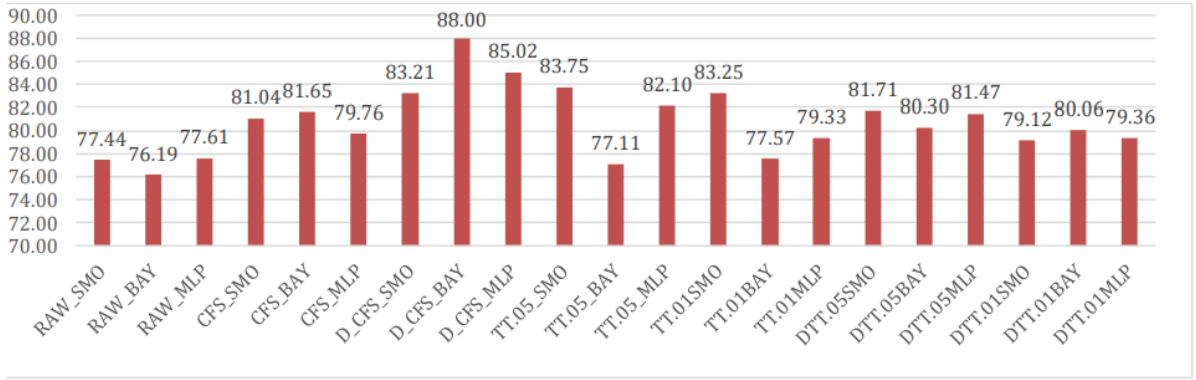


Figure 2. Classifiers mean efficiency using consensus threshold at 30%.

Paolizzo, F., Pichierri, N., Casali, D., Giardino, D., Matta, M. & Giovanni Costantini. (2019). *Multilabel Automated Recognition of Emotions Induced Through Music.*. Marie Sklodowska-Curie Actions Individual Fellowships Global Fellowships (MSCA-IF-GF) of the Horizon 2020

Ballroom

Methods	Preprocessing	Accuracy
MMCNN [35]	mel-spectrogram	87.6
MCLNN [30]	mel-spectrogram	90.4
Pons et al. [33]	mel-spectrogram	92.1
Marchand et al. [27]	MSS	93.1
SOTA [28]	MASSS	96.0
Ours	mel- spectrogram	**96.7**

Classification accuracy (%) on Ballroom dataset is compared across recently proposed methods

Genre	Precision	Recall rate	F-score
Cha Cha	100	96.2	98.0
Jive	97.9	94.1	96.0
Quickstep	96.4	100	98.1
Rumba	97.0	94.2	95.6
Samba	95.0	97.4	96.2
Tango	98.8	98.8	98.8
Viennese Waltz	98.5	94.3	96.4
Slow Waltz	94.3	100	97.0
Average	97.2	96.9	97.0

Precision (%), recall rate (%) and F-score(%) of each genre obtained on the Ballroom dataset

Liu, C., Feng, L., Liu, G., Wang, H. & Liu, S. (2020). *Bottom-up broadcast neural network for music genre classification . Multimedia Tools and Applications*

Extended ballroom

Methods	Preprocessing	Accuracy
Transform learning [3]	MFCC	86.7
RWCNN [34]	MFCC	89.8
BRNN+PCNNA [45]	STFT	92.7
DLR [16]	mel-spectrogram	93.7
SOTA [28]	MASSS	94.9
Ours	mel-spectrogram	**97.2**

Classification accuracy (%) on Extended Ballroom dataset is compared across recently proposed methods

Genre	Precision	Recall rate	F-score
Cha Cha	98.1	98.5	98.3
Foxtrot	98.8	99.6	99.2
Jive	98.5	99.4	98.9
Pasodoble	96.0	94.2	95.1
Quickstep	99.6	99.0	99.3
Rumba	96.7	94.5	95.6
Salsa	94.9	91.8	93.3
Samba	97.5	98.7	98.1
Slow Walz	80.7	71.1	75.6
Tango	99.0	95.3	97.1
Viennese Walz	96.8	97.7	97.3
Walz	93.1	98.1	95.5
Wcswing	78.5	57.8	66.6
Average	94.5	92.0	93.1

Precision (%), recall rate (%) and F-score of each genre obtained on the Extend Ballroom dataset

Liu, C., Feng, L., Liu, G., Wang, H. & Liu, S. (2020). *Bottom-up broadcast neural network for music genre classification . Multimedia Tools and Applications*

SeyerlehnerUnique

Table 4: Classification accuracy and standard error on the full training set using a 10-fold cross-validation.

