International Journal on Artificial Intelligence Tools
Vol. 27, No. 5 (2018) 1850016 (18 pages)

DOI: 10.1142/S0218213018500161

World Scientific
www.worldscientific.com

Classification of Music Mood Using MPEG-7 Audio Features and SVM with Confidence Interval

Riyanarto Sarno*, Johanes Andre Ridoean† and Dwi Sunaryono‡

Informatics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
**riyanarto@if.its.ac.id*
†johanes.andre13@mhs.if.its.ac.id
‡dwi@if.its.ac.id

Dedy Rahman Wijaya§

School of Applied Science, Telkom University, Bandung, Indonesia
Informatics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
dedyrw@tass.telkomuniversity.ac.id

Received 7 July 2017
Accepted 25 March 2018
Published 14 August 2018

Psychologically, music can affect human mood and influence human behavior. In this paper, a novel method for music mood classification is introduced. In the experiment, music mood classification was performed using feature extraction based on MPEG-7 features from the ISO/IEC 15938 standard for describing multimedia content. The result of this feature extraction are 17 low-level descriptors. Here, we used the Audio Power, Audio Harmonicity, and Audio Spectrum Projection features. Moreover, the discrete wavelet transform (DWT) was utilized for audio signal reconstruction. The reconstructed audio signals were classified by the new method, which uses a support vector machine with a confidence interval (SVM-CI). According to the experimental results, the success rate of the proposed method was satisfactory and SVM-CI outperformed the ordinary SVM.

Keywords: Music mood classification; MPEG-7; support vector machine; confidence interval.

1. Introduction

Music is an art form that allows people to express their creativity through sound. Music information retrieval (MIR) refers to multidisciplinary research efforts that seek to develop new content-based searching schemes, interfaces and network delivery mechanisms to make the large store of music in the world more accessible.[1] MIR methods have developed rapidly, for example, genre detection,[2] sound recognition,[3] and cover song detection,[4] etc. In this research, we focused on music mood classification, utilizing MPEG-7 for feature extraction. The advantage of using MPEG-7 for feature

§Corresponding author

extraction is that it stores features in the form of metadata according to the multimedia industry standard ISO/IEC 15938,[5] as well as several other features, which can be used to describe multimedia content for a variety of MIR systems.

Several MIR experiments have been performed using MPEG-7, for example, genre detection,[2] fingerprinting,[6] etc., but no experiments have been conducted yet that take advantage of features extracted from MPEG-7 for mood classification. Several interesting studies related to music mood classification have been carried out. Firstly, Ren and Wu introduced the short-term timbre and long-term modulation features for music mood classification.[7] This feature set performed satisfyingly using the MIREX dataset, with 69.50% accuracy. The drawback of this approach is that it works only for short-duration music clips. Secondly, Hu and Yang evaluated mood regression models using 15 features.[8] This work used a cross-dataset with Chinese and Western pop songs. The experimental results indicated that the loudness and timbre features got the best results for music mood regression in the valence and arousal dimensions. Thirdly, Kermanidis *et al.* compared jAudio and MIRtoolbox as audio feature extractors for music mood classification.[9] According to the experimental results, the feature set from MIRtoolbox got better performance than the jAudio feature set. However, this work did not discuss audio feature selection.

The present study makes two contributions to the field of MIR. Firstly, we propose a new method, Support Vector Machine with Confidence Interval (SVM-CI), to improve the accuracy of music mood classification. Secondly, this research demonstrates a music mood classification method that is based on extraction of industrial standard features from MPEG-7, using a combination of features, i.e. Audio Power, Audio Harmonicity, and Audio Spectrum Projection. The classification is expected not only to improve accuracy but also to be more efficient because the music mood prediction is based on signal characteristics only and does not depend on the content of the lyrics. The proposed method detects four mood labels: angry, happy, sad, and relaxed. These labels were taken from Russell's diagram.[10] A confidence interval is used on the classification result to improve the success rate.

Music can affect human mood and establish a personal attitude. The positive impact of background music on student concentration in a class has been investigated in a case study.[12] The terms "emotion" and "mood" have slightly different meanings.[11] Emotion is a feeling that emerges due to a direct stimulus, while mood is a feeling that emerges without direct stimulus. Table 1 shows a comparison between mood and emotion.

Table 1. Emotion vs. mood.

Emotion	Mood
Very short in duration (seconds/minutes)	Lasts longer than emotions (hours/days)
Specific and numerous in nature	More general
Usually accompanied by distinct facial expression	Generally not indicated by distinct expression

The rest of this paper is organized into the following sections: Section 2 provides a brief overview of related works on mood classification; Section 3 generally explains the MPEG-7 features used in this study; Section 4 explains the material and methods used in the experiment; Section 5 describes the results of the experiment, including improving the accuracy of the results using a confidence interval; and, finally, Section 6 contains the conclusion of this work.

2. Related Works

There are several previous works related to music mood/emotion recognition. A system using music emotion and the human face as features for drama video has been proposed.[13] The method concerned uses two high-level features (music emotion and the human face) and two low-level features (shot duration and motion magnitude) to extract highlights. The authors claim that the method is effective for video highlight extraction. A retrieval system between Chinese folk images and Chinese folk music based on a differential evolutionary-support vector machine (DE-SVM) has been proposed.[14] In another study, a music emotion recognition system based on modified gene expression programming (GEP) was developed.The performance of the system was reported as being better than ordinary GEP.[15] A model for music emotion recognition using feature selection and statistical models has been proposed.[16] The results obtained had a higher average accuracy rate for arousal compared to valence (80% for arousal vs. 63% for valence). The combination of audio, lyrics, and linguistic data to classify Greek songs into several valence and arousal categories has been introduced.[9] Another work attempted to use a cross-dataset (Chinese and English songs). It evaluated a mood regression model of the valence and arousal dimensions. It was reported that the loudness and timbre features had good performance for valence and arousal prediction.[8]

