

Oregon State University

Interpreting Emotion from Sound: A Comparison of Features and Models for Music Emotion Classification

Shaun Leib

May 16, 2025

soundbendor lab

Introduction

Automatic Emotion Classification

- Apply computational methods to classify emotional content
- Removes costs of time and money from manual annotation
- Disagreement between annotators
- Music recommendation systems, mood based playlist generation, affective music therapy, generative music composition
- MER stagnation

Task

- Find feature space with most emotionally relevant subset of features
- Features are major dimensionality reduction from raw audio
  - Example: pitch, bpm, spectral energy
- Obtained using extractor toolkits which analyze the audio signal
- Evaluate features using various ML models to classify emotion

Russel's Valence and Arousal Graph

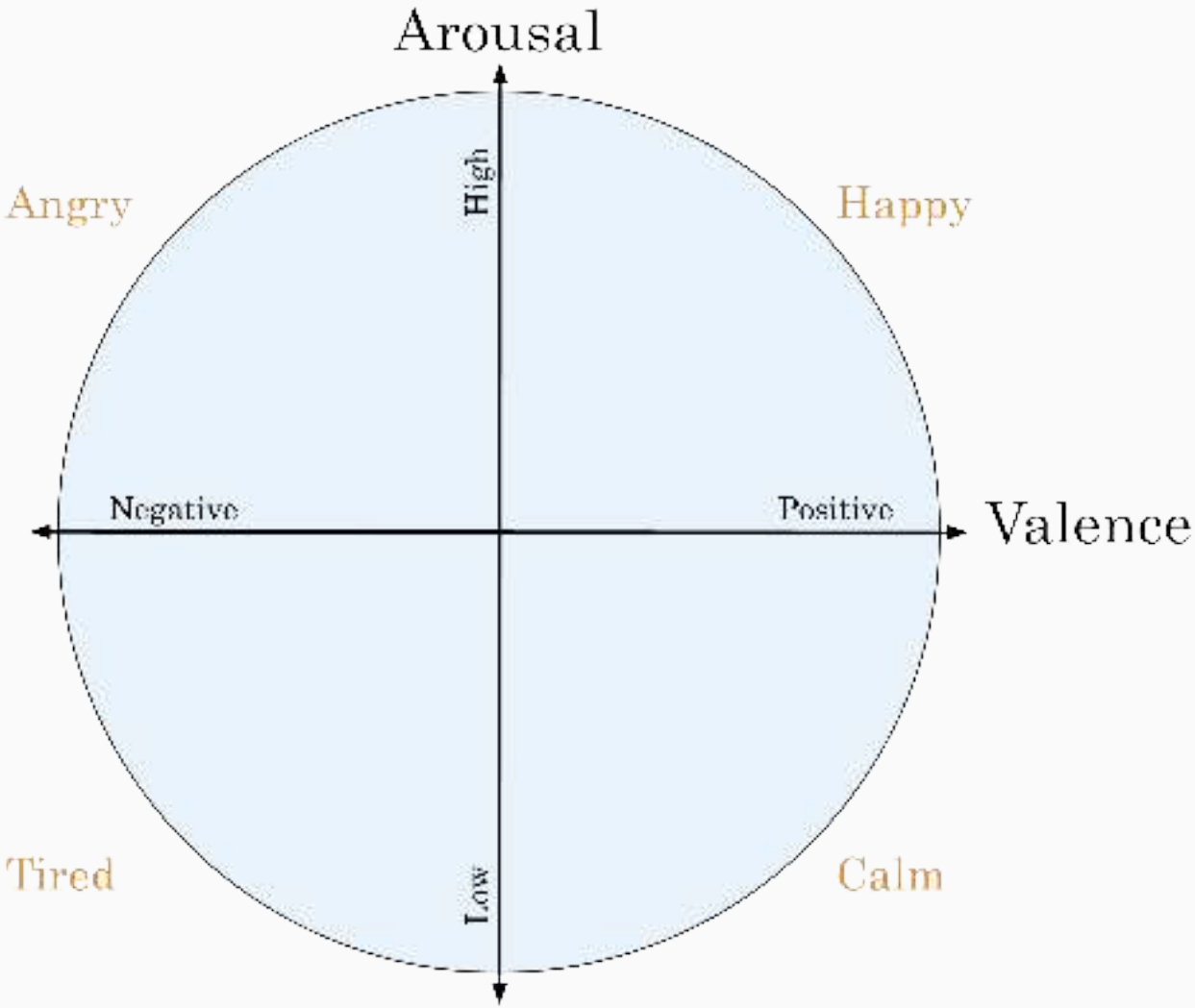


Figure 1: Russel's Valence and Arousal graph.

Related Works

MediaEval MultiMedia Evaluation Campaign

- 2013 - 2015
- DEAM [1]
  - Valence and Arousal regression
  - Highly imbalanced and high disagreement between annotators
- 260 low level features
- LSTM models consistently outperformed other models [2]
- CRNN models also demonstrated the ability to learn relevant temporal and frequency features from spectrograms [3]

MIREX