Dataset	Base Features	Genre Embedding	Tag Embedding
Artists	0.323 +/- 0.010	0.310 +/- 0.005	0.338 +/- 0.007
GTZAN	0.748 +/- 0.010	0.671 +/- 0.014	0.754 +/- 0.015
Homburg	0.580 +/- 0.012	0.561 +/- 0.009	0.584 +/- 0.008
Unique	0.651 +/- 0.006	0.634 +/- 0.005	0.666 +/- 0.006

Table 5: Precision at 10 for the music similarity task on different feature spaces. The genre embedding is learned using the 3 other genre datasets. The tag embedding is learned on the Magnatagatune dataset.

Dataset	Base Features	Genre Embedding	Tag Embedding
Artists	0.15	0.19	0.19
GTZAN	0.48	0.52	0.53
Homburg	0.36	0.41	0.40
Unique	0.53	0.52	0.54

Fig. 6. The proposed confidence-based late fusion using various β. (a) GTZAN dataset. (b) Unique dataset. (c) MASD.

Table III. Comparison of the Fusion (Two Feature Types) and Nonfusion (One Feature Type) Methods

Method	GTZAN	Unique	MASD
MLVF+GSV (confidence-based late fusion)	**88.60%**	**77.66%**	41.76%
MLVF+GSV (probability-based late fusion by the *max* rule)	87.70%	76.98%	40.60%
MLVF+GSV (probability-based late fusion by the *sum* rule)	88.10%	77.14%	41.49%
MLVF+GSV (probability-based late fusion by the *prod* rule)	88.20%	77.43%	41.53%
MLVF+GSV (early fusion)	87.00%	77.50%	**41.90%**
MLVF	85.70%	75.67%	40.28%
GSV	80.10%	73.58%	34.73%

Hamel, P., Davies, M., Yoshii, K. & Goto, M. (2013). TRANSFER LEARNING IN MIR: SHARING LEARNED LATENT REPRESENTATIONS FOR MUSIC AUDIO CLASSIFICATION AND SIMILARITY . *ISMIR 2013*

MIREX-like_mood

Table 3. Confusion matrix for multi-modal datasets.

	C1	C2	C3	C4	C5
C1	63.6%	15.9%	4.5%	4.5%	11.4%
C2	20.9%	60.5%	11.6%	7.0%	0.0%
C3	4.2%	18.8%	64.6%	8.3%	4.2%
C4	12.1%	18.2%	9.1%	51.5%	9.1%
C5	12.0%	3.0%	12.0%	8.0%	64.0%

Panda, R., Malheiro, R., Rocha, B., liveira, A. & Paiva, R. (2013). *Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. CMMR'2013*

Medley-solos

	piano	violin	dist. guitar	female singer	clarinet	flute	trumpet	tenor sax.	average
bag-of-features and random forest	**99.7 (± 0.1)**	**76.2 (± 3.1)**	**92.7 (± 0.4)**	81.6 (± 1.5)	49.9 (± 0.8)	22.5 (± 0.8)	63.7 (± 2.1)	4.4 (± 1.1)	61.4 (± 0.5)
spiral (36k parameters)	86.9 (± 5.8)	37.0 (± 5.6)	72.3 (± 6.2)	84.4 (± 6.1)	61.1 (± 8.7)	30.0 (± 4.0)	54.9 (± 6.6)	52.7 (± 16.4)	59.9 (± 2.4)
1-d (20k parameters)	73.3 (± 11.0)	43.9 (± 6.3)	91.8 (± 1.1)	82.9 (± 1.9)	28.8 (± 5.0)	**51.3 (± 13.4)**	63.3 (± 5.0)	**59.0 (± 6.8)**	61.8 (± 0.9)
2-d, 32 kernels (93k parameters)	96.8 (± 1.4)	68.5 (± 9.3)	86.0 (± 2.7)	80.6 (± 1.7)	81.3 (± 4.1)	44.4 (± 4.4)	68.0 (± 6.2)	48.4 (± 5.3)	69.1 (± 2.0)
spiral & 1-d (55k parameters)	96.5 (± 2.3)	47.6 (± 6.1)	90.2 (± 2.3)	84.5 (± 2.8)	79.6 (± 2.1)	41.8 (± 4.1)	59.8 (± 1.9)	53.0 (± 16.5)	69.1 (± 2.0)
spiral & 2-d (128k parameters)	97.6 (± 0.8)	73.3 (± 4.4)	86.5 (± 4.5)	**86.9 (± 3.6)**	**82.3 (± 3.2)**	45.8 (± 2.9)	66.9 (± 5.8)	51.2 (± 10.6)	71.7 (± 2.0)
1-d & 2-d (111k parameters)	96.5 (± 0.9)	72.4 (± 5.9)	86.3 (± 5.2)	91.0 (± 5.5)	73.3 (± 6.4)	49.5 (± 6.9)	67.7 (± 2.5)	55.0 (± 11.5)	73.8 (± 2.3)
2-d & 1-d & spiral (147k parameters)	97.8 (± 0.6)	70.9 (± 6.1)	88.0 (± 3.7)	85.9 (± 3.8)	75.0 (± 4.3)	48.3 (± 6.6)	67.3 (± 4.4)	**59.0 (± 7.3)**	**74.0 (± 0.6)**
2-d, 48 kernels (158k parameters)	96.5 (± 1.4)	69.3 (± 7.2)	84.5 (± 2.5)	84.2 (± 5.7)	77.4 (± 6.0)	45.5 (± 7.3)	**68.8** (± 1.8)	52.6 (± 10.1)	71.7 (± 2.0)