Wavelet transform has been employed in several works for audio processing tasks. Applying the wavelet coefficient with Daubechies, Coiflet, and Symlet families led to a significant improvement of error rate reduction.[17] The wavelet transform in tandem with a support vector machine has been used for voice activity detection under noisy environment; the experimental result showed that the method offers a promising performance.[18] Another research performed comparison of emotional expression based on loudness, tempo, spectral balance and perturbation. It was reported that these parameters have a significant contribution on emotion differentiation.[19]

We also found several works related to music mood classification. The timbre and modulation features have been used for music mood classification.[7] Statistical spectrum descriptors (SSD), MFCC, OSC, and SFM/SCM were employed for feature extraction. An overall accuracy of 50.91% in distinguishing five music mood clusters achieved by support vector machine was reported. Another work used timbre, intensity, and rhythm to classify several moods (happy, sad, exciting, and silent) in Bollywood music.[20] It used jAudio as feature extractor and K-means as classification technique. This work achieved a classification accuracy of 70%. Music mood classification based on signal processing theories has been proposed.[21] Entropy, Energy, Zero Crossing Rate, Spectral Roll-off,

Spectral Flux, Spectral Centroid, Root Mean Square, MFCC were used as features to distinguish three music moods (happy, sad, angry). The overall accuracy was 60% using a neural network classifier. Intensity, rhythm, and timbre have been used as features to detect the mood of Bollywood music.[22] A decision tree was used to differentiate four music moods (exuberance, anxious, serene, depression) with a classification accuracy of 60%. Acquiring mood information from songs in a large music database was performed in another work.[23] Several feature selection algorithms were employed for 5929 music clips, using rhythm, timbre, and intensity as features. A Gaussian mixture model (GMM) and a support vector machine were employed as classifiers. A classification accuracy of 84% was achieved in discriminating seven classes of music moods (sturdy, enthusiastic, lively, melodious, euphemistic, depressed, and anguished). Moreover, music mood trajectory estimation was also conducted. Lyrics and acoustic features have been used together in another work.[24] By using a clustering technique, this work could correctly match 21% of the songs. Lyrics have been used independently for music mood classification to distinguish two types of mood (happy and sad).[25] The accuracy achieved in this work was 80% using Naïve Bayes and SVM. Music mood classification of Hindi songs has also been reported. A decision tree technique in tandem with rhythm, timbre, and intensity as features were used to differentiate between five music clusters. The overall accuracy achieved was 51.56%.

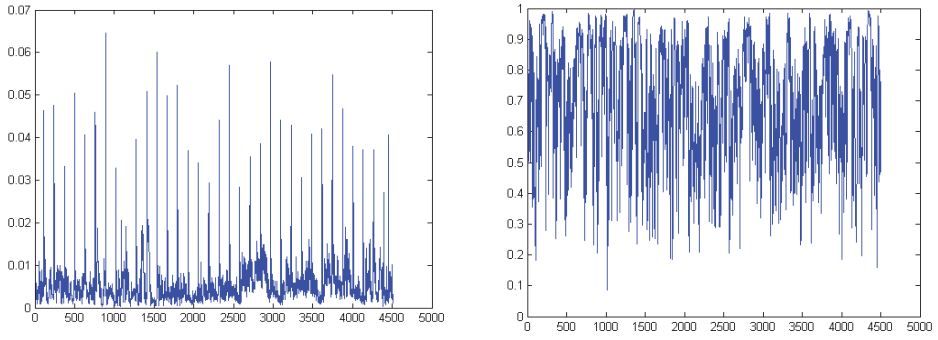
In the present experiment, we attempted to use MPEG-7 feature extraction for multimedia content description while also improving classifier performance by proposing SVM with a confidence interval.

3. MPEG-7 Content Description

MPEG-7 is an international multimedia content standard based on ISO/IEC 15938.[5,26] With MPEG-7 feature extraction, 17 low-level descriptors can be extracted. Each descriptor that is generated describes a characteristic of the music signal so that the content can be extracted, including music mood identification based on the characteristics of the signal. From a previous work, music mood is influenced by the rhythm and harmony of a musical piece.[27] In MPEG-7, these are represented by Audio Power (AP) and Audio Harmonicity (AH) respectively. Therefore, Audio Power and Audio Harmonicity were used in this experiment to classify music mood. Equation (1) provides the calculation to get the audio power of a music recording.

$$AP(l) = \frac{1}{N_{hop}} \sum_{n=0}^{N_{hop}-1} \left| s(n + lN_{hop}) \right|^2 \ (0 \ \leq l \leq L-1)\,, \tag{1}$$

where L is the total number of time frames, $s(n)$ is the average square waveform, l is the index frame and N_{hop} is the number of time samples between two successive frames. Audio Harmonicity is a feature that describes two properties of a spectrum. The first property the harmonic ratio, i.e. the ratio of the total harmonic power, and the second is



(a) (b)

Fig. 1. (a) Plot of Audio Power, (b) plot of Audio Harmonicity (*x*-axis is frequency and y-axis is magnitude).

upper limit harmonicity, i.e. the frequency spectrum that cannot be considered part of harmony. The goal is to distinguish harmonic sounds (e.g. musical instruments) and non-harmonic sounds (noise, unclear speech, etc.). Figure 1 is an example of the Audio Power and Audio Harmonicity plot that was used in this work.

4. Materials and Method

This section explains the dataset and the experimental setup used. Moreover, it also describes the audio signal reconstruction using the discrete wavelet transform and the proposed method, Support Vector Machine with Confidence Interval (SVM-CI), to improve the classification accuracy.