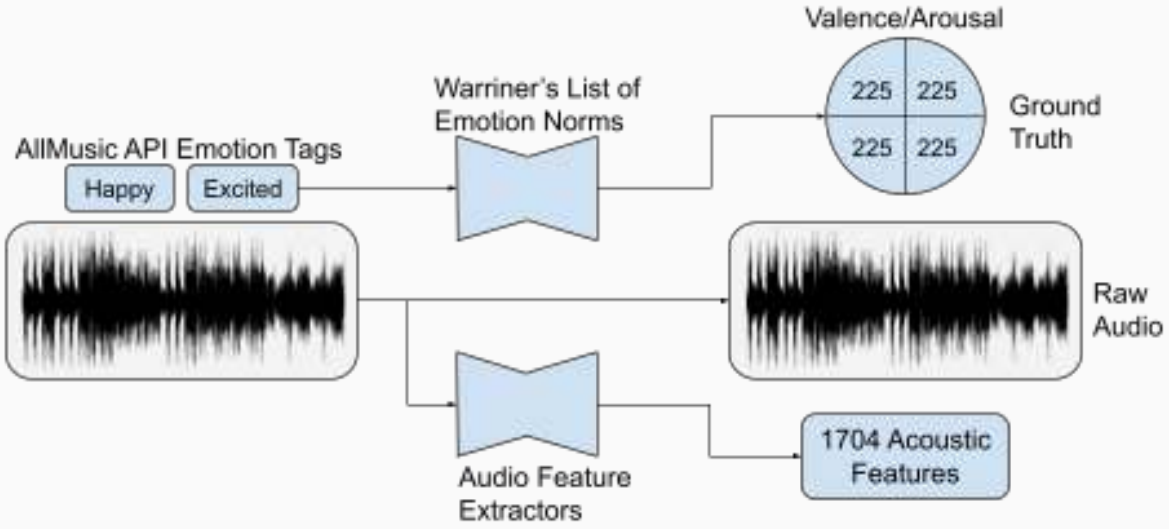
- 2012 - 2020
- Private dataset with five classes
- Panda et al. achieve 67.8% accuracy in 2012 [5]
- Models still struggled to surpass 68% accuracy in 2020
- Stagnation had become an issue [6]

Overview

Methodology

1. Replication study
2. Raw audio input
3. Feature extraction
4. Feature selection methods
5. Experimental results
6. Comparison of feature sets

Dataset



4Q

- 4Q dataset developed by Panda et al.
- Designed to improve upon the datasets in MediaEval and MIREX
- 900, 30 second audio clips obtained using AllMusic API
- Labeled in one of the four quadrants of valence and arousal graph
- 225 songs per quadrant
- 1704 extracted features, or top 100 subset

Methodology

Experimental Design

- Ten-fold cross validation across 20 independent runs
- Macro F_1-Score to evaluate the performance of our models

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Panda et al. Results

%	Q1	Q2	Q3	Q4	F_1-Score
Q1	81.87	6.34	3.79	8.00	78.0
Q2	10.64	84.69	3.11	1.56	86.6
Q3	6.31	3.73	69.89	20.07	70.9
Q4	10.82	0.73	20.38	68.07	68.8

Table 1: This table, recreated from [4], shows the confusion matrix from their results, and the F_1-Score by quadrant. The matrix has been row normalized to better compare against our results. They achieved a combined F_1-Score of 76.4%.