Table 2: Test set accuracies for all presented architectures. All convolutional layers have 32 kernels unless stated otherwise.

Panwar, S., Rad, P., Choo, K.-K. R., & Roopaei, M. (2018). *Are you emotional or depressed? Learning about your emotional state from your music using machine learning. The Journal of Supercomputing.* doi:10.1007/s11227-018-2499-y

CAL500

Table 3. (a) presents the results of Instrument, Instrument-Solo and Vocal tags, and (b) shows the result of Emotion tags. We use P, R, and F to denote per-tag precision, recall, and F-score, respectively.

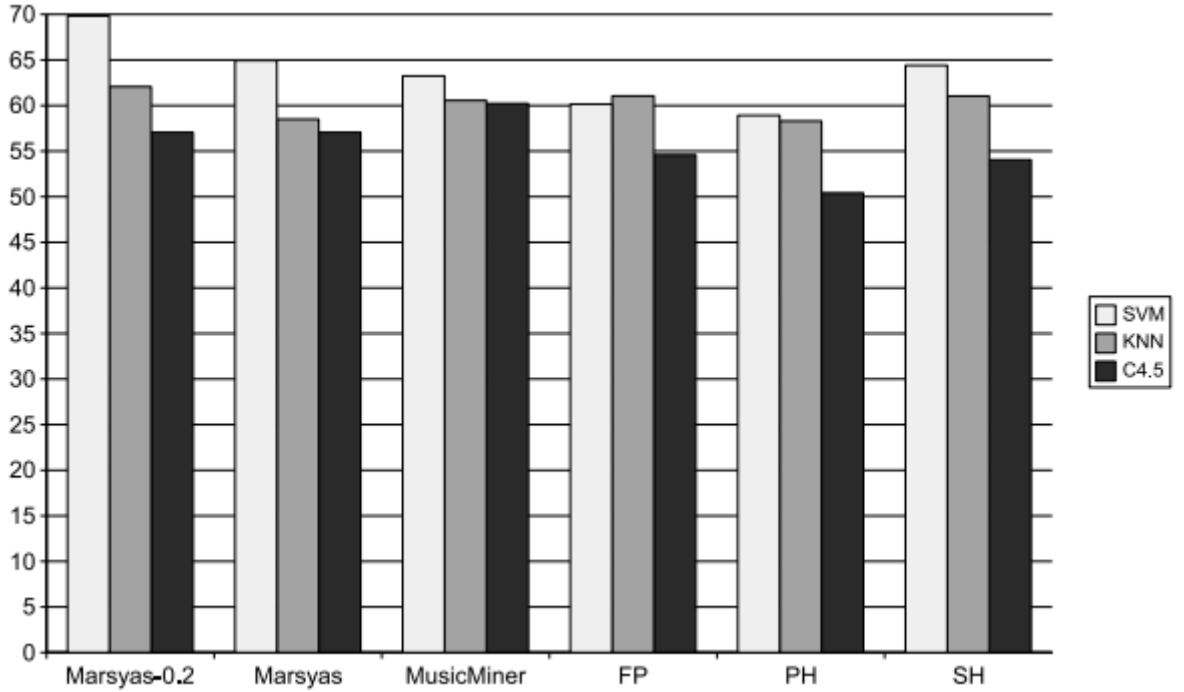
(a) instrument & vocal	P	R	F	AUC
CAL500 (CV)	0.157	0.371	0.213	0.833
CAL500exp (CV)	0.257	0.456	0.317	0.884
CAL500→CAL500exp	0.226	0.384	0.267	0.842
CAL500exp→CAL500	0.172	0.445	0.225	0.844
(b) emotion	P	R	F	AUC
CAL500 (CV)	0.301	0.701	0.417	0.735
CAL500exp (CV)	0.455	0.759	0.561	0.842
CAL500→CAL500exp	0.443	0.751	0.543	0.802
CAL500exp→CAL500	0.303	0.747	0.422	0.744