4.1. *Dataset and data acquisition*

The dataset was taken from a database containing 1000 songs.[28,29] It consists of music clips with 45 seconds of duration. In the dataset, not all instances are already labeled; only 310 instances have a mood label. Hence, the experiment was performed using the complete instances. Each music has a valence score and an arousal score. Arousal is a value used to measure the level of activity of a person, while valence is the value used to measure the level of pleasure/comfort. The dataset is a collection of MP3 files. However, MPEG-7 feature extraction requires WAV files, so the files first had to be converted to WAV. Figure 2 shows the steps of this experiment. MPEG-7 is used as feature extraction method for the data in the dataset. Then, the MPEG-7 XML metadata are obtained. Subsequently, the Audio Power and Audio Harmonicity features are obtained using XQuery (XQuery is a language for querying XML data). Both of these features are processed in the signal processing stage. Lastly, the processed features are used in the classification process.

The division of mood labels can be explained based on Russell's diagram: we divide moods using the arousal value as the y-axis and the valence value as the x-axis. From the dataset, every music fragment already has valence and arousal values, so they can be

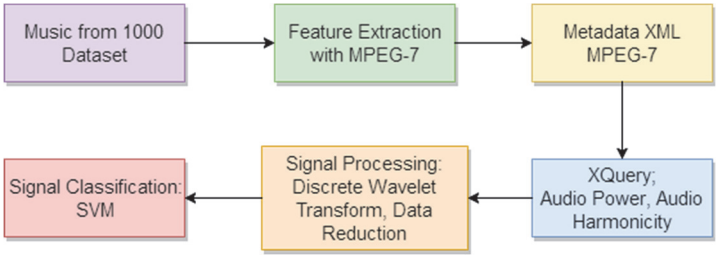


Fig. 2. General step of music mood classification.

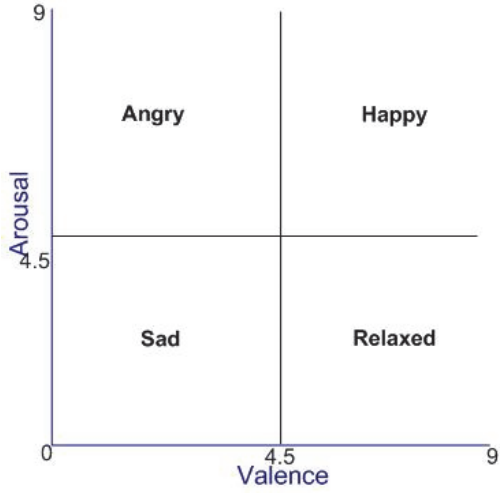


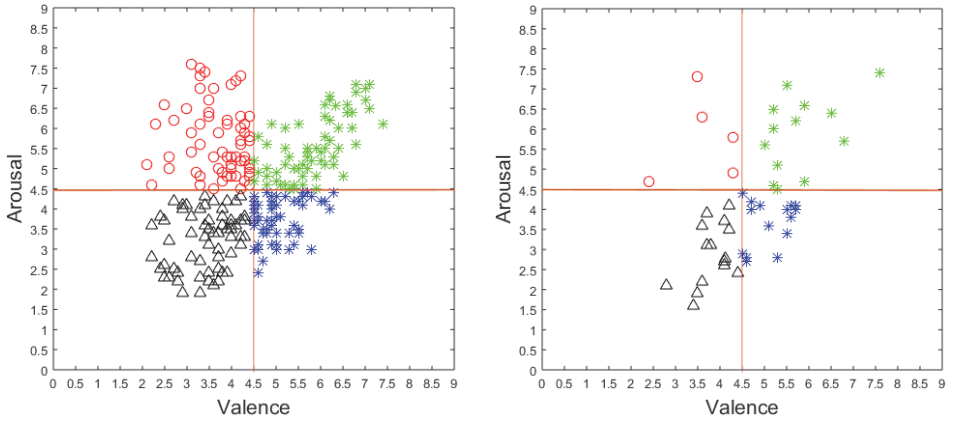
Fig. 3. Russell's diagram.

mapped according to Fig. 3. Both of these values range between 1 and 9. Then, we create two new dividing lines for the upper and lower limits, where the intersection is 4.5. The mood is categorized into four classes because we classify based on 2-dimensional Euclidean distance; these classes are considered sufficient to represent the mood of a song. In addition, Fig. 4 denotes the distribution of the dataset based on 2-dimensional label data. Several instances are on the boundaries between classes, which could lead to mood misclassification.

Once the music is divided according to label, feature extraction is performed by MPEG-7, which produces XML metadata. XQuery is employed to get the Audio Power and Audio Harmonicity features as XML metadata.

4.2. *Audio signal reconstruction*

The main purpose of this stage is to perform preprocessing to get better signal quality. Several works have reported performance improvement with a proper signal quality.[30–35]



(a) (b)

Fig. 4. Plot of the dataset based on 2-dimensional label data. The red, green, blue, and black symbols denote angry, happy, relaxed, and sad instances, respectively. (a) Data training, (b) data testing.

Audio signal reconstruction is essential because of the following reasons: variation in audio quality for example in compressed audio formats; the influence of the sound system; the presence of external noise; etc. In this study, the discrete wavelet transform (DWT) was used for noise filtering. DWT is a popular technique for non-stationary signal processing. Several wavelet families can be utilized for signal reconstruction so the best-suited mother wavelet (MWT) must be determined. In this study, we used the principle that signal reconstruction based on a particular MWT still has to keep the essential information from the original signal. Reconstruction of an audio signal $a(t)$ using DWT can be expressed as in Eq. (2):

$$dwt(j,l) = \langle a(t), \omega_{j,l}(t) \rangle = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} a(t)\omega^* \left(\frac{t - l2^j}{2^j} \right) dt \,, \tag{2}$$