Replication Study

Panda et al. [4] Recreation

- Download their top 100 features
- 71 standard and 29 MIDI
- Evaluate on ensemble of SVM classifiers using bagging
- Radial Basis Function kernel
- Panda's et al. F_1-Score: 76.4%
- Our F_1-Score: 76.8%

Initial Inputs

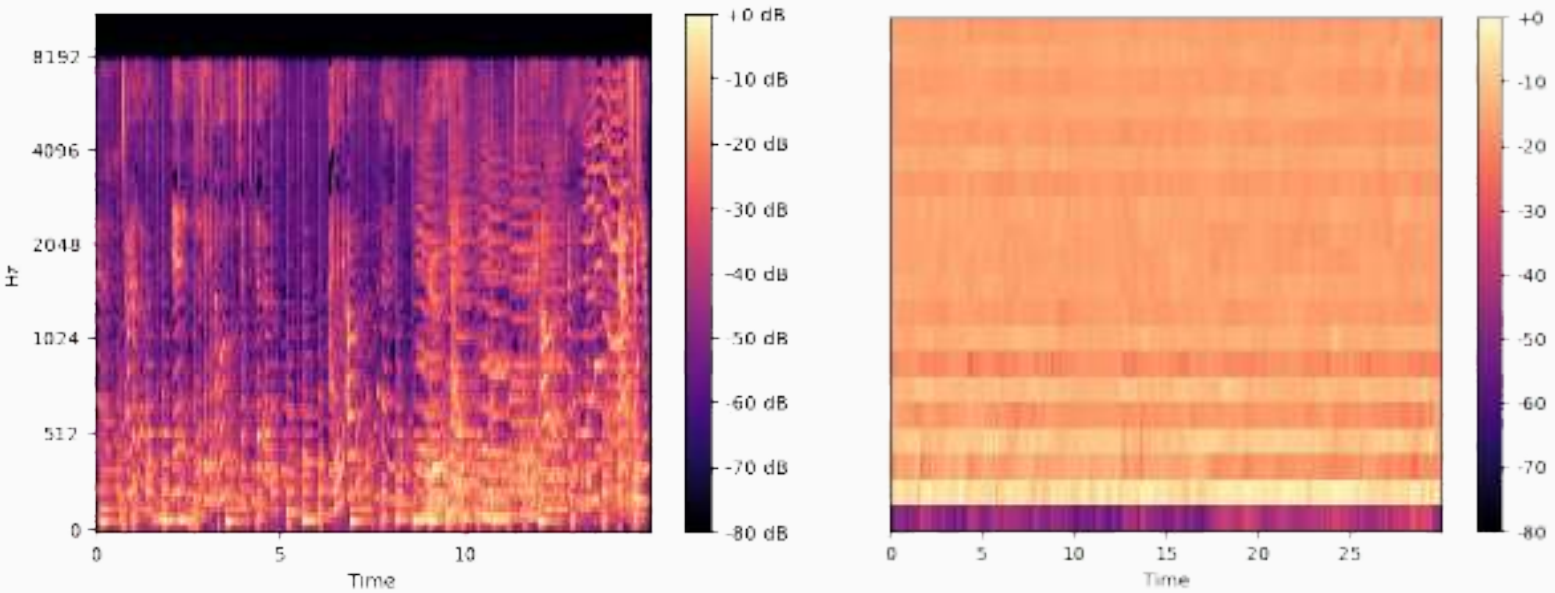


Figure 2: An example of a Mel spectrogram (left) and MFCC (right) from the dataset.

Raw Audio Input

Input Type	Model	F_1-Score	Trainable Parameters
Waveform	BLSTM	33.2	4,266,628
Mel spectrogram	CNN	54.6	48,768,708
Mel spectrogram	CRNN	37.7	86,634,436
MFCCs	CNN	48.4	48,768,708
MFCCs	CRNN	43.8	86,634,436

Table 2: F_1-scores for waveform, spectrogram, and MFCC representations of audio.