Wang, S.-Y., Wang, J.-C., Yang, Y.-H., & Wang, H.-M. (2014). Towards time-varying music auto-tagging based on CAL500 expansion. 2014 IEEE International Conference on Multimedia and Expo (ICME). doi:10.1109/icme.2014.6890290

ISMIR04

Collection	1NN	highest kNN (obtained at k)	Literature
Ballroom	88.4%	**89.2%** (k=5)	86.9% [7]
ISMIR'04 train	87.6%	**87.6%** (k=1)	84.0% [16]
ISMIR'04 1458	90.4%	**90.4%** (k=1)	83.5% [6]
HOMBURG	50.8%	**57.0%** (k=10)	55% [12]

Table 2. Accuracies obtained by the "unified" algorithm on the various collections.



Moerchen, F., Mierswa, I., & Ultsch, A. (2006). *Understandable models Of music collections based on exhaustive feature generation with temporal statistics. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '06.* doi:10.1145/1150402.1150523

Good-sounds

Table 2: Results of the **pre-training** in terms of classification accuracy. *Classes* is the number of different class labels. *Hours* describes the amount of recorded material in the subset that we used. The *Train* and *Test* columns contain the accuracy on the train and test sets.

Dataset	Task (Recognition)	Classes	Hours	Accuracy	
				Train	Test
Librispeech	Speech	1000	100	97.6	91.8
IEMOCAP (1 sec)	Speech Emotion	4	7.3	85.6	51.9
IEMOCAP (4 sec)	Speaker	5	7.3	99.8	96.5
Good-Sounds	Instrument	12	14	100.0	100.0
Nsynth	Instrument	11	66	98.1	69.9
MS-SNSD	Noise Type	11	20	100.0	99.8

Table 5: Accuracy results for **sound goodness estimation (SGE)**. For the target task we use the Good-Sounds dataset. We compare no transfer learning (None), weight initialization (WI) and anti-transfer (AT). In particular, we compare anti-transfer with pre-training on the same dataset (Good-Sounds) and on a bigger dataset (Nsynth). We test 2 different train/validation/test splits: random (Rand) and instrument-wise (Instr). The best results per column are highlighted in bold font.

Transfer	Pre-training		Train accuracy Split Type		Test accuracy Split Type	
Type	Task	Dataset	Rand	Instr	Rand	Instr
None	n/a	n/a	91.8	42.2	83.8	22.8
WI	Instrument	Nsynth	93.4	40.5	84.7	29.6
WI	Instrument	Good-Sounds	93.3	**42.3**	84.9	23.9
AT	Instrument	Nsynth	**96.8**	41.0	**86.3**	30.0
AT	Instrument	Good-Sounds	93.9	36.4	85.7	**34.3**

Eric Guizzo, Tillman Weyde, & Giacomo Tarroni . (2020). *Anti-Transfer Learning for Task Invariance in Convolutional Neural Networks for Speech Processing. Neural Networks Journal*

MICM

- 10generated music samples (five violin and five straw samples) are given to ten musicians randomly to rate the quality of the generated samples and the overall result was 76.5%.