where j is the scaling parameter of the mother wavelet. Furthermore, $\omega * (.)$ is the complex conjugation of the used mother wavelet, l is the shifting parameter on the time axis. Meanwhile, ω describes the MWT used. The main problem when performing DWT is to determine the decomposition level and the best-suited mother wavelet for a particular signal reconstruction. Thus, in the first step, we need to find the frequency characteristic of the audio signal to determine the most appropriate decomposition level. Equation (3) shows how to get the frequency characteristic of the particular signal:

$$[maxvalue, indexmax] = \max\left(abs\left(FFT(S - mean(S)) \right) \right), \tag{3}$$

where S is the feature of Audio Power and Audio Harmonicity, *maxvalue* is the maximum frequency of the signal, and *indexmax* is an index of the maximum frequency of the signal. The goal of getting these two values is to obtain the frequency range. The purpose of performing FFT on the signal is to change the signal from the time domain to the frequency domain so that we can know the useful information contained in the

signal's frequency domain.[30] Equation (3) is used to obtain the maximum frequency and the maximum frequency index, which are used to find the frequency range corresponding to the table with wavelet decomposition levels. The table with wavelet decomposition levels is obtained by the following rule[36]:

$$\frac{f_q}{2^N + 1} \leq f_{char} \leq \frac{f_q}{2^N}, \tag{4}$$

where f_q is the sampling frequency, f_{char} is the dominant/maximum frequency, and N is the level of decomposition. Table 2 shows the level of decomposition based on Eq. (4).

Table 2. Wavelet decomposition level.

Decomposition Level (L)	Frequency Range (Hz)
1	256–512
2	128–256
3	64–128
4	32–64
5	16–32
6	8–16
7	4–8
8	2–4
9	1–2
10	0.5–1
11	0.25–0.5
12	0.125–0.25
13	0.0625–0.125

To find a range of frequency values in the above table, do the following calculation:

$$Fh = indexmax * Fs/L, \tag{5}$$

where Fs is the sampling frequency = 1024 and L is the length of the signal. Thus, the best decomposition level for both features is obtained. After determining the decomposition level, it is necessary to find the best-suited MWT for the audio signal samples. In this work, the information quality ratio (IQR) is used to find the best-suited MWT for signal reconstruction, which has the best capability to keep essential information. IQR can be expressed as in Eq. (6)[37]:

$$IQR\big(a(t), b(t)\big) = \frac{\sum_{a_i \in a(t)} \sum_{b_j \in b(t)} p\big(a_i, b_j\big) log_2(p(a_i)p(b_j))}{\sum_{a \in a(t)} \sum_{b_j \in b(t)} p\big(a_i, b_j\big) log_2(p(a_i, b_j))} - 1, \tag{6}$$

where $a(t)$, $b(t)$, a_i, b_i are original audio signal, reconstructed audio signal, particular value of $a(t)$, particular value of $b(t)$, respectively. $p(a_i)$ and $p(b)$ are the marginal probability and $P(a_i, b_i)$ is the joint probability of a_i and b_j. In this experiment, 38 MWTs were compared based on the IQR value and bior 2.8 as the best-suited mother wavelet for audio signal reconstruction. The approximate coefficient was taken at three levels of decomposition for the further processes. Hence, these two features were

Table 3. Length of features.

Feature	Length
Audio Power	4498
Audio Harmonicity	4493

combined in a single list. Length uniformity is necessary because the signals generated by MPEG-7 have different lengths. Equalization of the signal length is done for the classification process. We have done a long analysis of 310 music fragments; the minimum length is shown in Table 3. Combining these two features in one list, the first part (0-4449) is Audio Power and the rest is Audio Harmonicity.

For example, L_{AP} and L_{AH} are the length of Audio Power and Audio Harmonicity, respectively. After performing DWT, data reduction with length equalization of the signals is performed. Then, equalization of signal length is performed by Eqs. (7) and (8):

$$L'_{AP} = \min\{4.498\,,\ L_{AP}\}\,. \tag{7}$$

$$L'_{AH} = \min\{4.493,\ L_{AH}\}\,. \tag{8}$$

The length of L_{AP} and L_{AH} are never below 4.498 and 4.493 during 45 seconds of the duration of the extracted music. L'_{AP} and L'_{AH} are the new lengths of the readily processed signals. This process only removes a few milliseconds of the audio signal, which will certainly not eliminate the signal's characteristics. We also confirmed that feature comparison will not be messed up when classification is performed because of length equalization. The next process is classification using a support vector machine. SVM was utilized because of its satisfactory performance in music mood classification in previous studies. SVM has been reported as having the best performance for classifying music mood based on timbre and modulation.[7] In a previous study, it has also been used for music mood annotation.[38] Moreover, a differential evolutionary algorithm has been employed to optimize the SVM training parameters. Successful implementation in building up an emotion-driven Chinese folk music retrieval system has been reported.[14] Also, support vector regression with RBF kernel has been successfully implemented for a regression model using a cross-dataset and cross-cultural music mood.[8] The number of training data for each label is 65 instances because the dataset only contains 70 angry mood instances. Table 4 provides a breakdown of the amount of training data. After this phase, the machine learning algorithm can predict new data entries.

Table 4. Number of training data.

Data Training Details		
Module	Label	Number of Data
Mood	Angry	65
	Happy	65
	Relaxed	65
	Sad	65

4.3. *Proposed Support Vector Machine with Confidence Interval* (*SVM-CI*)

In this study, the library from scikit-learn was used. The kernel used was the radial basis function (RBF) with auto kernel coefficient, the penalty parameter C of the error term was 1.0, and *tol score* was 0.001. The other settings were default. The kernel coefficient/gamma defines how much influence a single training example has, while *tol* is the tolerance for the stopping criteria, which tells when to stop searching for a minimum or maximum once a certain tolerance level has been reached. In this study, we did not focus on optimizing the SVM parameters but used a confidence interval for performance improvement. SVM-CI is proposed to improve the classification accuracy for music mood detection. Figure 5 shows the steps of SVM-CI.