Feature Extraction

Dimension	Count
Melody	14
Harmony	100
Rhythm	45
Dynamics	64
Tone Color (Timbre)	448
Expressivity	27
Texture	8
Form	2
High level & Voice	2
Total	710

Toolkit	Count
Essentia	220
Yaafe	95
MIR ToolBox	100
PsySound3	77
Marsyas	218
Total	710

Table 3: Distribution of extracted features by musical dimension (Left) and toolkit (Right). The total feature size is represented by a (710, 467) matrix.

Feature Extraction

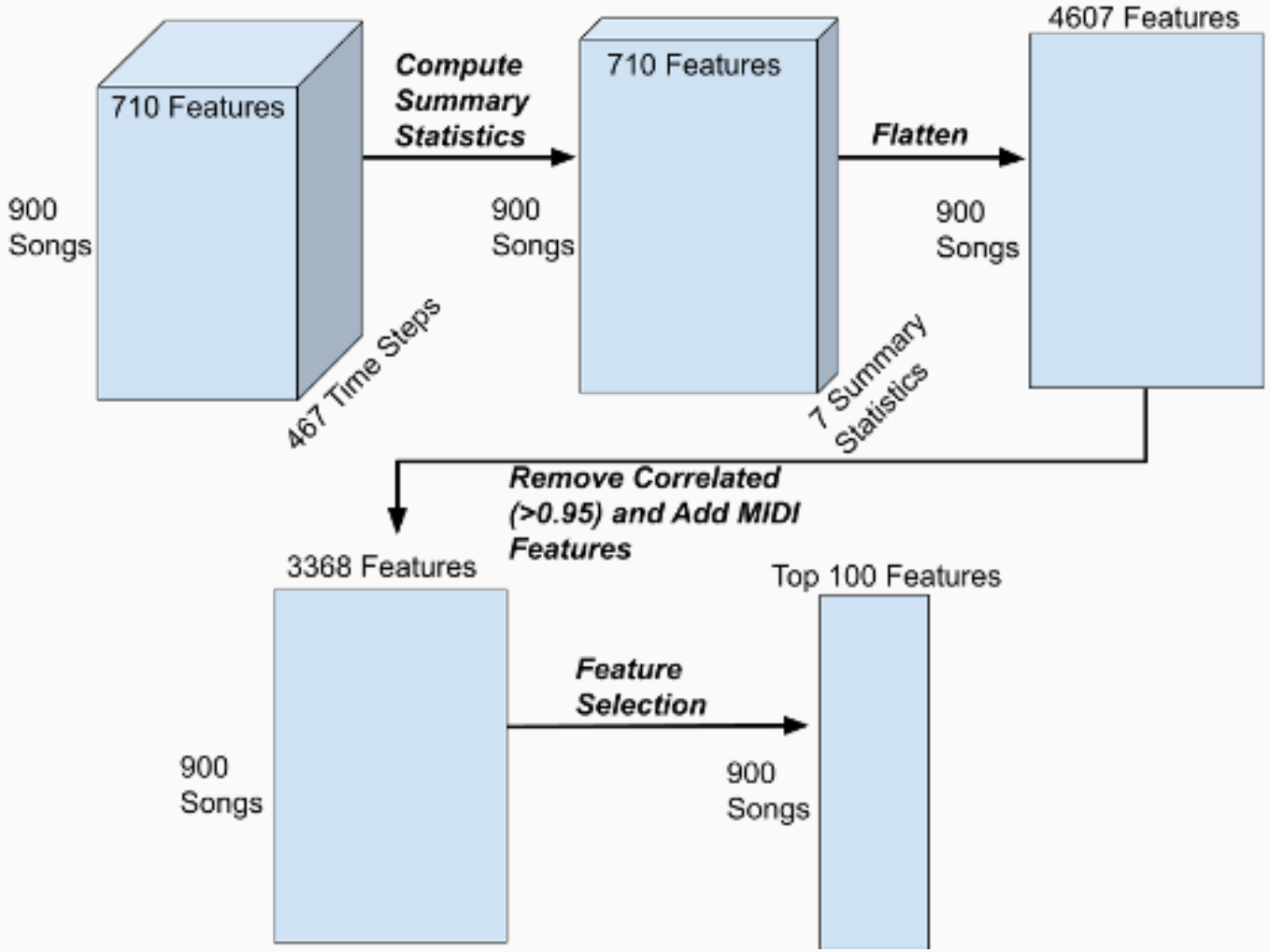


Figure 3: Flow chart representing how we preprocess our features.

Feature Selection Methods

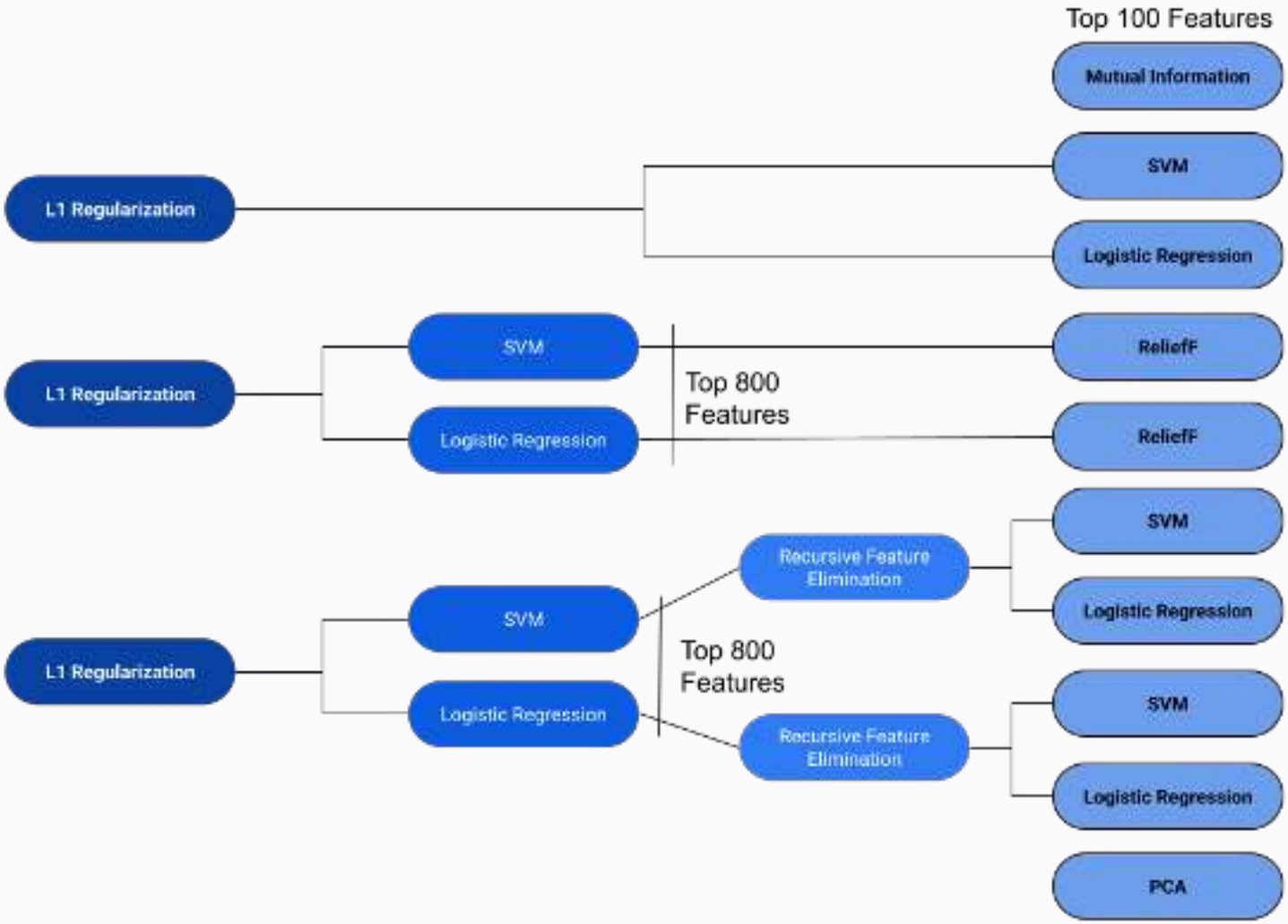


Figure 4: The 10 feature selection/reduction methods we test.