TABLE IV. MAIN OVERALL RESULTS

Name of Dastgah	Precision	Recall	F1
Shour	97.42	87.52	92.21
Homayoun	76.22	82.72	79.34
Mahour	91.53	89.92	90.72
Segah	83.58	84.95	84.26
Chahargah	63.04	91.53	74.66
Rastpanjgah	90.66	90.30	90.48
Nava	96.12	87.92	91.84

Azar, S., Malekzadeh, S., Ahmadi, A. & Samami, M. (20198). *Instrument-Independent Dastgah Recognition of Iranian Classical Music Using AzarNet. 27th Iranian Conference on Electrical Engineering (ICEE 2019)*

PCMIR

TABLE II. TRAIN AND TEST PRECISION CHANGES VS THE NUMBER OF NEURONS FOR ALL INSTRUMENTS (PROPOSED METHOD VS [35]) ON PROPOSED DATABASE

Hidden layer neurons	Training classification accuracy		Testing classification accuracy	
-	*Proposed Method*	*[35]*	*Proposed Method*	*[35]*
10	75.6 %	74.6 %	63.3 %	65.8 %
20	79.3 %	80.1 %	66.7 %	66.4 %
30	81.0 %	82.2 %	71.9 %	71.1 %
40	85.7 %	83.2 %	78.2 %	79.9 %
100	94.2 %	94.0 %	**82.5** %	81.4 %

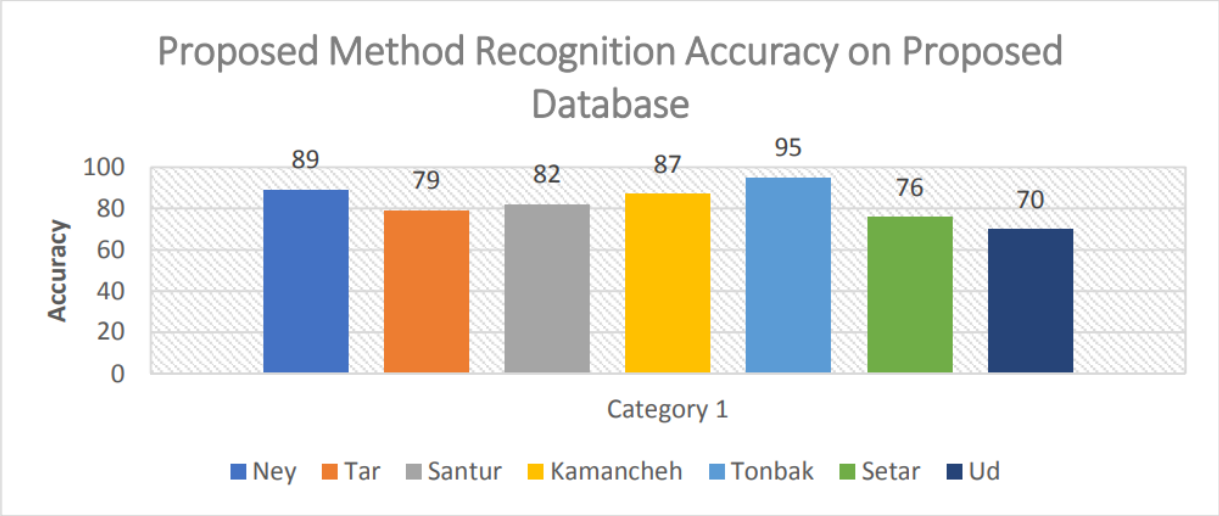


Fig. 12. Proposed method's recognition accuracy on proposed database

Mousavi, S. M. H., Prasath, V. B. S., & Mousavi, S. M. H. (2019). *Persian Classical Music Instrument Recognition (PCMIR) Using a Novel Persian Music Database. 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE).* doi:10.1109/iccke48569.2019.8965166

giantsteps

Dataset	E_0	E_1	E_2	$E_{1/2}$	E_3	$E_{1/3}$	E_4	$E_{1/4}$	$E_{3/2}$	$E_{2/3}$	$E_{5/4}$	$E_{4/5}$	$E_{4/3}$	$E_{3/4}$
ACM˙MIRUM	0.7	73.5	16.0	7.0	0.8	0.0	0.1	0.0	1.8	0.1	0.0	0.0	0.0	0.1
Ballroom	0.4	64.3	1.0	29.7	0.0	1.1	0.0	0.7	0.1	2.3	0.0	0.0	0.0	0.3
Hainsworth	8.6	64.4	1.4	19.4	0.0	0.5	0.0	0.0	1.8	1.8	0.9	0.5	0.5	0.5
GTZAN	4.0	72.2	15.7	5.1	0.4	0.1	0.0	0.0	1.0	0.4	0.1	0.4	0.5	0.1
ISMIR04 Songs	4.3	64.7	19.4	6.3	1.1	0.2	0.0	0.0	2.4	0.4	0.4	0.2	0.2	0.4
SMC	29.0	37.8	6.0	10.6	0.0	2.3	0.0	0.0	5.1	2.3	0.5	2.3	0.9	3.2
Combined	3.9	68.1	12.3	11.2	0.5	0.4	0.0	0.1	1.5	0.8	0.1	0.3	0.2	0.4
GiantSteps	7.1	63.1	4.1	21.5	0.0	0.0	0.0	0.0	0.6	0.3	0.8	0.9	0.5	1.2

Table 1. Error class distribution for tempo estimates T (given in BPM) for different datasets in percent.