According to Fig. 5, standard deviation values are calculated from the valence and arousal values of each instance. This is the basis to build the rules of the confidence interval for music mood. In our proposed method, three rules are needed to improve the flexibility of the classifier. Misclassification often occurs in the border areas, which leads to severe performance degradation of the classifier. This method aims to address such errors by way of remapping the output of the classifier based on vertical, horizontal, and diagonal rules.

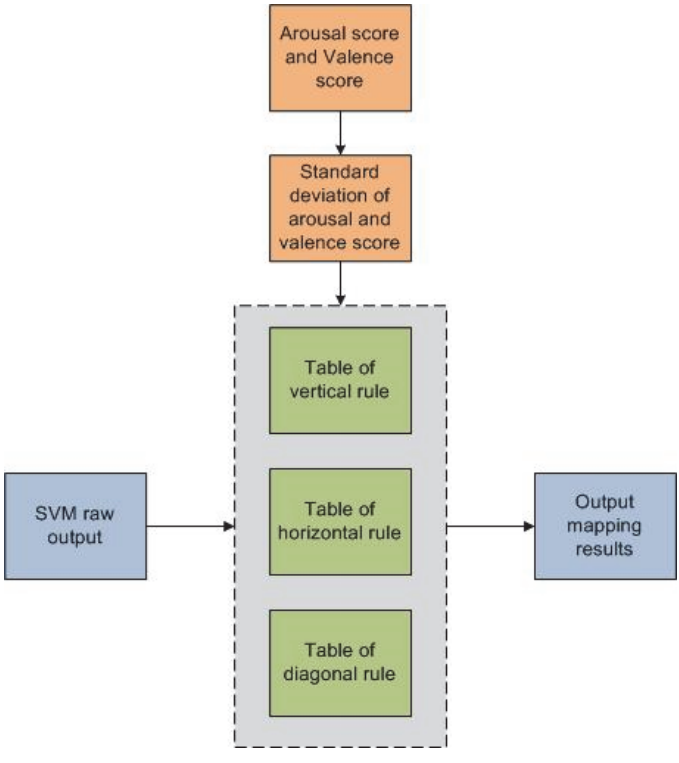


Fig. 5. SVM with Confidence Interval (SVM-CI).

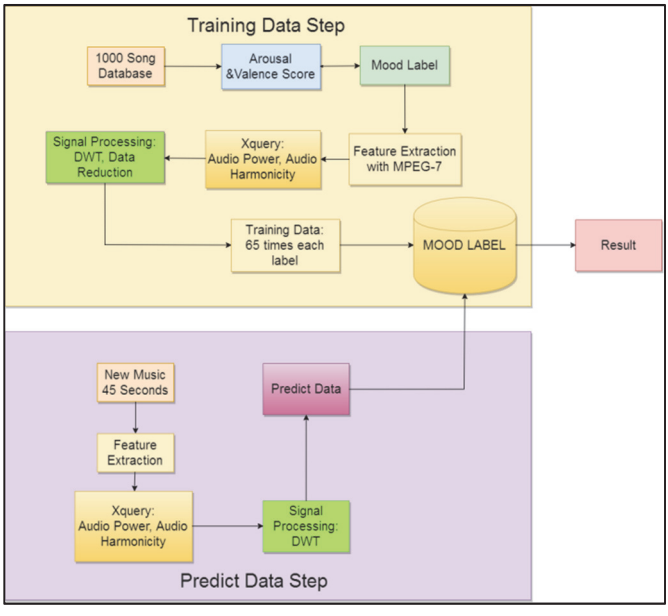


Fig. 6. Testing scenario.

5. Results and Discussion

Figure 6 is a step-by-step visualization of the experiment performed. The music data used in the testing phase of the experiment were taken from the same dataset but with different training data.

The classification accuracy, recall/TPR, and precision of the classification were calculated by Eqs. (9), (10) and (11), respectively.

$$Accuracy = \frac{TP + TN}{TOTAL\ DATA} \times 100\% . \tag{9}$$

$$TPR = \frac{TP}{TOTAL\ CONDITION\ POSITIVE} \times 100\% . \tag{10}$$

$$Precision = \frac{TP}{TOTAL\ PREDICTION\ POSITIVE} \times 100\% . \tag{11}$$

TP and TN are the total of true positives and the total true negatives, respectively. We tested 15 data for each label except angry mood because there were only 70 instances in the dataset and 65 for the training process. So for angry mood only 5 data were tested.

5.1. *Music mood classification based on Audio Power and Audio Harmonicity features*

The overall success rate was 72%. Table 5 displays the experimental results obtained, including the recall and precision values.

Table 5. Results of mood classification using AP and AH features.

Testing / Actual	Angry	Happy	Relaxed	Sad	Recall/ TPR
Angry	5	0	0	0	100%
Happy	0	15	0	0	100%
Relaxed	6	0	1	8	6.67%
Sad	0	0	0	15	100%
Precision	45.45%	100%	100%	65.22%	

The use of these features looks promising for accurately detecting angry, happy, and sad mood. However, it is not satisfactory yet for relaxed mood detection with only 6.67% of recall.

5.2. *Music mood classification based on Audio Spectrum Projection feature*

In the experiment, we also performed a comparison of classification performance based on Audio Spectrum Projection (ASP). Audio Projection is a characteristic of the signal used for the classification of any music type. In previous works,[39,40] it is explained that the ASP feature is part of the MPEG-7 feature set and consists of a normalized audio spectrum envelope (NASE) and basic decomposition algorithm. The steps taken are the same as in the data reduction step. The minimum length of the ASP feature is 20 240. When the test was done, the results obtained were less accurate. The overall success rate was only 38%. Table 6 is the result of the experiment using the ASP feature.

Based on the obtained results, the ASP feature is only useful for detecting happy mood and relaxed mood, but it has a very low true positive rate in detecting angry and sad mood.