Feature Selection

Feature Selection Algorithm	F_1-Score
Mutual Information	62.04
L1 Regularization (SVM)	75.5
L1 Regularization (logistic regression)	75.6
L1 Regularization (SVM) + ReliefF	70.74
L1 Regularization (logistic regression) + ReliefF	67.77
L1 Regularization (SVM) + RFE (SVM)	**78.57**
L1 Regularization (SVM) + RFE (logistic regression)	78.54
L1 Regularization (logistic regression) + RFE (SVM)	77.77
L1 Regularization (logistic regression) + RFE (logistic regression)	77.92
PCA (100)	68.58
Panda et al. [4]	76.4
Replication Study	76.8

Table 4: F_1-Scores obtained for our SVM model on different feature sets all with 100 features.

ML Models

Model	Panda's et al. 100	Our 100
SVM	76.8	78.57
Random Forest	71.7	73.1
Multi-Layer Perceptron (MLP)	74.8	**79.89**

Table 5: A comparison of F_1-scores between the feature set Panda et al. extracted versus the feature set we extracted, across multiple different machine learning models. MLP contains 275,460 trainable parameters across 4 hidden layers.

Feature Set Size

Feature Size	SVM	MLP
70	78.04	78.20
100	78.57	79.89
150	79.78	80.58
200	79.51	81.51
300	77.64	**82.03**
500	76.45	81.20
Panda et al. 100	76.4	74.8

Table 6: F_1-Scores comparing different feature set sizes evaluated on SVM and MLP. All sets were obtained using the best performing features selection method. MLP contains 275,460 trainable parameters across 4 hidden layers.

Confusion Matrix

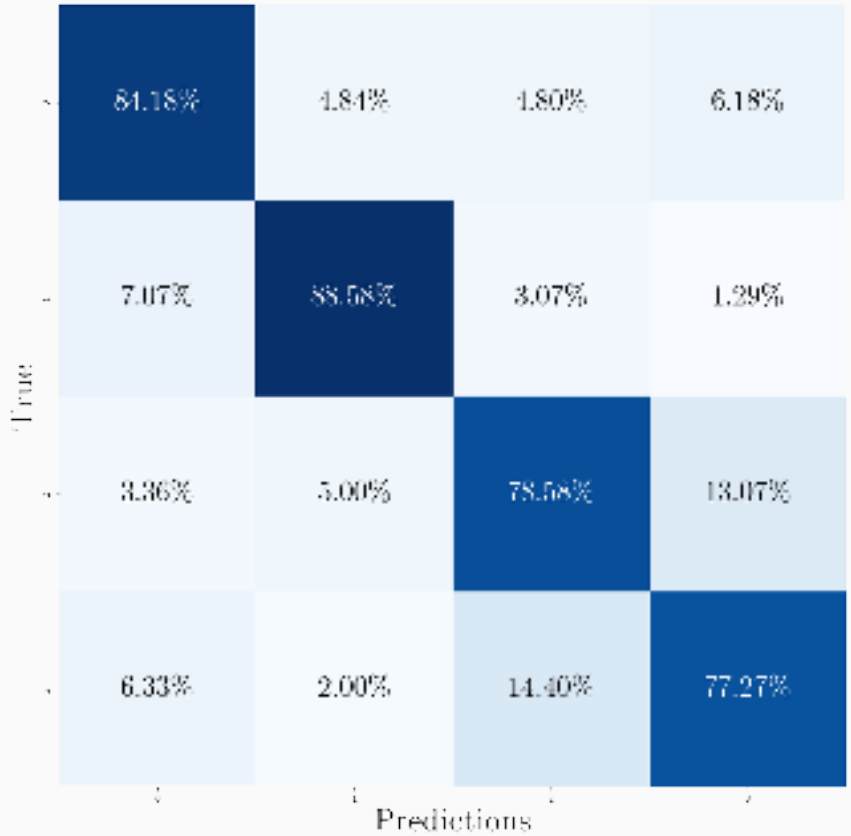


Figure 5: The confusion matrix for the predictions of the MLP model trained on 300 features.