Dataset	base	base+new	stem	stem+new	böck	böck+new	schr	schr+new
ACM MIRUM	73.6	**81.8+**	74.3	79.9+	74.5	76.1+	76.3	81.3+
ISMIR04 Songs	65.5	69.2	60.1	62.3	55.4	58.4+	**74.1**	70.7-
Ballroom	64.3	85.1+	64.0	81.8+	84.8	**90.4+**	67.0	85.0+
Hainsworth	64.9	73.4+	69.8	74.8+	84.2	**85.6**	73.0	75.7
GTZAN	73.5	**78.8+**	77.9	76.9	70.7	71.1	78.0	76.2
SMC	45.2	39.6	29.5	29.5	51.1	**51.6**	41.5	35.5-
Dataset Average	64.5	71.3	62.6	67.5	70.1	**72.2**	68.3	70.7
Combined	69.0	**77.4+**	69.1	74.4+	72.4	74.5+	72.8	76.6+
GiantSteps	64.5	64.0	47.0	65.0+	61.5	**70.9+**	58.0	60.1
GiantSteps+Combined	68.4	**75.5+**	66.0	73.1+	70.8	74.0+	70.7	74.3+

Table 4. Tempo results for *Accuracy1* in percent. The '+' and '−' signs indicate a statistically significant difference between an algorithm and the same algorithm enhanced with `new`. Bold numbers mark the best-performing algorithm(s) for a dataset. *Dataset Average* is the mean of the algorithms' results for each dataset except GiantSteps.

Hendrik Schreiber & Meinard Muller. (2017). *A POST-PROCESSING PROCEDURE FOR IMPROVING MUSIC TEMPO ESTIMATES USING SUPERVISED LEARNING. ISMIR 2017*

MagnaTagATune

TABLE III
MODELS PERFORMANCE FOR MAGNATAGATUNE DATASET.

Model	AUC/#param	Model	AUC/#param
Small-rectangular	0.865 / 75k	Choi *et al.* [5]	**0.894** / 22M[9]
Dieleman *et al.* [7]	0.881 / 75k	Proposed x2	**0.893** / 191k
Proposed	0.889 / 75k	Proposed x4	0.887 / 565k

	AUC
FCN-3, mel-spectrogram	.852
FCN-4, mel-spectrogram	**.894**
FCN-5, mel-spectrogram	.890
FCN-4, STFT	.846
FCN-4, MFCC	.862

Table 3. The results of the proposed architectures and input types on the MagnaTagATune Dataset

Methods	AUC
The proposed system, FCN-4	.894
2015, Bag of features and RBM [18]	.888
2014, 1D convolutions [6]	.882
2014, Transferred learning [28]	.88
2012, Multi-scale approach [5]	.898
2011, Pooling MFCC [8]	.861

Table 4. The comparison of results from the proposed and the previous systems on the MagnaTagATune Dataset

Pons, J., Slizovskaia, O., Gong, R., Gomez, E., & Serra, X. (2017). *Timbre analysis of music audio signals with convolutional neural networks. 2017 25th European Signal Processing Conference (EUSIPCO).* doi:10.23919/eusipco.2017.8081710
Keunwoo Choi, Gyorgy Fazekas, Mark Sandler. (2016) *AUTOMATIC TAGGING USING DEEP CONVOLUTIONAL NEURAL NETWORKS. ISMIR 2016*

GMD

- Since the metronome provides a standard measurement of where the musical beats and subdivisions lie in time, we can deterministically quantize all notes to the nearest musical division, yielding a musical score. Recording to a metronome also allows us to take advantage of the prior structure of music by modeling relative note times (quarter note, eighth note, etc.) so as to free models from the burden of learning the concept of tempo from scratch. The main drawback of the metronome is that we enforce a consistent tempo within each individual performance (though not across performances) so we do not capture the way in which drummers might naturally change tempo as they play.

Gillick, J., Roberts, A., Engel, J., Eck, D. & Bamman, D. (2019). Learning to Groove with Inverse Sequence Transformations. *Proceedings of the 36th International Conference on Machine Learning*