Table 6. Results of mood classification using ASP feature.

Testing / Actual	Angry	Happy	Relaxed	Sad	Recall/ TPR
Angry	0	4	1	0	0%
Happy	0	10	5	0	66.67%
Relaxed	0	5	9	1	60%
Sad	0	6	9	0	0%
Precision	0%	40%	37.5%	0%	

5.3. *Music mood classification based on Audio Power, Audio Harmonicity, and Audio Spectrum Projection features*

We also combined three features: AP, AH, and ASP, but the results obtained were less accurate. Almost half of the testing data were detected as happy mood. The overall

Table 7. Results of mood classification using AP, AH, and ASP features.

Testing / Actual	Angry	Happy	Relaxed	Sad	Recall/ TPR
Angry	0	5	0	0	0%
Happy	0	15	0	0	100%
Relaxed	0	14	0	1	0%
Sad	0	15	0	0	0%
Precision	0%	30.61%	0%	0%	

success rate obtained was only 30%. Table 7 demonstrates the result of the experiment using the AP, AH, and ASP features.

From Table 5 we can see that the combination of AP and AH yielded the most promising results, with a success rate of 72%, where most errors occurred in relaxed mood detection.

5.4. *Improving classification accuracy using SVM-CI*

To improve classification accuracy, we added a confidence interval to the support vector machine to classify the music mood based on AP and AH. The confidence interval is a range of estimated values in a population, which is derived from a sample collected from that particular population. In this experiment, the confidence values were obtained from the standard deviation (STD) of the arousal and valence values from the training data for each module. The calculation of the valence and arousal STD values can be expressed as in Eqs. (12) and (13):

$$STD_{aro} = \sqrt{\frac{1}{N}\sum_{1}^{N}(aro_i - \overline{aro})^2} \tag{12}$$

$$STD_{val} = \sqrt{\frac{1}{N}\sum_{1}^{N}(val_i - \overline{val})^2}\,, \tag{13}$$

where aro_i, $\overline{aro}$, val_i, $\overline{val}$, N are arousal score, arousal score mean, valence score, valence mean, and amount of data, respectively. Table 8 shows the STD values obtained for each label.

Table 8. Standard deviation (STD) of arousal and valence.

	Arousal	Valence
Angry	0.9	0.7
Happy	0.8	0.8
Relaxed	0.5	0.5
Sad	0.7	0.6

Table 9. Rules to determine the confidence interval.

Axis	Friction			Rule
y	Happy	to	Relaxed	A >= 3.7
	Relaxed	to	Happy	A <= 5
	Angry	to	Sad	A >= 3.6
	Sad	to	Angry	A <= 5.2
x	Angry	to	Happy	V <= 5.2
	Happy	to	Angry	V >= 3.7
	Sad	to	Relaxed	V <= 5.2
	Relaxed	to	Sad	V >= 4

Table 10. Rule for determining the confidence interval diagonally.

Diagonally			Rule	
Angry	to	Relaxed	V <= 5.4	A >= 3.6
Relaxed	to	Angry	V >= 4	A <= 5
Sad	to	Happy	V <= 5.2	A <= 5.1
Happy	to	Sad	V <= 3.7	A >= 3.7

The STD values were used to shift the label to get the local tolerance/doubt. The obtained doubt areas were used to determine a new label. The method is to shift the original threshold to $arousal\_score \pm STD_{aro}$ or $valence\_score \pm STD_{val}$. Table 9 demonstrates the rules to determine the confidence interval for the shift on an axis. The *x*-axis determines the valence value while the y-axis determines the arousal value.

Table 10 is a reference to determine the confidence interval for mood label diagonally, for example: angry to relaxed, or sad to happy. Hence, choosing the confidence interval produces values for arousal and valence.

With a rule derived from Tables 9 and 10, the mood division diagram is as shown in Fig. 7, where the red lines indicate the area of the confidence interval.

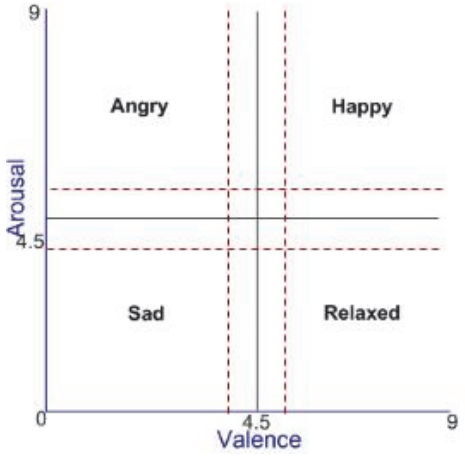


Fig. 7. Russell's diagram with confidence interval.

Table 11. Results of mood classification using AP, AH, and ASP features and SVM with confidence interval.

Testing / Actual	Angry	Happy	Relaxed	Sad	Recall/ TPR
Angry	2	3	0	0	40%
Happy	0	15	0	0	100%
Relaxed	0	5	9	1	0%
Sad	0	14	0	1	0%
Precision	100%	40.54%	100%	50%	

Table 12. Results of mood classification using AP-AH features and SVM with confidence interval.

Testing / Actual	Angry	Happy	Relaxed	Sad	Recall/ TPR
Angry	5	0	0	0	100%
Happy	0	15	0	0	100%
Relaxed	3	0	7	5	46.67%
Sad	0	0	0	15	100%
Precision	62.50%	100%	100%	75%	

Table 13. Comparison between ASP, AP-AH-ASP, AP-AH, and AP-AH with confidence interval.