Feature Presence

Toolbox	Novel Feature Types
Essentia	24
MIR ToolBox	11
PsySound3	2
Yaafe	2
Marsyas	3
MIDI	-18

Table 7: Number of feature types by toolbox that were present in our top 100 but not in Panda's et al. feature set.

Toolbox Distribution

Toolbox	Our 300	Our 100	Panda et al. 100
Essentia	111	33	0
MIR ToolBox	48	17	17
PsySound3	16	6	16
Yaafe	30	12	0
Marsyas	79	21	38
MIDI	16	11	29
Total	300	100	100

Table 8: Toolbox distribution for our 100 and 300 feature subsets and Panda's et al. feature set.

Musical Dimension Distribution

Musical Dimension	Our 300	Our 100	Panda et al. 100
Melody	16	4	1
Harmony	49	16	6
Rhythm	20	4	4
Dynamics	12	4	4
Tone Color (Timbre)	182	61	56
Expressivity	13	8	20
Texture	4	2	8
Form	0	0	0
Voice	1	1	1
High-Level	3	0	0
Total	300	100	100

Table 9: The distribution of features within the musical dimensions for our datasets and Panda's et al.

Limitations

Dataset

- 900 songs
- Emotional tags mapped to quadrants
- Easy task: four classes, evenly balanced
- Our features are overfit to this dataset

MIDI Features

- Not able to extract MIDI features
- Limits potential to run the same experiment on other datasets
- Feature extraction limits real time applications

Future Work

Datasets

- DEAM and PmEmo
  - Dynamic regression in VA space
- MagnaTagaTune
  - Multi-tagging labels

Models

- Dynamic features for temporal information
- Autoencoders for dynamic dimensionality reduction
- Attention to identify which time frames are most important

Conclusion

- Outperforms the state-of-art on 4Q dataset achieving the highest F_1-Score to date on this dataset
- Best results from MLP model trained on 300 acoustical features
  - 82.03% F_1-Score
- 100 dimensions is not a limit for features
- Stronger learning models such as MLP are able to capture emotion in higher dimensional spaces
- Advancements in music recommendation systems, mood based playlist generation, affective music therapy, and generative music composition

References i

[1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. **"Developing a benchmark for emotional analysis of music".** In: *Plos One* (2017). DOI: 10.1371/journal.pone.0173392.

[2] Xinxing Li et al. **"A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction".** In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 544–548. DOI: 10.1109/ICASSP.2016.7471734.

[3] Miroslav Malík et al. **"Stacked Convolutional and Recurrent Neural Networks for Music Emotion Recognition".** In: *Sound and Music Computing Conference (SMC)*. Vol. 14. Espoo, Finland, July 2017. DOI: 10.48550/arXiv.1706.02292.

[4] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. **"Novel Audio Features for Music Emotion Recognition".** In: *IEEE Transactions on Affective Computing* 11.4 (2020), pp. 614–626. DOI: 10.1109/TAFFC.2018.2820691.

[5] Renato Panda and Rui Pedro Paiva. **"MIREX 2012: Mood Classification Tasks Submission".** In: *Music Information Retrieval Exchange (MIREX)*. Porto, Portugal, Oct. 2012.

[6] Ricardo Scholz, Geber Ramalho, and Giordano Cabral. **"Cross Task Study on MIREX Recent Results: an Index for Evolution Measurement and Some Stagnation Hypotheses".** In: *17th International Society for Music Information Retrieval Conference (ISMIR)*. Aug. 2016. DOI: 10.5281/zenodo.1416897.