		ASP	AP-AH-ASP	AP-AH	AP-AH-ASP with Confidence Interval	AP-AH with Confidence Interval
Recall	Angry	0%	0%	100%	40%	100%
	Happy	66.67%	100%	100%	100%	100%
	Relaxed	60%	0%	6.67%	0%	46.67%
	Sad	0%	0%	100%	0%	100%
Precision	Angry	0%	0%	45.45%	100%	62.50%
	Happy	40%	30.61%	100%	40.45%	100%
	Relaxed	37.50%	0%	100%	100%	100%
	Sad	0%	0%	65.22%	50%	75%
Accuracy		38%	30%	72%	54%	84%

Using the confidence interval predicts the capability of tolerating data by shifting the labels based on Tables 7 and 9. Hence, the prediction results can be said to be *true*. In this experiment, the combination of the AP, AH, and ASP features yielded the worst performance. The performance improvement based on SVM-CI can be seen in Table 11. The accuracy was 54%, which means an increase of 24% compared to before using the confidence interval.

The utilization of a confidence interval for AP-AH features is demonstrated in Table 12. Using SVM-CI, an overall success rate of 84% could be achieved with a significant improvement in relaxed mood detection. Table 12 also shows the details of recall and precision for each class.

5.5. *Overall comparison of classification recall, precision, and accuracy*

According to Table 13, SVM-CI performed better than the other methods. It improved relaxed mood detection based on the AP-AH features. SVM-CI with AP-AH features outperformed ordinary SVM with a classification accuracy of 84%. It also had the highest performance in detecting angry, happy, relaxed, and sad mood based on the values of recall and precision.

6. Conclusion

According to the experimental results of this study, the mood type of a musical piece is influenced by its rhythm and tone harmonization as represented by the Audio Power and Audio Harmonicity features in MPEG-7. Combining the Audio Power and Audio Harmonicity features in Support Vector Machine with Confidence Interval (SVM-CI), the method proposed in this paper, produced the most satisfactory results. An overall success rate of 72% was achieved using ordinary SVM, with only 6.67% for detecting relaxed mood. Meanwhile, an overall success rate of 84% was achieved using SVM with a confidence interval. The proposed method is good for music mood classification of angry, happy, and sad mood, achieving a success rate of 100%. The success rate of relaxed mood detection was 47%. This means that SVM-CI significantly improves relaxed mood detection, while there are still opportunities for further improvement. In a future work, we plan to develop a method to improve the accuracy of relaxed mood detection.

References

1. F. Wiering, Can humans benefit from music information retrieval?, in *Proc. of the 4th Int. Conf. on Adaptive Multimedia Retrieval: User, Context, and Feedback*, eds. S. Marchand-Maillet, A. Nürnberger, M. Detyniecki, pp. 82–94, doi:https://doi.org/10.1007/978-3-540-71545-0_7.
2. S. Li, H. Li and L. Ma, Music genre classification based on MPEG-7 audio features, in *Proc. of the Second Int. Conf. on Internet Multimedia Computing and Service*, pp. 185–188, doi:10.1145/1937728.1937772.
3. C.-H. Lin, M.-C. Tu, Y.-H. Chin, W.-J. Liao, C.-S. Hsu, S.-H. Lin, J.-C. Wang and J-F. Wang, SVM-based sound classification based on MPEG-7 audio LLDs and related enhanced features, in *Convergence and Hybrid Information Technology*, eds. G. Lee, D. Howard, D. Ślęzak and Y. S. Hong, pp. 536–543.
4. N. Chen, J. S. Downie, H. Xiao and Y. Zhu, Cochlear pitch class profile for cover song identification, *Applied Acoustics* **99** (2015) 92–96, doi:10.1016/j.apacoust.2015.06.003.
5. ISO/IEC (2001), Information technology — Multimedia content description interface — Part 4: Audio, in FDIS 15938-4:2001(E), n.d.

6. S. D. You, W-H. Chen and W-K. Chen, Music identification system using MPEG-7 audio signature descriptors, *The Scientific World Journal* **2013** (2013), e752464, doi:10.1155/2013/752464.
7. J. M. Ren, M. J. Wu and J. S. R. Jang, Automatic music mood classification based on timbre and modulation features, *IEEE Transactions on Affective Computing* **6** (2015) 236–246, doi:10.1109/TAFFC.2015.2427836.
8. X. Hu and Y.-H. Yang, Cross-dataset and cross-cultural music mood prediction: A case on western and Chinese pop songs, *IEEE Transactions on Affective Computing* **8** (2017) 228–240, doi:10.1109/TAFFC.2016.2523503.
9. K. L. Kermanidis, I. Karydis, A. Koursoumis and K. Talvis, Combining language modeling and LSA on Greek song "Words" for mood classification, *International Journal on Artificial Intelligence Tools* **23** (2014) 1440007, doi:10.1142/S0218213014400077.
10. M. Nardelli, G. Valenza, A. Greco, A. Lanata and E. P. Scilingo, Recognizing emotions induced by affective sounds through heart rate variability, *IEEE Transactions on Affective Computing* **6** (2015) 385–394, doi:10.1109/TAFFC.2015.2432810.
11. D. Hume, Emotion and moods, in *Organizational Behavior,* pp. 258–297, http://vig.prenhall.com/samplechapter/0132431564.pdf.
12. D. Strachan, The Space between the notes: The effects of background music on student focus, Masters of Arts in Education Action Research Papers (2015).
13. K.-S. Lin, A. Lee, Y.-H. Yang, C.-T. Lee and H. H. Chen, Automatic highlights extraction for drama video using music emotion and human face features, *Neurocomputing* **119** (2013) 111–117, doi:10.1016/j.neucom.2012.03.034.
14. B. Xing, K. Zhang, S. Sun, L. Zhang, Z. Gao, J. Wang and S. Chen, Emotion-driven Chinese folk music-image retrieval based on DE-SVM, *Neurocomputing* **148** (Elsevier, 2015) 619–627, doi:10.1016/j.neucom.2014.08.007.
15. K. Zhang and S. Sun, Web music emotion recognition based on higher effective gene expression programming, *Neurocomputing* **105** (2013) 100–106, doi:10.1016/j.neucom.2012.06.041.
16. J.L. Zhang, X.L. Huang, L. Yang and L. Nie, Bridge the semantic gap between pop music acoustic feature and emotion: Build an interpretable model, *Neurocomputing*, **208**(2015), 1–9, doi:10.1016/j.neucom.2016.01.099.
17. E. Didiot, I. Illina, D. Fohr and O. Mella, A wavelet-based parameterization for speech/music discrimination, *Computer Speech and Language* **24** (2010) 341–357, doi:10.1016/j.csl.2009.05.003.
18. S. Chen, R. C. Guido, T. Truong and Y. Chang, Improved voice activity detection algorithm using wavelet and support vector machine, *Computer Speech and Language* **24** (2010) 531–543, doi:10.1016/j.csl.2009.06.002.
19. K. R. Scherer, J. Sundberg, L. Tamarit and G. L. Salomão, Comparing the acoustic expression of emotion in the speaking and the singing voice, *Computer Speech and Language* **29** (2015) 218–235, doi:10.1016/j.csl.2013.10.002.
20. A. Ujlambkar, O. Upadhye, A. Deshpande and G. Suryawanshi, Mood based music categorization system for Bollywood music, *International Journal of Advanced Computer Research* **4** (2014).
21. S. Sharma and R. S. Jadon, Mood based music classification, *International Journal of Innovative Science, Engineering & Technology* **1** (2014).
22. V. Hampiholi, A method for music classification based on perceived mood detection for Indian bollywood music, *International Journal of Computer, Electrical, Automation, Control and Information Engineering* **6** (2012) 507–5014, doi:urn:dai:10.1999/1307-6892/15269.
23. Y. Liu and Y. Gao, Acquiring mood information from songs in large music database, in *2009 Fifth Int. Joint Conf. on INC, IMS and IDC*, pp. 1485–1491 doi:10.1109/NCM.2009.311.

24. N. Nishikawa, K. Itoyama, H. Fujihara, M. Goto, T. Ogata and H. G. Okuno, A musical mood trajectory estimation method using lyrics and acoustic features, in *Proc. of the 1st Int. ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies,* pp. 51–56, doi:10.1145/2072529.2072543.
25. E. I. V. Ascalon and R. Cabredo, Lyric-based music mood recognition, in *Proc. of the DLSU Research Congress,* Vol. 3 (2015), http://www.dlsu.edu.ph/conferences/dlsu_research_congress/2015/proceedings/HCT/009-HCT_Ascalon_EV.pdf.
26. H.-G. Kim, N. Moreau and T. Sikora, MPEG-7 audio and beyond: Audio content indexing and retrieval (2005).
27. Z. W. Ras and A. Wieczorkowska, *Advances in Music Information Retrieval,* 1st edn. (2010).
28. M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha and Y.-H. Yang, 1000 songs for emotional analysis of music, in *Proc. of the ACM Int. Multimedia Conference and Exhibition* **6** (2015) 1–14.
29. M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha and Y-H. Yang, Emotion in music database — MediaEval 2013 – aka 1000 songs (n.d.).
30. R. Sarno, B. T. Nugraha and M. N. Munawar, Real time fatigue-driver detection from electroencephalography using Emotiv EPOC+, *International Review on Computers and Software* (*IRECOS*) **11** (2016) 214–223, doi:10.15866/irecos.v11i3.8562.
31. M. N. Munawar, R. Sarno, D. A. Asfani, T. Igasaki and B. T. Nugraha, Significant preprocessing method in EEG-Based emotions classification, *Journal of Theoretical and Applied Information Technology* **87** (2016) 176–190.
32. B. T. Nugraha, R. Sarno, D. A. Asfani, T. Igasaki and M. N. Munawar, Classification of driver fatigue state based on EEG using Emotiv EPOC+, *Journal of Theoretical and Applied Information Technology* **86** (2016) 347–359.
33. D. R. Wijaya, R. Sarno, and E. Zulaika, Sensor array optimization for mobile electronic nose: Wavelet transform and filter based feature selection approach, *International Review on Computers and Software* **11** (2016) 659–671, doi:https://doi.org/10.15866/irecos.v11i8.9425.
34. D. R. Wijaya, R. Sarno and E. Zulaika, Gas concentration analysis of resistive gas sensor array, in *2016 IEEE Int. Symp. on Electronics and Smart Devices*, pp. 337–342, doi:10.1109/ISESD.2016.7886744.
35. D. R. Wijaya, R. Sarno, E. Zulaika and S. I. Sabila, Development of mobile electronic nose for beef quality monitoring, in *4th Information Systems International Conference 2017* (*ISICO*), *Procedia Computer Science* **124** (2017) 728–735, doi:10.1016/j.procs.2017.12.211.
36. R. X. Gao and R. Yan, *Wavelets: Theory and Applications for Manufacturing* (New York; London, 2010).
37. D. R. Wijaya, R. Sarno and E. Zulaika, Information quality ratio as a novel metric for mother wavelet selection, *Chemometrics and Intelligent Laboratory Systems* **160** (2016) 59–71, doi:10.1016/j.chemolab.2016.11.012.
38. P. Saari, G. Fazekas, T. Eerola, M. Barthet, O. Lartillot and M. Sandler, Genre-adaptive semantic computing and audio-based modelling for music mood annotation, *IEEE Transactions on Affective Computing* **7** (2016) 122–135, doi:10.1109/TAFFC.2015.2462841.
39. M. Casey, MPEG-7 sound-recognition tools, *IEEE Transactions on Circuits and Systems for Video Technology* **11** (2001) 737–747, doi:10.1109/76.927433.
40. H.-G. Kim and T. Sikora, Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation, in *2004 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 5 (2004), pp. 7–10, doi:10.1109/ICASSP.2004.1327263.